

Probability and Statistics



SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES
INDIAN INSTITUTE OF TECHNOLOGY MANDI
KAMAND-175075, INDIA

Contents

1	Python Tutorial	1
1.1	Python manual	1
1.1.1	Types of Variables	1
1.1.1.1	Integer	1
1.1.1.2	Float	2
1.1.1.3	String	2
1.1.1.4	Boolean	2
1.1.2	Comments	3
1.1.3	Indentation	3
1.1.4	Operators	3
1.1.4.1	Arithmetic Operators	4
1.1.4.2	Comparison Operators	4
1.1.4.3	Logical Operators	5
1.1.5	Conditional Statements	5
1.1.6	Loops	5
1.1.6.1	For Loops	6
1.1.6.2	While Loops	7
1.1.6.3	Break and Continue Statements	7
1.1.6.4	Nested Loops	7
1.1.7	User Input	7
1.1.8	Typecasting	8
1.1.9	List	9
1.1.10	Tuple	10
1.1.11	Sets	10
1.1.12	Dictionaries	11
1.1.13	Function	11
1.2	Data Analysis and Visualization	13

1.2.1	Pandas	13
1.2.2	Numpy	15
1.2.3	Matplotlib	16
1.3	Practice Problems-1	17
2	Set Theory	22
2.1	Set	22
2.1.1	Definitions	22
2.1.1.1	Sample Space	22
2.1.1.2	Event	22
2.1.1.3	Mutually Exclusive Event	23
2.1.1.4	Empty set	23
2.1.1.5	Set Membership Symbol	24
2.1.1.6	Cardinality of a set	24
2.1.2	Operations	24
2.1.2.1	Union	24
2.1.2.2	Intersection	25
2.1.2.3	Complement	25
2.1.2.4	Set difference	26
2.1.3	Demorgan's Law	26
3	Introduction to Probability	27
3.1	Short History	27
3.2	Basic Terminology	27
3.2.1	Random Experiment	27
3.2.2	Sample Space	28
3.2.3	Deterministic Phenomena	28
3.2.4	Non-deterministic Phenomena	28
3.2.5	Mutually exclusive events	30
3.2.6	Combinations of Events	30
3.2.7	Disjoint Events	31
3.2.8	Conditional Probability	31
3.2.9	Independence	33
3.3	Correlation	33
3.4	Return level estimation	44
3.5	Lab Assignment 1	49
3.6	Lab Assignment 2	57

3.7	Theory Assignment 1	61
3.8	Theory Assignment 2	64
3.8.1	Theory Assignment-2-solutions	65
4	Bayes' Theorem and Counting	67
4.1	Bayes' Theorem	67
4.1.1	Bayes' Theorem Examples	68
4.2	Counting techniques	75
4.2.1	Finite uniform probability space	76
4.2.2	Techniques for counting	76
4.2.3	Rule 1	76
4.2.4	Rule 2	77
4.2.5	The Multiplicative Rule of Counting	78
4.2.6	Permutations:	79
4.2.7	Permutations of size $k (< n)$:	80
4.2.8	Combinations of size $(k < n)$:	81
4.3	Birthday Paradox	87
4.3.1	Experiment, Observations and results	87
4.3.2	Results of above experiment by simulations	88
4.4	Theory Assignment-3	88
4.4.1	Questions	88
4.4.2	Answers	89
4.5	Lab Assignment 3	91
5	Discrete and Continuous Random Variables	98
5.1	Introduction	98
5.2	Definition	98
5.3	Examples	99
5.4	Types	99
5.5	Discrete Random Variables	100
5.5.1	Probability Mass Function	100
5.5.2	Cumulative Distribution Function	103
5.5.3	Functions of Random Variables	104
5.5.4	Special Distributions	106
5.5.4.1	Bernoulli Distribution	106
5.5.4.2	Binomial Distribution	107
5.5.4.3	Hypergeometric distribution	108

5.5.4.4	Poisson Distribution	108
5.6	Practice Problem Set-1	110
5.7	Practice Problem Set-1 Solutions	110
5.8	Practice Problem Set-2	112
5.9	Practice Problem Set-2 Solutions	113
5.10	Practice Problem Set-3	115
5.11	Practice Problem Set-3 Solutions	118
6	Random Variables	130
6.1	Introduction	130
6.2	Definition	130
6.3	Examples	131
6.4	Types	132
6.5	Discrete Random Variables	132
6.5.1	Probability Mass Function	133
6.5.2	Cumulative Distribution Function	136
6.5.3	Functions of Random Variables	140
6.5.4	Special Distributions	141
6.5.4.1	Bernoulli Distribution	141
6.5.4.2	Binomial Distribution	143
6.5.4.3	Geometric Distribution	144
6.5.4.4	Hypergeometric distribution	145
6.5.4.5	Poisson Distribution	146
6.6	Continuous Random Variable	148
6.6.1	Probability Density Function	149
6.6.2	Cumulative Distribution Function	152
6.6.3	Special Distributions	153
6.6.3.1	Uniform Distribution	153
6.6.3.2	Exponential Distribution	155
6.6.3.3	Normal Distribution	157
6.7	Theory Assignment 6	171
6.7.1	Theory Assignment 6 Solutions:	172
6.8	Lab Assignment 4	175
6.9	Mock-Mid term exam	184
6.9.1	Mock-Mid Term Exam-Solutions	186
6.10	Bivariate Distributions	190
6.10.1	Joint Probability Distributions	190

6.10.2 Marginal Probability Distributions	193
6.10.3 Conditional Probability Distributions	195
6.10.4 Independence	197
6.11 Summary	199
6.12 Practice Problem Set-1	205
6.13 Practice Problem Set-1 Solutions	205
6.14 Practice Problem Set-2	208
6.15 Practice Problem Set-2 Solutions	209
6.16 Practice Problem Set-3	211
6.17 Practice Problem Set-3 Solutions	214
6.18 Practice Problem set-4	227
6.19 Practice set -4 solutions	229
6.20 Lab Assignment 6	233
7 Expectation	245
7.1 What is expected value?	245
7.1.1 Formal definitions:	246
7.1.2 Example:	246
7.1.3 Theorem 1	247
7.1.4 The properties of $E(X)$ are as follows.	247
7.1.5 Expectation of Two-dimensional Random Variable	248
7.1.5.1 Definition	248
7.1.5.2 Theorem 2	248
7.1.5.3 Theorem 3	249
7.1.6 Relation between Expectation and Moments	249
7.1.6.1 Definition	249
7.1.6.2 Definition	249
7.1.6.3 Definition	249
7.2 Variance	251
7.2.1 Variance Definition:	251
7.2.2 Variance Interpretation	251
7.3 Expectation of $g(X)$	251
7.4 Mean / Expected value Function of a random variables	252
7.4.1 Definition:	252
7.5 Mean / Expected value Function of two random variables	252
7.5.1 Problem:	252
7.6 Expectation of XY : the definition of $\mathbb{E}(XY)$	253

7.7	Properties of Expectation	253
7.8	Probability as an Expectation	254
7.9	Variance, covariance, and correlation	254
7.9.1	Covariance	255
7.9.2	Properties of Variance	256
7.10	Conditional Expectation and Conditional Variance	256
7.10.1	Conditional expectation as a random variable	257
7.10.2	Conditional expectation of X given random Y is a random variable:	258
7.10.3	Conditional variance	259
7.11	Risk and Return: Example of expectations and Variance	259
7.11.1	Expected Return	259
7.11.2	Measures of Risk	260
7.11.3	Portfolio Risk and Return	261
7.12	Midterm question paper with solutions	265
7.13	Practice Problems	270
7.14	Lab Assignment 5	277
7.15	Lab Assignment 7	295
8	Estimation Theory	303
8.1	Estimation	303
8.1.1	Difference between likelihood and probability.	303
8.1.2	Normal Parameter	304
8.1.3	Sample variance	305
8.1.4	Criteria for selection of estimator:	306
8.2	Statistical Inference	307
8.3	Estimation	308
8.3.1	Point Estimator	308
8.3.2	Interval Estimator	309
8.4	Quality of Estimators	309
8.4.1	Unbiasedness	309
8.4.2	Consistency	310
8.4.3	Efficiency	310
8.5	Estimating μ When σ^2 is Known	310
8.6	It follows that for a given α , we have	311
8.7	Interpretation:	312
8.8	Width of Confidence Interval	312
8.9	Details	313

8.10 Selecting the Sample Size	314
9 Sampling Theory and Hypothesis Testing	315
9.1 Sampling	315
9.1.1 Statistical Inference	315
9.1.2 Parameter estimation	316
9.1.3 Notation	316
9.1.4 Random Sampling	317
9.1.5 Law of Large Numbers	320
9.1.6 The Central Limit Theorem	321
9.2 Inference with Sample Mean	322
9.3 Estimating the difference between two mean	322
9.4 Confidence interval (CI)	323
9.4.1 Population Mean	324
9.4.1.1 σ - known	326
9.4.1.2 σ - unknown	328
9.4.2 Proportion	330
9.5 What is Hypothesis Testing?	331
9.5.1 Why do we use it?	332
9.5.2 What are the basics of Hypothesis?	332
9.5.3 Which are important parameter of hypothesis testing?	333
9.6 Testing a hypothesis about the mean of a population	334
9.6.1 Testing hypothesis for the mean μ	335
9.6.2 Test Types	335
9.7 The Use of P–Values in Decision Definition Making	337
9.7.1 What is P–Values?	337
9.7.2 An Alternative Decision Rule using the P-value Definition	338
9.8 Hypothesis test for the population correlation coefficient ρ	339
9.8.1 Steps for Hypothesis Testing for ρ	339
9.8.1.1 Hypotheses	339
9.8.1.2 Test Statistic	339
9.8.1.3 P–Value	340
9.8.1.4 Decision	340
9.9 Problems based on rainfall data	340
9.9.1 Question 1	341
9.9.1.1 Solution/explanation	341
9.9.2 Question 2	343

9.9.2.1	Solution/explanation	343
9.9.3	Question 3	346
9.9.3.1	Solution/explanation	346
9.9.4	Question 4	349
9.9.4.1	Solution/Explanation	349
9.10	Problems	352
9.10.1	Task 1: Hypothesis Testing	352
9.10.1.1	Solution	352
9.10.1.2	Output	352
9.10.2	Task 2: Confidence Intervals	353
9.10.2.1	Solution	353
9.10.2.2	Output	353
9.10.3	Task 3: Sample Size Calculation	353
9.10.3.1	Solution	354
9.10.3.2	Output	354
9.10.4	Task 4	354
9.10.4.1	Solution	354
9.10.4.2	output	355
10	Regression Analysis	357
10.1	Introduction	357
10.1.1	Assumptions of Regression	357
10.1.1.1	Linearity	357
10.1.1.2	Independence	358
10.1.1.3	Homoscedasity	358
10.1.1.4	No endogeneity	358
10.1.1.5	No perfect Multicollinearity	358
10.1.1.6	Normality of Residuals	358
10.1.1.7	No autocorrelation	358
10.1.2	Mean Squared Error	359
10.1.3	R ² score	359
10.2	Scatter Plots	359
10.3	Simple Linear Regression	360
10.4	Multivariate Regression	360
10.5	Polynomial Regression	360
10.6	Applying the regression on a Real-World Data	360
10.6.0.1	Conclusion	364

10.6.1	Covid Data	365
10.6.1.1	Conclusion	372
10.6.2	Practice Problems	373
10.7	Regression Questions	375
11	Descriptive Statistics	380
11.1	What is Descriptive Statistics?	380
11.1.1	Measures of Central Tendency	380
11.1.1.1	Mean	380
11.1.1.2	Median	381
11.1.1.3	Mode	381
11.1.2	Measures of Dispersion	383
11.1.2.1	Range	383
11.1.2.2	Variance	383
11.1.2.3	Standard Deviation	383
11.1.3	Skewness and Kurtosis	383
11.1.3.1	Skewness	384
11.1.3.2	Kurtosis	387
11.1.4	Boxplots and Interquartile Range	389
11.1.4.1	Boxplots	389
11.1.4.2	Interquartile Range	389
11.1.4.3	Identifying Outliers by help of boxplot	390
11.1.5	Descriptive analysis on Rainfall data using Python	390
11.1.5.1	Understanding how outliers affects the data	390
11.1.5.2	Using different ways to analysis Data	392
11.1.5.3	Different parameters variance comparison	396
11.1.5.4	Cumulative analysis	398
11.1.5.5	Using spatial plots for location based data	399
11.2	Lab Assignment 10	403

List of Tables

3.1	Table for Question-3	60
6.1	Table for Question-7a	189
6.2	Joint probability mass function	192
6.3	Joint cumulative function	192
6.4	Marginal probability mass function	194
6.5	Probability Distribution of goals	227
6.6	Joint Probability Mass Function	230
10.1	Exam and Homework Scores	376

List of Figures

1.1	Flow Diagram of if else block	6
1.2	Flow Diagram of break statement	8
1.3	Flow Diagram of continue statement	9
3.1	Correlation between monthly maximum and minimum temperature from 1981-2021 over 525 grids of NWH.	36
3.2	Mean of annual maximum temperature from 1981-2021 over 525 grids of NWH.	38
3.3	Standard deviation of annual maximum temperature from 1981-2021 over 525 grids of NWH.	40
3.4	The best-fitted distribution for maximum temperature from 1981-2021 over 525 grids of NWH.	43
3.5	Return level estimates for 100 years return period 525 grids of NWH.	49
3.6	Plot: Q3	55
3.7	Plot: Q4	57
6.1	Random Variable	131
6.2	Probability Mass Function	135
6.3	Probability mass function of a Bernoulli random variable	142
6.4	Probability density function	149
6.5	Histogram of weights	150
6.6	Histogram of weights	150
6.7	Probability density function	151
6.8	Finding probability using pdf curve	151
6.9	Cumulative distribution function	152
6.10	Probability density function of a uniform random variable	154
6.11	Probability density function of an exponential random variable with $\lambda = 1$	155
6.12	Memoryless property of an exponential random variable	156
6.13	Effect of the parameters of a normal random variable	158
6.14	Probability density function of a normal random variable	158

6.15	Probability density function of a standard normal random variable	159
6.16	Cumulative distribution function of a standard normal random variable	159
6.17	Symmetry of the standard normal random variable	160
6.18	<i>Z</i> table example	161
6.19	Right skewed distribution	165
6.20	Left skewed distribution	165
6.21	Symmetric distribution	165
6.22	Symmetric distribution of marks	166
6.23	<i>Q – Q</i> plot of a data which isn't normally distributed	166
6.24	<i>Q – Q</i> plot of a data which is normally distributed	167
6.25	<i>Q – Q</i> plot of a left skewed distributed data	168
6.26	<i>Q – Q</i> plot of a right skewed distributed data	168
6.27	Comparison of the <i>pmf</i> of $\text{Bin}(16, 0.5)$ with the <i>pdf</i> of $N(8, 4)$	171
6.28	The random walk problem. <i>Left:</i> Equal probability of moving in both directions. <i>Right:</i> Because of the inclination, the person has more probability of moving towards right as compared to the left direction.	177
6.29	Marginal probability density function of X	194
6.30	Marginal probability density function of Y	195
6.31	<i>Z</i> distribution table-1	203
6.32	<i>Z</i> distribution table-2	204
6.33	Joint probability mass function	227
6.34	Plot: Q2	237
7.1	Plot: Q2	299
7.2	Plot: Q3	301
9.1	<i>Distribution of Sample Means with 21 Samples</i>	320
9.2	<i>Distribution of Sample Means with 96 Samples</i>	320
9.3	<i>Distribution of Sample Means with 170 Samples</i>	321
9.4	<i>Classification of Confidence interval</i>	324
9.5	<i>Confidence interval distribution</i>	324
9.6	<i>Confidence interval distribution</i>	325
9.7	<i>Confidence interval distribution</i>	325
9.8	<i>Confidence level and corresponding z - values</i>	325
9.9	<i>t - distribution table</i>	329
9.10	Normal Curve images with different mean and variance	332
9.11	Standardised Normal curve image and separation on data in percentage in each section	333

9.12 Locations where rainfall is more in summer than in winter	343
9.13 locations where Rainfall in recent years is more	345
9.14 monthly rainfall at location no. 50	345
9.15 monthly rainfall at location no. 150	346
9.16 locations where temperature has increased in recent years	348
9.17 monthly average of temperature at location 150	348
9.18 monthly average of temperature at location 50	349
9.19 locations where correlation between rainfall and Humidity exists	351
9.20 correlation(r-value) at every location	351
9.21 Sampling distribution of sample mean	355
9.22 Sampling distribution of sample mean with p-value	356
10.1 Actual vs Predicted Data Comparison via Simple Linear Regression	362
10.2 Actual vs Predicted Data Comparison via Multivariate Linear Regression	363
10.3 Actual vs Predicted Data Comparison via Polynomial Regression	365
10.4 Recovery Rate plot in all three scenarios	372
10.5 Vaccination coverage plot in all three scenarios	373
10.6 Retail mobility plot in all three scenarios	374
10.7 Transit mobility plot in all three scenarios	375
11.1 Mode of Average Temp of each day in 2020	382
11.2 Postively skewed data	384
11.3 Negatively skewed data	385
11.4 Skewness of histogram of Average Temperature of each day in 2020	386
11.5 Diffrent types of kurtosis	388
11.6 Boxplot	391
11.7 Boxplot without outlier	392
11.8 Comparing 5 year data 2017-2021 with 2022	395
11.9 Specific humidity and precipitation with there respective means	397
11.10Cumulative rainfall for last 5 years 2017-2021	399
11.11Total average rainfall in an year in mm	401
11.12Daily average windspeed	401
11.13Daily average specific humidity	402
11.14Correlation Between rainfall and specific humidity	403
11.15(For mean of temperature (yearwise))	406

Chapter 1

Python Tutorial

1.1 Python manual

If you have an interest in Data Science, Web Development, Robotics, or IoT you must learn Python. Python has become the fastest-growing programming language due to its heavy usage and wide range of applications. Python is a high level and object oriented programming language which uses an interpreter to convert into a low level language and execute the program. Let's start with installation. We need a python interpreter and an ide which helps us to easily write our code. Ide have their own advantages like syntax highlighting(Syntax- The syntax of the programming language is the set of rules that defines how a program will be written.), console and many more. There are many IDEs for python but the best is to install anaconda.

Youtube video

Let's start with writing our first code we will use spider ide for this print("Hello world")

When you run the code the console will have the following output

Hello world

In this line of code we will have two things first is print statement or other is "Hello world"(string) Let's discuss variables and type of variables A variable is a way of referring to a memory location used by a computer program. Well in most programming languages you need to assign the type to a variable. But in Python, you don't need to, so it is called a dynamically typed language. For example, to declare an integer in C++, the following syntax is used:

```
1 int num = 5;. In Python it's num = 5 where "=" is the assigning operator.
```

1.1.1 Types of Variables

1.1.1.1 Integer

Numerical values that can be positive, negative, or zero without a decimal point.

```
1     Code :
2     num=5
3     print(num)
4     print(type(num))
5     Output:
6     5
7     <class 'int'>
```

1.1.1.2 Float

Similar to an integer but with one slight difference – floats are a numerical value with a decimal place.

```
1     num=5.0
2     print(num)
3     print(type(num))
4     Output:
5     5
6     <class 'float'>
```

1.1.1.3 String

A formation of characters or integers. They can be represented using double or single quotes.

```
1     num='Hello world'
2     print(num)
3     print(type(num))
4     Output:
5     Hello world
6     <class 'str'>
```

1.1.1.4 Boolean

A binary operator with a True or False value.

```
1     num=True
2     print(num)
3     print(type(num))
4     Output:
5     True
6     <class 'bool'>
```

There is one more type you can try by assigning None keyword to a variable. Keywords are the reserved words. We cannot use a keyword as a variable name, function name.

1.1.2 Comments

Comments make it easy to write code as they help us (and others) understand why a particular piece of code was written. These are sentences which are ignored by the compiler.

Single line comment:

By using `#` we can write single line comments

Multiline comment:

By using `"""` for opening and `"""` for closing the multiline comment

```
# single line comment
```

```
"""
```

Multiline comments

```
"""
```

So we add some space in the beginning of the above code like this

```
1 print(" h e l l o )
```

And run the code we will get indentation error.

1.1.3 Indentation

Another interesting part of this language is indentation. Why? Well, the answer is simple: It makes the code readable and well-formatted. It is compulsory in Python to follow the rules of indentation.

If proper indentation is not followed you'll get the following error:

```
IndentationError: unexpected indent
```

1.1.4 Operators

Operators are special symbols in Python that carry out arithmetic or logical computation.

Operator	Name	Example
+	Addition	a+b
-	Subtraction	a-b
*	Multiplication	a*b
/	Division	a/b
%	Modulo	a%b
**	Exponentiation	a**b
//	Floor Division	a//b
=	Assignment	a=10
+=	Addition Shorthand	a+=2 (same as a=a+2)
-=	Subtraction Shorthand	a-=2 (same as a=a-2)
=	Multiplication Shorthand	a=2 (same as a=a*2)
/=	Division Shorthand	a/=2 (same as a=a/2)
==	Equal	a==b
!=	Not Equal	a!=b
>	Greater than	a>b
<	Less than	a < b
≥	Greater than or equal to	a≥b
≤	Less than or equal to	a≤b
and	True if both statements are true	a < 5 and b > 10
or	True if either statement is true	a < 5 or b > 10
not	Reverses the result. If result is true then false and vice versa	not(a<5)

1.1.4.1 Arithmetic Operators

These include addition, subtraction, deletion, exponentiation, modulus, and floor division. Also the shorthand syntax for some operators.

```

1 a=12
2 b=2
3 print(12-2)
4 Output:
5 10

```

1.1.4.2 Comparison Operators

These include equal to, greater than, and less than.

1.1.4.3 Logical Operators

These operators include not, and, & or.

```
1 Code
2 print(12>10)
3 Output:
4 True
5 You can try the rest of the operators.
```

1.1.5 Conditional Statements

As the name suggests, conditional statements are used to evaluate if a condition is true or false. Many times when you are developing an application you need to check a certain condition and do different things depending on the outcome. In such scenarios conditional statements are useful. If, elif for(else if)and else are the conditional statements used in Python.

We can compare variables, check if the variable has any value or if it's a boolean, then check if it's true or false.

All Integers except 0 are considered as true. All strings considered as true.

```
1 if logical_expression_1:
2     Block1
3     elif logical_expression_2:
4         block2
5     else
6         Block3
```

The flow of if else statements will go like this if logical expression 1 is true then the first indented code block will execute and the rest of elif or else statement will be ignored. But if the first logical expression is false it will go to the elif part and check logical expression 2. And if all the logical statements are false then the else statements will get executed.

Now u can play with the above code and see what will be the output for different combinations of true and false.

1.1.6 Loops

Another useful method in any programming language is an iterator. If you have to implement something multiple times, what will you do?

Well, that's one way to do it, write these things multiple times. But imagine you have to do it a hundred or a thousand times. Well, that's a lot of print statements we have to write. There's a better way called iterators or loops. We can either use a for or while loop.

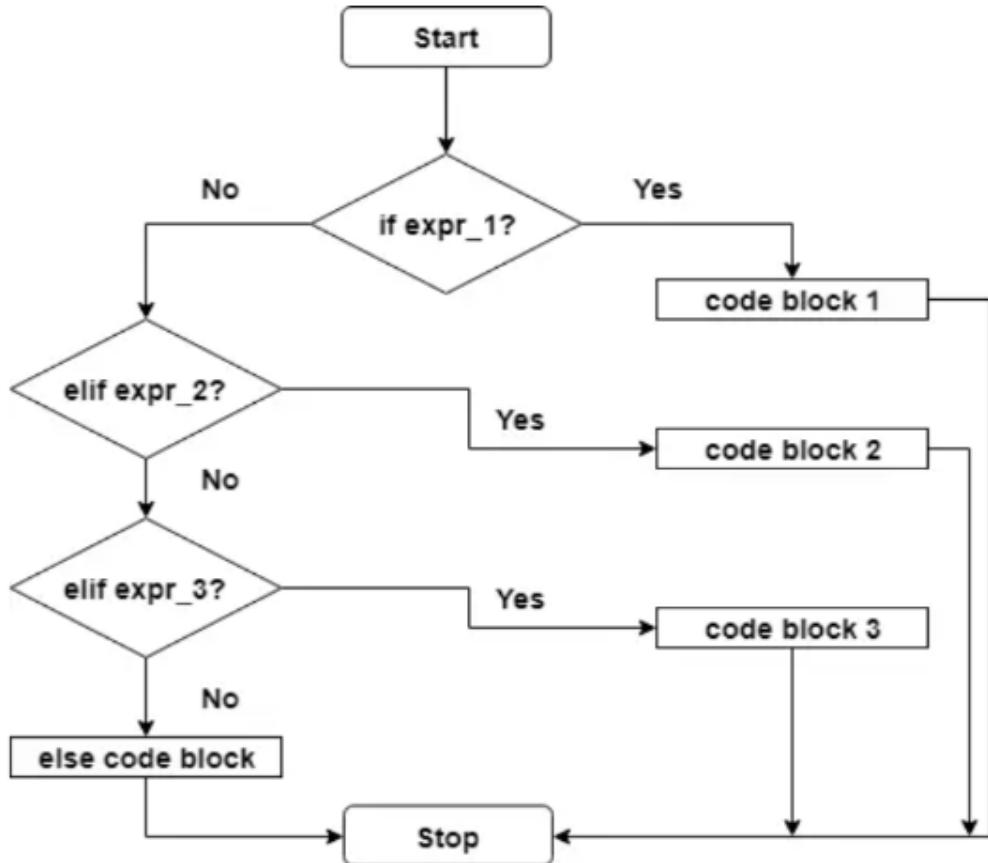


Fig. 1.1: Flow Diagram of if else block

1.1.6.1 For Loops

```

1  Code:
2  for i in range(5):
3      print("Hello")
4  Code :
5  for i in range(3,15,3):
6      print(i)

```

For loops have two components one is iterator and other is range function range function allows the iterator to change its value over the range. Range function asks for three arguments: start, end and step. In the second code the value of i should start from 5 and it should be strictly less than 15 and incremented by 2 in each step. While in the first code we only give the end parameter the start value

by default is 0 and step is 1 so the value of i starts from 0 and increments by 1 in each step. You can try to work out What will happen when we give two values for the in range function that will be considered as the end value and step or start and end.

1.1.6.2 While Loops

```
1 Code:  
2 while (logical_expression) :  
3     Indentation block
```

While the logical expression is true the intended block will execute if the logical expression is true for forever then the intended block will execute for infinite times and the program will run for forever. There is one more kind of loop “do while” but we do not use it frequently though you can read about that on the internet.

1.1.6.3 Break and Continue Statements

Break : The break statement terminates the loop containing it. Control of the program flows to the statement immediately after the body of the loop. If the break statement is inside a nested loop (loop inside another loop), the break statement will terminate the innermost loop.

Continue : The continue statement is used to skip the rest of the code inside a loop for the current iteration only. Loop does not terminate but continues on with the next iteration.

1.1.6.4 Nested Loops

When a loop is inside another loop they form nested loops.

```
1 Code:  
2 for i in range(5):  
3     for j in range(3):  
4         print(i, j)
```

The inner loop will run for each i so we will get 15 print statements.

1.1.7 User Input

Sometimes you have to take the user’s input and act accordingly. To do that you can use Python’s inbuilt input method.

When you use the input method and press enter, you’ll be prompted with the text that you enter in the input method.

```
1 name=input()  
2 print(name)  
3 print(type(name))
```

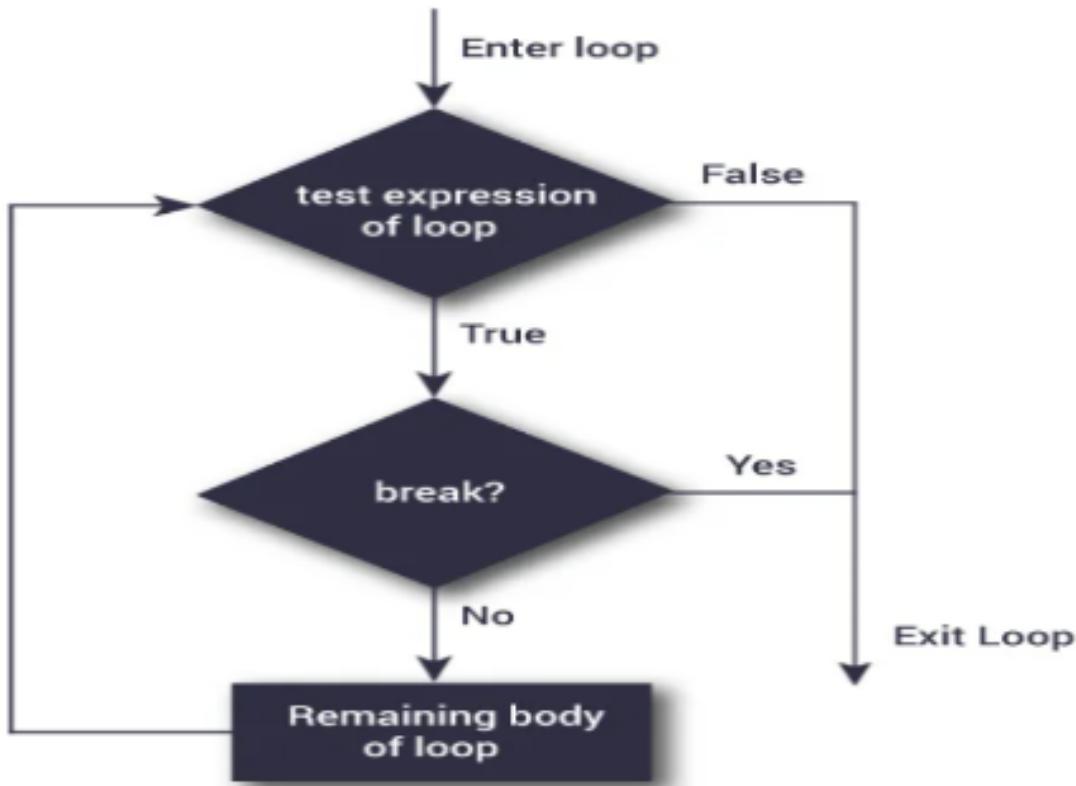


Fig. 1.2: Flow Diagram of break statement

Even if we entered an integer 12 and it's still giving us its type as a string. It's not a bug. It's how input is intended to work, it always takes input as a string. To convert the string to integer we will use typecasting.

1.1.8 Typecasting

By typecasting we can convert one type of variable to another if they are convertible. Eg you cannot convert "Hello" string to int because it's not a valid integer but we can convert "12" to int 12.

```

1 name=int(input())
2 print(name)
3 print(type(name))
4 The answer of print(type(num)) in this case is < class      int      >.
  
```

There are some more data types which help us to deal with data.

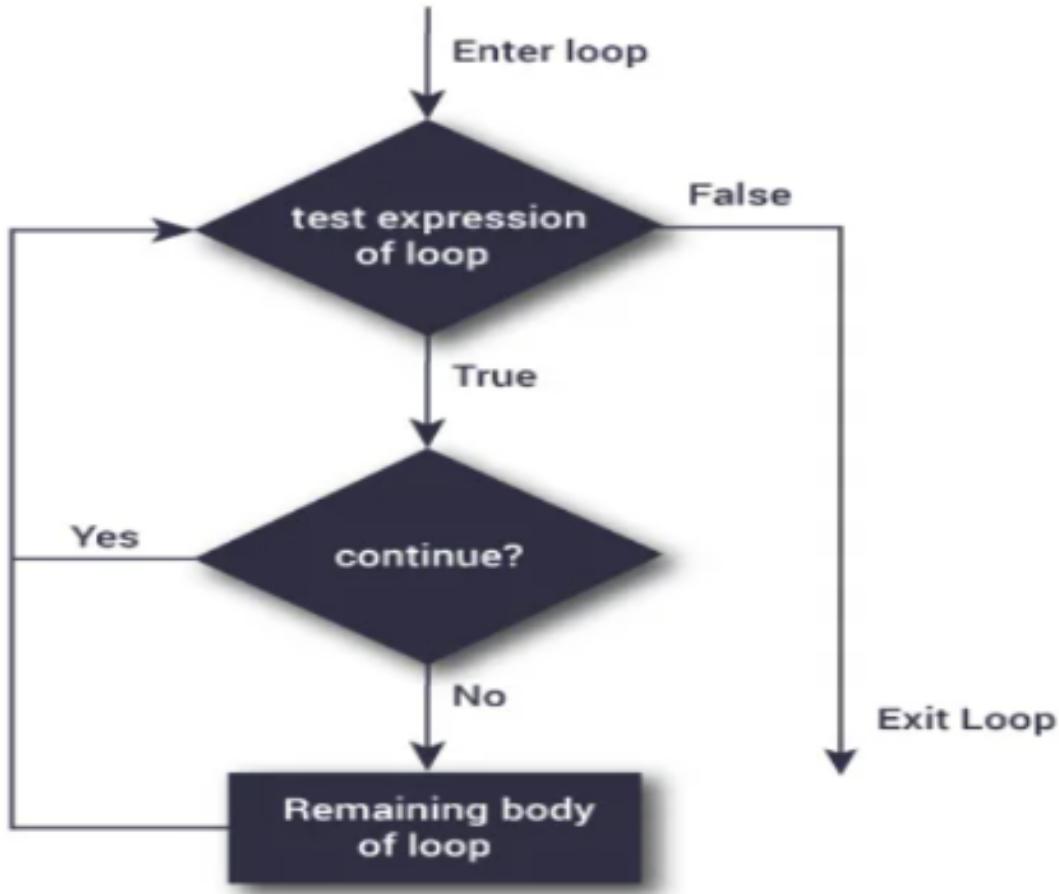


Fig. 1.3: Flow Diagram of continue statement

1.1.9 List

Imagine you have a bunch of data that is not labeled. In other words, each piece of data doesn't have a key that defines it. So how will you store it? Lists to the rescue. They are defined as follows:

```
1 data = [1,5,'XYZ',True]
```

A list is a collection of random, ordered, and mutable data (i.e., it can be updated).

The indexing of the position of the elements begins from zero.

To access the first element we need to access it as follows:

```
1 data = [0]
2 >> 1
```

You can also specify a range to access the element between those positions.

```
1 data[2:4]
2 >>['xyz', True]
```

Here, the first value represents the start while the last value represents the position until which we want the value.

To add an item in the list we need to use the append method provided by python.

```
1 data.append("Hello")
2 data=[ 1, 5, "xyz", True , Hello ]
```

We can also loop through the list to find a certain element and operate on it.

```
1 Code:
2 data = [ 1, 5, "xyz", True ]
3 for i in data:
4     print(i)
5 Output:
6 1
7 xyz
8 True
```

There are many more things you can take a look like removing an element from a list, checking a element in the list etc.

1.1.10 Tuple

The tuple is a data type which is ordered and immutable (i.e. data cannot be changed).

Let's create a tuple:

```
1 data =left(1, 3 , 5, "bye")
```

We can access elements in the tuple the same way as we access them in a list:

```
1 data[3]
2 >>'bye'
```

We can access the index range as follows:

```
1 data[2:4]
2 >> (5, 'bye')
```

We cannot change the value of tuple as it is immutable.[Note: Once a tuple is created a new value cannot be added to it.].

1.1.11 Sets

Sets are another data type in Python which are unordered and unindexed. Sets are declared as follows:

```

1   data = { "hello", "bye", 10, 15}
2   print(data)
3   >> {10, 15, 'hello', 'bye'}

```

As sets are unindexed we cannot directly access the value in a set.

Thus to access the value in the set you need to use a for loop.

```

1 code:
2 for i in data:
3     print(i)
4 Output:
5 10
6 15
7 hello
8 bye

```

Once the set is created, values cannot be changed but we can add and remove values from a set using the add and remove function

1.1.12 Dictionaries

Python dictionary is an unordered collection of items. Each item of a dictionary has a key/value pair.

Creating a dictionary:

```

1 my_dict = {1: 'apple', 2: 'ball'}

```

1.1.13 Function

Functions are a block of code that helps us in the reusability of repetitive logic. Functions can be both inbuilt and user-defined.

To declare a function we use the def keyword. Following is the syntax of the functions:

```

1 def hello_world():#function declaration
2     print("Hello world")

```

Here we are declaring a function which, when called, prints a "Hello world" statement. To call a function we use the following syntax:

```

1 hello_world()#function calling

```

We will get the following output:

```

1 Hello world

```

Practice Problems: Hackerrank

Reference:

Method	Description
<code>clear()</code>	Removes all items from the dictionary.
<code>copy()</code>	Returns a shallow copy of the dictionary.
<code>fromkeys(seq[, v])</code>	Returns a new dictionary with keys from <code>seq</code> and value equal to <code>v</code> (defaults to <code>None</code>).
<code>get(key[,d])</code>	Returns the value of the <code>key</code> . If the <code>key</code> does not exist, returns <code>d</code> (defaults to <code>None</code>).
<code>items()</code>	Return a new object of the dictionary's items in (key, value) format.
<code>keys()</code>	Returns a new object of the dictionary's keys.
<code>pop(key[,d])</code>	Removes the item with the <code>key</code> and returns its value or <code>d</code> if <code>key</code> is not found. If <code>d</code> is not provided and the <code>key</code> is not found, it raises <code>KeyError</code> .
<code>popitem()</code>	Removes and returns an arbitrary item (<code>key, value</code>). Raises <code>KeyError</code> if the dictionary is empty.
<code>setdefault(key[,d])</code>	Returns the corresponding value if the <code>key</code> is in the dictionary. If not, inserts the <code>key</code> with a value of <code>d</code> and returns <code>d</code> (defaults to <code>None</code>).
<code>update([other])</code>	Updates the dictionary with the key/value pairs from <code>other</code> , overwriting existing keys.
<code>values()</code>	Returns a new object of the dictionary's values

Python Documentation

Free Code Camp

Programiz

For machine learning or data analysis jupyter notebook is quite useful so i would suggest you all to go through this:

Getting Started With Jupyter Notebook for Python

1.2 Data Analysis and Visualization

Various Libraries in Python are used for data analysis, manipulation and visualization. Some of the most common libraries include Pandas, Numpy, Matplotlib, Seaborn and Scikit learn. You will be using these libraries throughout the lab work of this course, so it is recommended that you explore these libraries in depth on your own. If you are stuck with any syntactic error, you can refer to stackoverflow or the original documentation of these libraries. We will look briefly into each of these libraries and their applications.

1.2.1 Pandas

Open-source library that is made mainly for working with relational or labeled data.

Functions for analyzing, cleaning, exploring, and manipulating data

To install Pandas just type ‘pip install pandas’ in the terminal (given that you already have python and pip installed in your system).

Series - A Pandas Series is like a column in a table.

```
1 Code :  
2 import pandas as pd  
3 a = [1, 7, 2]  
4 myvar = pd.Series(a)  
5 print(myvar)  
6 Output :  
7 0    1  
8 1    7  
9 2    2  
10 dtype: int64
```

This is an array sort of data type where you can access the series elements by index (starting from 0). We can also give our own labels rather than accessing from the indices by using the “index” argument in pd.Series().

```
1 Code :  
2 import pandas as pd  
3 a = [1, 7, 2]  
4 myvar = pd.Series(a, index = ["x", "y", "z"])  
5 print(myvar[    y    ])  
6 Output :  
7 7
```

We can also give key-value objects while creating a Series.

```

1 Code :
2 import pandas as pd
3 calories = {"day1": 420, "day2": 380, "day3": 390}
4 myvar = pd.Series(calories)
5 print(myvar)
6 Output:
7 day1    420
8 day2    380
9 day3    390
10 dtype: int64

```

DataFrames - Datasets in Pandas are usually multi-dimensional tables called DataFrames. These are the most important in Pandas and are used frequently.

```

1 Code :
2 import pandas as pd
3 data = [
4     {"calories": [420, 380, 390], "duration": [50, 40, 45]}
5 #load data into a DataFrame object:
6 df = pd.DataFrame(data)
7 print(df)
8 Output :
9          calories   duration
10      0        420         50
11      1        380         40
12      2        390         45

```

We can use .loc[r,c] to locate any particular element in the dataframe (r = row index, c = column index).

df.head() is used to obtain the first 5 entries in the dataframe

Reading CSV Files - We can read CSV files with the help of pandas which loads all the data into a data frame.

```

1 import pandas as pd
2 df = pd.read_csv('data.csv')

```

The argument in the read_csv function is the path of the file we are trying to read
To Learn More:

Pandas Tutorial (w3schools.com)

Python - Pandas (tutorialspoint.com)

Pandas Documentation — Pandas 1.5.3 documentation (Pydata.org)

1.2.2 Numpy

Numpy stands for numerical python and is used for working with arrays.

It has functions for working in the domain of linear algebra, fourier transform, and matrices. It is vastly used for performing matrix manipulations in python.

Numpy arrays are almost 50 times faster than python arrays. The array object in NumPy is called ndarray.

To install Numpy just type ‘pip install numpy’ in the terminal (given that you already have python and pip installed in your system).

```
1 import numpy as np    # importing numpy library under np alias
2 arr = np.array([1, 2, 3, 4, 5]) # converts normal Python list to numpy array
3 print(arr)
```

We can use arrays having multiple dimensions

```
1 Eg.
2 arr = np.array(42) # 0-D np array
3 arr = np.array([1, 2, 3, 4, 5]) # 1-D np array
4 arr = np.array([[1, 2, 3], [4, 5, 6]])# 2-D
5 arr = np.array([[[1, 2, 3], [4, 5, 6]], [[1, 2, 3], [4, 5, 6]]]) # 3-D
```

arr.ndim() # gives the dimension of the arr array

Elements in the array are accessed just like in any other normal array.Eg. arr[0], arr[1,2].

Joining 2 np arrays

```
1 Eg.
2 import numpy as np
3 arr1 = np.array([[1, 2], [3, 4]])
4 arr2 = np.array([[5, 6], [7, 8]])
5 arr = np.concatenate((arr1, arr2), axis=1)
6 print(arr)
```

np.sort(arr) - To sort the numpy array.

To learn more:

NumPy Tutorial ([w3schools.com](https://www.w3schools.com/python/numpy/))

NumPy documentation — NumPy v1.24 Manual

1.2.3 Matplotlib

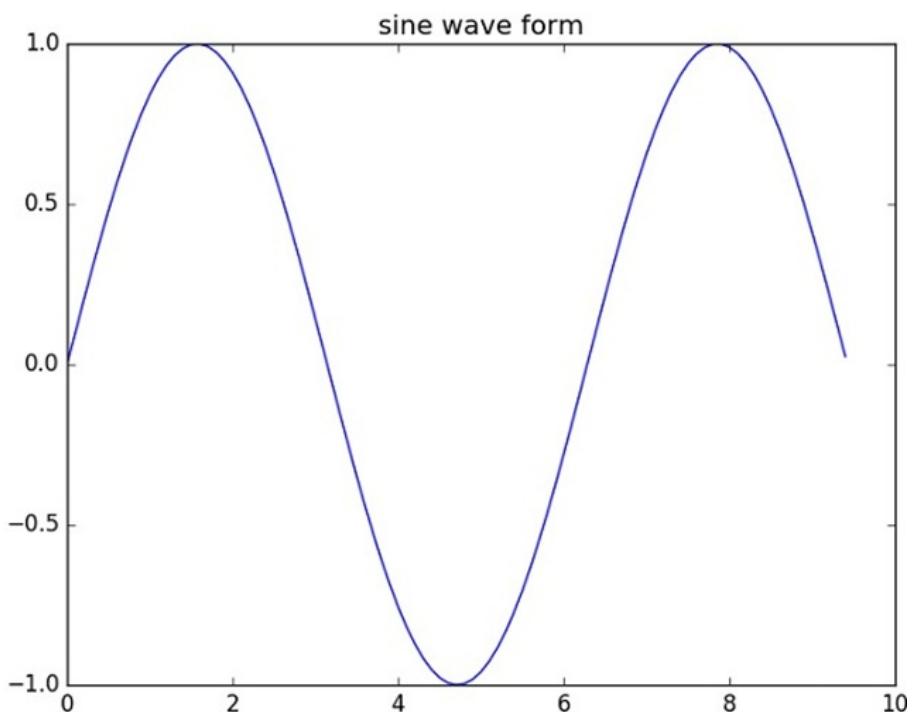
Matplotlib is a graph plotting library in python that serves as a visualization utility. Most of the Matplotlib utilities lies under the pyplot submodule.

To install matplotlib just type ‘pip install matplotlib’ in the terminal (given that you already have python and pip installed in your system).

It supports a very wide variety of graphs and plots namely - histogram, bar charts, scatter plots, heat maps, power spectra, error charts etc.

To learn more about how to use color and markers in your graph refer to: Matplotlib Markers

```
1 Eg:  
2 import numpy as np  
3 import matplotlib.pyplot as plt  
4 # Compute the x and y coordinates for points on a sine curve  
5 x = np.arange(0, 3 * np.pi, 0.1) # arange(start,end,step) generates number  
       starting from      start      to      end      with a step value of      step  
6 y = np.sin(x)  
7 plt.title("sine wave form")  
8 # Plot the points using matplotlib  
9 plt.plot(x, y)  
10 plt.show()
```

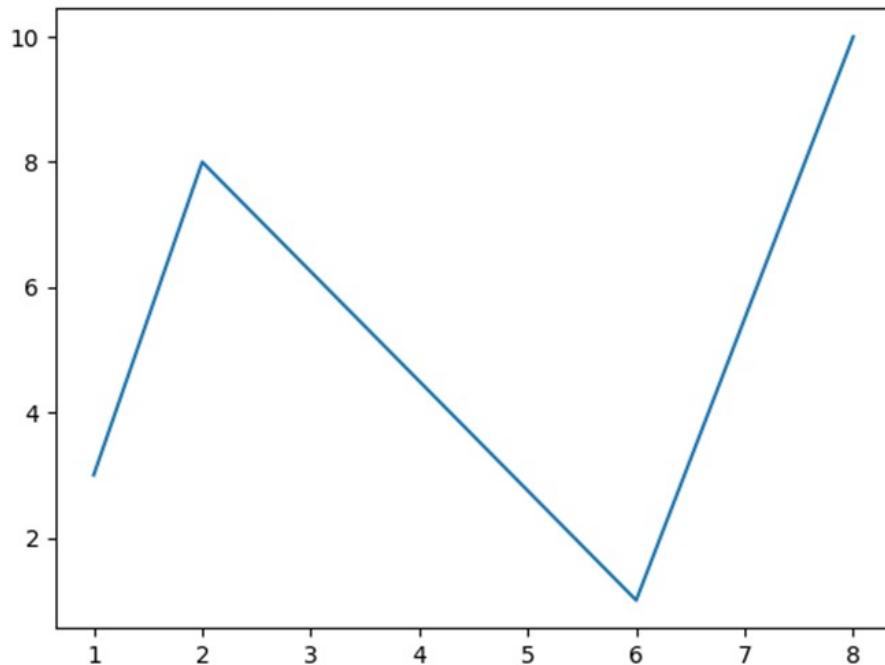


```
1 Eg:  
2 xpoints = np.array([1, 2, 6, 8])
```

```

3 ypoints = np.array([3, 8, 1, 10])
4 plt.plot(xpoints, ypoints)
5 plt.show()

```



To learn more :

free Code Camp Crash course on Matplotlib

Matplotlib — Visualization with Python

Matplotlib Pyplot (w3schools.com)

1.3 Practice Problems-1

A data file named ‘random_integer_data’ has been provided containing an array of 4000 random integers. Attempt the following questions to familiarize with the basic python syntax.

```

1 data = []      # creating a list for storing the data extracted from the txt file.
2 with open(r'random\_integer\_data.txt', 'r') as f:
3     for lines in f:

```

```
4     data.append(int(lines))    #storing the values in data array.
```

question 1 Write a Python code and Find the sum of :

- a) First 100 numbers
- b) Last 100 numbers
- c) First 2000 numbers
- d) Last 2000 numbers
- e) 'Sum' and 'mean' of numbers in chunks of 100 numbers

```
1 #A
2 sum_first_100 = sum(data[0:100])      # returns the sum of first 100 elements
3 print(sum_first_100)      # prints the sum of first 100 solutions
4
5 # OR
6
7 sum_first_100 = 0          # using bruteforce method
8 for i in range(100):
9     sum_first_100 = sum_first_100 + data[i]    # summing up each data value after each
10    iteration.
11 print(sum_first_100)
12 # %%
13 #B
14 length = len(data)          # returns the length of the list data.
15 sum_last_100 = sum(data[length-100:length])      # returns the sum of last 100
16    elements
17
18 sum_last_100 = 0          #brute force approach of adding each data element.
19 for i in range(100):
20     sum_last_100 = sum_last_100 + data[length-i-1]
21 print(sum_last_100)
22 # %%
23 # C
24 sum_first_2000 = sum(data[0:2000])      # summing up the first 2000 elements.
25 print(sum_first_2000)      #printing sum of first 2000 elements
26
27 sum_last_2000 = sum(data[length-2000:length])      # summing up the last 2000 elements
28 print(sum_last_2000)      # printing the sum of last 2000 elements.
29 # %%
30 #E
31 sum_100_chunks = [] #defining a list variable to store the sum of each chunk of
32    length 100(or bin of 100)
33 mean_100_chunks = [] #defining a list variable to store the mean of each chunk of
```

```

32             length 100(or bin of 100)
33 for i in range(length//100): # a loop which after iteration stores the sum of
34 100 element and mean of 100 elements and finally append it to the list defined above.
35     sum_100 = sum(data[i*100:(i+1)*100])
36     mean_100 = sum_100/100
37     sum_100_chunks.append(sum_100)
38     mean_100_chunks.append(mean_100)
39 print(sum_100_chunks) #printing the list
40 print(mean_100_chunks) #printing the list

```

question 2 Write a Python code to :

- a) Find the number of times number 89 is repeated
- b) Find the positions (can be called, line number in the file) where the number 89 is placed in the array. This would be a list numbers signifying positions.
- c) Find the number of times all the numbers are repeated.
- d) Find the number which is repeated the maximum number of times.

```

1 #A
2 count_89 = 0          # counts the number of times 89 is repeated.
3 position_89 = []    # a list which stores the position of 89
4 for i in range(length):    # a loop which iterates over
5     if data[i]==89:      #checks if the data list element is equal to 89
6         count_89 += 1    #iterates the count of 89
7         position_89.append(i)    #stores the position of 89
8     else:
9         continue      # if the value of data element not equal to 89 then skip this
10        iteration of loop
11 print(count_89)    #printing the count of 89
12 print(position_89)    # printing the positions of 89
13 #
14 #C
15 frequency = {}       # creating a dictionary named frequency to store the count of each
16 element.
17 for number in data:  #iterating over each number in data list
18     if (number in frequency):  # checking if the number is already stored in our
19         dictionary
20         frequency[number] += 1 #incrementing the frequency by 1
21     else:
22         frequency[number] = 1  #else creating a new dictionary element
23 print(frequency)    #printing the dictionary
24 for key , value in frequency.items():

```

```

23     print(" , key , , value , )")
24 print("the no. which is repeated the most is:",max(frequency,key=frequency.get),"it is
      repeated:",
25 frequency.get(25,"times") #printing the maximum frequency element and the frequency
      of the element

```

question 3 Write a Python code to find :

- a) Largest and smallest number amongst the first 100 numbers.
- b) Largest and smallest number amongst 101st position to 201st position.
- c) Largest and smallest number in chunks(orderwise viz. 1-100, 101-200, 201-300, so on and so forth...) of 100 numbers taken at a time.
- d) Find the mean of the two numbers(largest and smallest) in each chunk obtained in the part c).

```

1 print("largest in first 100 numbers",max(data[0:100]))    #printing the largest element
      among the
2 first 100 numbers
3 print("smallest in first 100 numbers",min(data[0:100]))  # printing the smallest
      number among the
4 first 100 numbers.
5 print("largest in 101th and 202th numbers",max(data[101:202])) #printing maximum
      number among
6 the elements b/w 101-202
7 print("smallest in 101th and 202th numbers",min(data[101:202])) #printing minimum
      number among
8 the elements b/w 101-202
9 # %%
10 max_list_100 = []          #stores max among each bin of 100 numbers.
11 min_list_100 = []          #stores min among each bin of 100 numbers.
12 mean_list_100 = []         #stores mean among each bin of 100 numbers.
13 for i in range(length//100):
14     max_100 = max(data[i*100:(i+1)*100])
15     min_100 = min(data[i*100:(i+1)*100])
16     max_list_100.append(max_100)
17     min_list_100.append(min_100)
18     mean_list_100.append((max_100+min_100)/2)
19
20 #printing max, min, and mean of each bin of 100 numbers.
21 for i in range(length//100):
22     print("the largest value in",i*100," , ",(i+1)*100,"is",max_list_100[i])
23     print("the smallest value in",i*100," , ",(i+1)*100,"is",min_list_100[i])

```

```
24     print("the mean of largest and smallest value in", i*100, ", ", (i+1)*100, "is",
mean_list_100[i])
```

question 4 Write a Python code and Find :

- a) Number of odd integers.
- b) Number of even integers.
- c) the list of odd integers.
- d) the list of even integers.
- e) Find the average of the odd integers and the even integers and compare both.

```
1 odd_list = []           #list stores the odd integers
2 even_list = []          #list stores the even integers
3 for i in range(length): #loop which goes through each element in data list
4     num = data[i]
5     if num%2 ==0 :        #checks if the num is even
6         even_list.append(num)
7     else :                # if odd append the number in odd_list
8         odd_list.append(num)
9 print("number of odd integers =",len(odd_list))    #prints the number of odd integers
10 print("number of even integers =",len(even_list))   #prints the number of even
     integers
11 print("average of odd integers =",sum(odd_list)/len(odd_list)) #prints the
12 average/mean of odd integers
13 print("average of even integers =",sum(even_list)/len(even_list)) #prints the
     average/mean of even integers
14
```

Chapter 2

Set Theory

2.1 Set

A set is a collection of distinct elements or objects or items. The elements, items or objects in a set can be anything: numbers, people, animals, books, or other sets. The idea of a set is a fundamental concept in mathematics used in various fields such as computer science, engineering, and physics.

Sets can be defined in various ways. One way to define a set is to simply list its elements between curly brackets, separated by commas. For example, we could define a set of fruits as:

{apple, banana, orange, mango, kiwi}

Another way to define a set is to use set-builder notation. Set-builder notation involves specifying a rule or a condition for membership in the set. For example, we could define a set of even numbers using set-builder notation as:

{ $x|x$ is an even number}

This would read as the set of all x such that x is an even number.

2.1.1 Definitions

2.1.1.1 Sample Space

In probability theory, the sample space is the set of all possible outcomes of a random experiment.

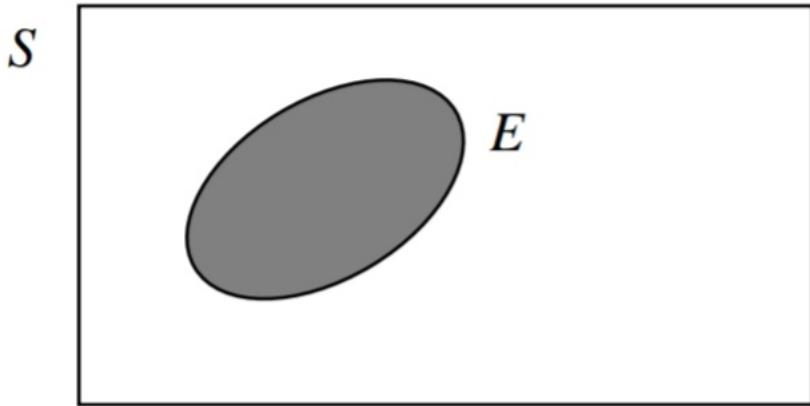
Example: Tossing a coin, $S = \{\text{Head}, \text{Tail}\}$

Rolling a die, $S = \{1,2,3,4,5,6\}$

2.1.1.2 Event

An event is a subset of the sample space, which represents a particular outcome or a collection of outcomes that we are interested in observing or analyzing.

Example: Getting an even number in rolling a die, $E = \{2,4,6\}$



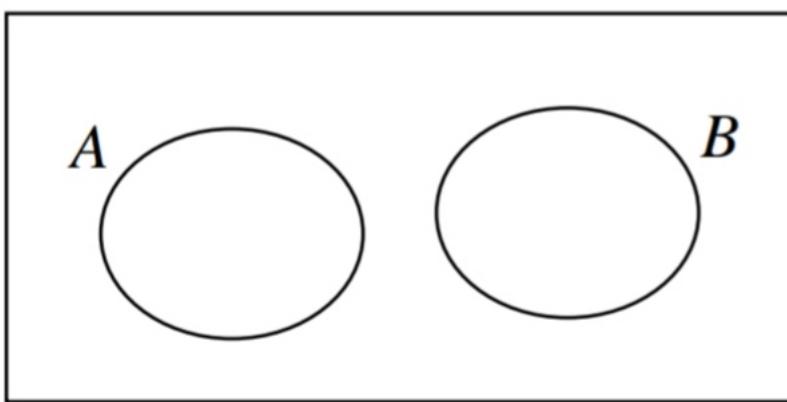
2.1.1.3 Mutually Exclusive Event

In probability theory, two events A and B are said to be mutually exclusive if they cannot occur at the same time. In other words, if event A occurs, then event B cannot occur, and vice versa. Mathematically, this can be expressed as:

$$P(A \cap B) = 0$$

where $P(A \cap B)$ represents the probability of both events A and B occurring simultaneously. If $P(A \cap B) = 0$, it means that the intersection of A and B is an empty set, or in other words, they have no elements in common.

In simpler terms, mutually exclusive events are events that have no outcomes in common. For example, when flipping a coin, the events of getting heads and getting tails are mutually exclusive since it is impossible to get both outcomes at the same time.



2.1.1.4 Empty set

The empty set is a set containing no objects. It is written as a pair of curly braces with nothing inside {} or by using the symbol ϕ .

As the set of all humans that high at least twelve foot, for example, is the empty set. Sets whose definition contains a contradiction or impossibility are often empty.

2.1.1.5 Set Membership Symbol

The set membership symbol \in is used to say that an object is a member of a set. It has a partner symbol \notin which is used to say an object is not in a set.

For ex, $S = \{1, 2, 3\}$ then $3 \in S$ and $4 \notin S$.

The set membership symbol is often used in defining operations that manipulate sets. The set $T = 2, 3, 1$ is equal to S because they have the same members: 1, 2, and 3. (Note: We say two sets are equal if they have exactly the same members.)

2.1.1.6 Cardinality of a set

For a finite set, the total number of elements present in a set is the cardinality. Let S be the set then, the cardinality will be denoted by $|S|$.

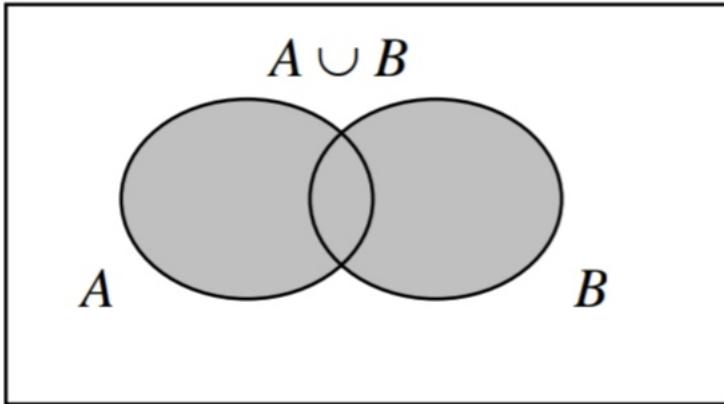
For ex, $S = \{1, 2, 3\}$, then, $|S| = 3$.

2.1.2 Operations

2.1.2.1 Union

The union of two sets S and T is the collection of all objects that are in either set. The \cup symbol is used to denote the union of two sets.

Syntax: $A \cup B$

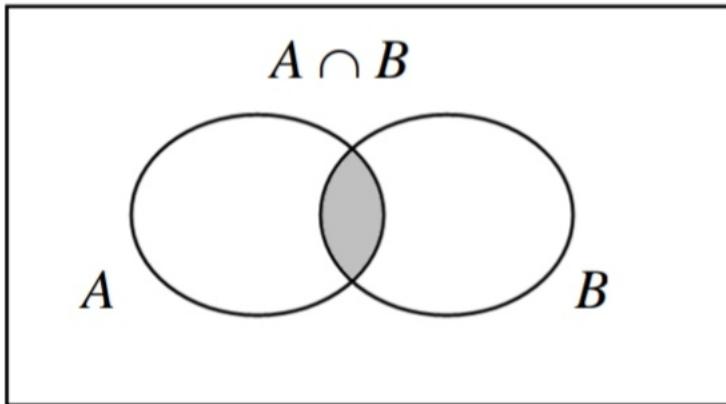


Example: If A is the set $\{1, 2, 3\}$ and B is the set $\{3, 4, 5\}$, then the union of A and B is:

$$A = \{1, 2, 3\} \quad B = \{3, 4, 5\} \quad A \cup B = \{1, 2, 3, 4, 5\}$$

2.1.2.2 Intersection

The intersection of two sets S and T is the collection of all objects that are in both sets. The \cap symbol is used to denote the intersection of two sets. Syntax: $A \cap B$

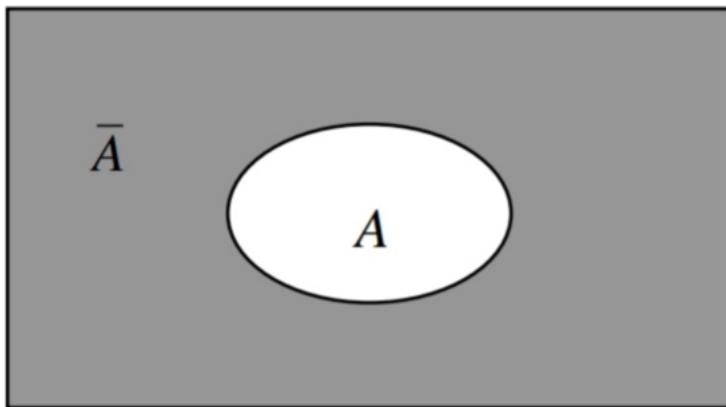


Example: If A is the set $\{1, 2, 3\}$ and B is the set $\{3, 4, 5\}$, then the intersection of A and B is:
 $A = \{1, 2, 3\}$ $B = \{3, 4, 5\}$ $A \cap B = \{3\}$

2.1.2.3 Complement

The compliment of a set S is the collection of objects in the universal set that are not in S . The ' $'$ symbol is used to denote the complement of a set.

Syntax: A'



Example: If A is the set $\{1, 2, 3\}$, then the complement of A is:

$A = \{1, 2, 3\}$ $A' = \{x | x \text{ is not an element of } A\} = \{x | x \neq 1 \text{ and } x \neq 2 \text{ and } x \neq 3\}$

2.1.2.4 Set difference

The difference of two sets S and T is the collection of objects in S that are not in T . The difference is written $S - T$.

Example: If A is the set {1, 2, 3, 4, 5, 6} and B is the set {1, 2, 7}

$$A - B = \{4, 5, 6\}$$

2.1.3 Demorgan's Law

Let A and B be events in a sample space S. Then,

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

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

Proof:

Let x be an arbitrary element in S. Then by definition,

$$x \in (A \cup B)^c \text{ if and only if } x \notin A \cup B.$$

By the distributive law of set theory,

$$x \notin A \cup B \text{ if and only if } x \notin A \text{ and } x \notin B.$$

Therefore,

$$x \in (A \cup B)^c \text{ if and only if } x \in A^c \text{ and } x \in B^c.$$

Since x was arbitrary, this holds for all elements in S. Hence,

$$(A \cup B)^c = A^c \cap B^c.$$

Proof:

Let x be an arbitrary element in S. Then by definition,

$$x \in (A \cap B)^c \text{ if and only if } x \notin A \cap B.$$

By the distributive law of set theory,

$$x \notin A \cap B \text{ if and only if } x \notin A \text{ or } x \notin B.$$

Therefore,

$$x \in (A \cap B)^c \text{ if and only if } x \in A^c \text{ or } x \in B^c.$$

Since x was arbitrary, this holds for all elements in S. Hence,

$$(A \cap B)^c = A^c \cup B^c.$$

Chapter 3

Introduction to Probability

3.1 Short History

Galileo (1564-1642), an Italian mathematician, was the first to attempt at a quantitative measure of probability while dealing with some problems related to the theory of dice in gambling. But the first foundation of the mathematical theory of probability was laid in the mid-seventeenth century by two French mathematicians, B. Pascal (1623-62) and P. Fermat (1601-65), while solving a number of problems posed by French gambler and noble man Chevalier-De-Mere to Pascal. Two mathematicians, after a lengthy correspondence between themselves, ultimately solved this problem and this correspondence laid the first foundation of the science of probability. The next famous person in this field was J. Bernoulli (1654-1705) whose 'Treatise on Probability' was published by his nephew N. Bernoulli in 1713. De-Moivre (1667-1754) also did considerable work in this field and published his famous 'Doctrine of Chances' in 1718. Russian mathematicians also have made very valuable contributions to the modern theory of probability. Chebychev (1821-94) who founded the Russian school of Statisticians; A Markoff (1856-1922); Liapounoff (central limit theorem); A. Khintchine (law of large numbers) and A. Kolmogorov, who gave axioms the calculus of probability.

3.2 Basic Terminology

3.2.1 Random Experiment

If in each trial of an experiment conducted under identical conditions, the outcome is not unique, but may be any one of the possible outcomes, then such an experiment is called a random experiment. Examples of a random experiment are tossing a coin, selecting a card from a pack, throwing a die, etc. The result of a random experiment is called the outcome. And any particular performance of a random experiment is called the trial and the outcome or combination of outcomes are termed an

event. For example, tossing a coin is a random experiment because we are not aware of the exact outcome but we know the possible outcomes. So tossing a coin one time is a trial of a random experiment and getting head or tail is an event.

As we know that an event is an outcome or a set of outcomes of a random experiment. One more example, tossing a coin three times event A = getting exactly two heads = {HTH, HHT, THH}. Similarly, tossing a fair dice event A = result is an even number = {2, 4, 6}.

3.2.2 Sample Space

The sample space S of a random process is the set of all possible outcomes. For example, one coin toss $S = \{H, T\}$. And three coin tosses $S = \{HHH, HTH, HHT, TTT, HTT, THT, TTH, THH\}$. Same way, roll a six-sided dice $S = \{1, 2, 3, 4, 5, 6\}$. The total number of possible outcomes of a random experiment are known as exhaustive events. For example, tossing a coin there are two exhaustive events head and tail. Throwing a six face die there are 6 exhaustive events.

The idea of probability comes in the context of carrying out a random experiment.

3.2.3 Deterministic Phenomena

There exists a mathematical model that allows “perfect” prediction of the phenomena outcome. Many examples exist in Physics, Chemistry (the exact sciences).

3.2.4 Non-deterministic Phenomena

No mathematical model exists that allows “perfect” prediction of the phenomena’s outcome. In general, the probability of an event is given by the ratio of the number of ways it can happen divided by the total number of outcomes. Non-deterministic Phenomena further may be divided into two groups.

1. Random phenomena: Unable to predict the outcomes, but in the long run, the outcomes exhibit statistical regularity.

2. Haphazard phenomena: unpredictable outcomes, but no long-run, the exhibition of statistical regularity in the outcomes.

For instance, tossing a coin the set of possible outcomes is $S = \text{Head, Tail}$. We are unable to predict on each toss whether is Head or Tail. But in the long run, can predict that 50% of the time heads will occur and 50% of the time tails will occur. Probability can be defined by classical and relative frequency methods. In the classical method, assigning probabilities based on the assumption of equally likely outcomes while in the relative frequency method, assigning probabilities based on experimentation or historical data.

In the classical method, if an experiment has n possible outcomes, this method would assign a probability of $1/n$ to each outcome. For instance, rolling a die the sample space S is 1, 2, 3, 4, 5, 6 and probabilities each sample point has a $1/6$ chance of occurring.

On the other hand, in the relative frequency method, Lucas Tool Rental would like to assign probabilities to the number of car polishers it rents each day. Office records show the following frequencies of daily rentals for the last 40 days.

Number of polishers rented	Number of days
0	4
1	6
2	18
3	10
4	2

Each probability assignment is given by dividing the frequency (number of days) by the total frequency (total number of days).

Number of polishers rented	Number of days	Probabilities
0	4	$4/40 = .10$
1	6	.15
2	18	.45
3	10	.25
4	2	.05
	total = 40	1

Example: When rolling a die the set of possible outcomes is $S = \{1, 2, 3, 4, 5, 6\}$. We are unable to predict the outcome but in the long run can one can determine that each outcome will occur $1/6$ of the time.

Example: What are the chances of rolling a "4" with a die?



The number of ways it can happen is 1 (there is only 1 face with a "4" on it) and the total number of outcomes is 6 (there are 6 faces altogether). So the probability = $1/6$.

Example: There are 5 marbles in a bag 4 are blue, and 1 is red. What is the probability that a blue marble will be picked?



The number of ways it can happen is 4 (there are 4 blues) and the total number of outcomes is 5 (there are 5 marbles in total). So the probability = 4/5.

Probability does not tell us exactly what will happen, it is just a guide. For example, toss a coin 100 times, how many Heads will come up? Probability says that heads have a 1/2 chance, so we would expect 50 Heads. But when you actually try it out you might get 48 heads, or 55 heads ... or anything really, but in most cases, it will be a number near 50. Similarly, before planning for an outing or a picnic, we always check the weather forecast. Suppose it says that there is a 60% chance that rain may occur. Do you ever wonder where this 60% comes from? Meteorologists use a specific tool and technique to predict the weather forecast. They look at all the other historical databases of the day, which have similar characteristics of temperature, humidity, and pressure, etc. And determine that on 60 out of 100 similar days in the past, it had rained.

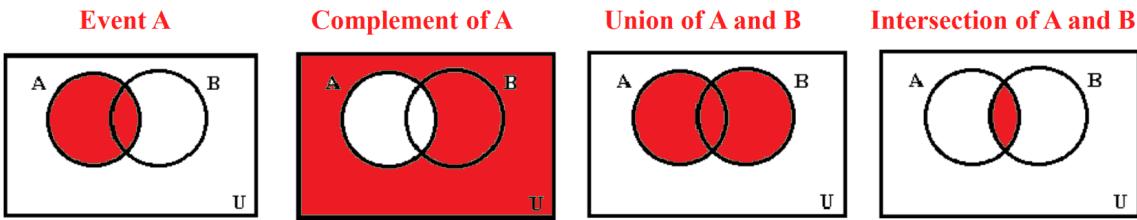
3.2.5 Mutually exclusive events

Events are said to be mutually exclusive if happening of one precludes the happening of other i.e. if two events cannot happen simultaneously. Mathematically, two events A and B are mutually exclusive if $A \cap B = \phi$. For example, tossing a coin the head and tail are mutually exclusive. In climate science, mutually exclusive events are events that cannot occur at the same time or under the same conditions. For example, in the context of extreme weather events, it is impossible for a location to experience both a severe drought and a flood at the same time. Similarly, certain atmospheric conditions that lead to hurricanes may preclude the formation of tornadoes. By identifying and analyzing mutually exclusive events, climate scientists can gain insights into the underlying mechanisms that drive climate patterns and variability. This information can help improve climate models and our ability to predict extreme weather events, which can have significant impacts on human populations and the environment.

3.2.6 Combinations of Events

The complement i.e. A^c of an event A is the event that A does not occur. The union of two events A and B is the event that either A or B or both occurs. The intersection of two events A and B is

the event that both A and B occur.



3.2.7 Disjoint Events

Two events are called disjoint if they can not happen at the same time. Events A and B are disjoint means that the intersection of A and B is zero (mutually exclusive). For example, the coin is tossed twice then $S = \{\text{HH}, \text{TH}, \text{HT}, \text{TT}\}$. Now define, events $A = \{\text{HH}\}$ and $B = \{\text{TT}\}$ are disjoint. And events $A = \{\text{HH, HT}\}$ and $B = \{\text{HH}\}$ are not disjoint. These events are unlikely to happen at the same time and are considered to be disjoint events. For instance, in a given location, a record high temperature of 110°F and a record low temperature of 10°F are unlikely to occur simultaneously. Therefore, the events of extreme heat and extreme cold are considered disjoint events.

Now let A be any event of sample space S. Then

1. $0 \leq P(A) \leq 1$ for any event A,
2. The probability of the whole sample space is 1 $P(S) = 1$,
3. $P(A^c) = 1 - P(A)$
4. If A and B are disjoint events then $P(A \text{ or } B) = P(A) + P(B)$.

Example: Saskatoon and Moncton are two of the cities competing in the World university games. (There are also many others). The organizers are narrowing the competition to the final 5 cities. There is a 20% chance that Saskatoon will be among the final 5. There is a 35% chance that Moncton will be amongst the final 5 and an 8% chance that both Saskatoon and Moncton will be among the final 5. What is the probability that Saskatoon or Moncton will be among the final 5?

Solution: Let A = the event that Saskatoon is amongst the final 5. Let B = the event that Moncton is amongst the final 5. Given $P[A] = 0.20$, $P[B] = 0.35$, and $P[A \cap B] = 0.08$ What is $P[A \cup B]$?

$$P[A \cup B] = P[A] + P[B] - P[A \cap B] = 0.20 + 0.35 - 0.08 = 0.47.$$

3.2.8 Conditional Probability

As discussed in the class, the probability $P(A)$ of an event A represents the likelihood that a random experiment will result in an outcome in set A relative to the sample space S of the random experiment. Mathematically, let A and B be two events in the sample space. The conditional probability that

event B occurs given that event A has occurred is:

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

Example: Given the imperfect diagnostic test for a disease

	Disease+	Disease-	total
Test+	30	10	40
Test-	10	50	60
total	40	60	100

What is the probability that a person has the disease given that they tested positive?

Solution:

$$P(\text{disease}+ | \text{test}+) = \frac{P(\text{disease}+ \text{ and test}+)}{P(\text{test}+)} = \frac{30/100}{40/100} = 0.75.$$

Now we discuss the conditional probability axioms

1. $P(A|B) = P(A, B)/P(B) \geq 0.$
2. If S is sample space and B is an event $P(S|B) = P(S, B)/P(B) = P(B)/P(B) = 1.$
3. Suppose A and B are two disjoint events $P((A + B)|M) = P((A + B)M)/P(M) = P(AM + BM)/P(M) = P(AM)/P(M) + P(BM)/P(M) = P(A|M) + P(B|M).$

Example: Suppose we are performing a fair die experiment. It has got six faces (f1,f2,f3,f4...). f1 is a face marked as 1. If we are given an event A={f2}=1/6. Consider another event M={even}={f2,f4,f6} subject to the fact M has occurred, what is the probability of A?

$$P(A|M) = P(A, M)/P(M) = P(A)/P(M) = 1/3.$$

Example: Suppose we have a set of temperature data for a particular region that ranges from 1 to 10 degrees Celsius. If we randomly select a day from this data, we are told that the temperature on that day is at least five, then what is the probability that it is ten?

Solution: Let E denote the event that the selected temperature is ten degree Celsius and let F be the event that it is at least five. The desired probability is $P(E|F).$

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

The probability of getting a temperature of 10 degrees Celsius and at least 5 is 1/10 (since there is only one temperature observation with 10 degrees Celsius and six observations with a number greater than or equal to 5 degrees Celsius). The probability of getting a temperature that is at least 5 degrees Celsius is 6/10 (since there are six observations with temperatures greater than or equal to 5 out of

ten total temperature observations).

$$P(E|F) = \frac{1/10}{6/10} = 1/6.$$

3.2.9 Independence

Several events are said to be independent if the happening of an event is not affected by the supplementary knowledge concerning the occurrence of any number of remaining events. If A and B are independent, then $P(A \cap B) = P(A) \times P(B)$ which means that conditional probability is:

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

We have a more general multiplication rule for events that are not independent

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

Example: Suppose we toss two fair dice. Let E denote the event that the sum of the dice is six and F denote the event that the first die equals four. Prove E and F are not independent.

Solution: $P(E) = \text{number of ways to get a sum of six} / \text{total possible outcomes} = 5/36$ (There are five ways to get a sum of six: (1,5), (2,4), (3,3), (4,2), and (5,1), and there are 36 possible outcomes.)

$P(F) = \text{number of ways to get the first die to be four} / \text{total possible outcomes} = 1/6$ ((4,1), (4,2), (4,3), (4,4), (4,5), or (4,6), and there are 36 possible outcomes.)

Now we need to calculate the probability of the intersection of events E and F:

$P(E \cap F) = \text{number of ways to get a sum of six when the first die is four} / \text{total possible outcomes}$
 $= 1/36$ (There is only one way to get a sum of six when the first die is four: (4,2), and there are 36 possible outcomes.) and $P(E)P(F) = \frac{5}{36} \cdot \frac{1}{6}$. Hence, E and F are not independent.

3.3 Correlation

Correlation is a statistical measure that shows the strength and direction of the relationship between two variables. It measures the degree to which two variables are related to each other. The correlation coefficient is a value that ranges from -1 to +1, where -1 indicates a perfect negative correlation, +1 indicates a perfect positive correlation, and 0 indicates no correlation between the variables. A positive correlation indicates that as one variable increases, the other variable also increases, while a negative correlation indicates that as one variable increases, the other variable decreases. We use the data of monthly maximum and minimum temperature from 1981-2021, which contains the 525 grids of NWH, to calculate the correlation at each grid between monthly maximum and minimum temperatures. Below code was used to complete the task. The Spatial plot of correlation between

monthly maximum and minimum temperature from 1981-2021 over NWH is given in Figure 3.1.

```
1 library(moments)
2 library(ggplot2)
3 library(rgdal) #loading the rgdal package
4 max = read.csv(file.choose()) #read the temp data over each grid
5 min = read.csv(file.choose())
6
7 corr = numeric(525)
8 for(i in 1:525){
9   corr[i] = cor(as.numeric(max[i,]), as.numeric(min[i,]))
10 }
11 #read dataset latlongNWH to find latitude and longitude
12 data = read.csv(file.choose())
13 #bind the lat long and calculated correlation in a single dataset
14 data=cbind(data,corr)
15
16 #now using the ggplot for grid wise data plot
17 obj = ggplot(data ,aes(x= long, y= lat, fill = corr)) +geom_tile() +theme_classic() +
18   theme(panel.border = element_rect(color = "black",fill = NA,size = 1)) +
19   xlab("Longitude") +ylab("Latitude") + labs(fill="Correlation") +
20   scale_fill_gradientn(breaks = seq(min(data$corr), max(data$corr), length.out = 10),
21   colors = rainbow(20)) + coord_fixed() +
22   theme(legend.key.height= unit(3.7, 'cm'), legend.key.width= unit(1, 'cm')); obj
23
24 shf = readOGR("C:\\\\Users\\\\Neeraj Poonia\\\\Desktop\\\\neeraj\\\\MA605\\\\NWH_temp_plot
25 \\\\NWH ","4-17-2018-899072")
26 obj.shf = geom_polygon(data=shf, aes(x= long, y= lat),
27                         colour ="black", fill="white", alpha=0)
28 obj + obj.shf
```

```
1 import pandas as pd
2 #long,latitude of laocation
3 location = pd.read_csv(r'New folder (2)/new_latlongNWH.csv')
4 location
5 maxtemp = pd.read_csv(r'New folder (2)/Temperature(2M)_Maximum_NWH_1981-2021 - Copy.
6   csv')
6 mintemp = pd.read_csv(r'New folder (2)/Temperature(2M)_Minimum_NWH_1981-2021 - Copy.
7   csv')
7 correlation = maxtemp.corrwith(mintemp, axis = 1)
8 correlation
9 import geopandas as gpd
10
11 shapefile = gpd.read_file(r"NWH/NWH/4-17-2018-899072.shp")
12 print(shapefile)
13 import matplotlib.pyplot as plt
14 fig, ax = plt.subplots()
```

```

15 shapefile.to_crs(epsg=4326).plot(ax=ax)
16 plt.show()
17 import matplotlib.pyplot as plt
18 shxfile = gpd.read_file(r"NWH/NWH/4-17-2018-899072.shx")
19 fig, ax = plt.subplots()
20 shxfile.plot(ax=ax)
21 plt.show()
22 shxfile
23 from shapely.geometry import Point, Polygon
24 crs = {'init': 'EPSG:4326'}
25 geometry = [Point(xy) for xy in zip(location['long'], location['lat'])]
26 geo_df = gpd.GeoDataFrame(location,
27                           crs = crs,
28                           geometry = geometry)
29 geo_df.head()
30
31
32 geo_df['correlation'] = correlation
33 geo_df
34 fig, ax = plt.subplots(figsize = (10,10))
35 shapefile.to_crs(epsg=4326).plot(ax=ax, color='lightgrey')
36 geo_df.plot(column = 'correlation', ax=ax,
37               legend = True,
38               markersize = 10)
39 ax.set_title('Kings County Price Heatmap')
40 plt.savefig('Heat Map')

```

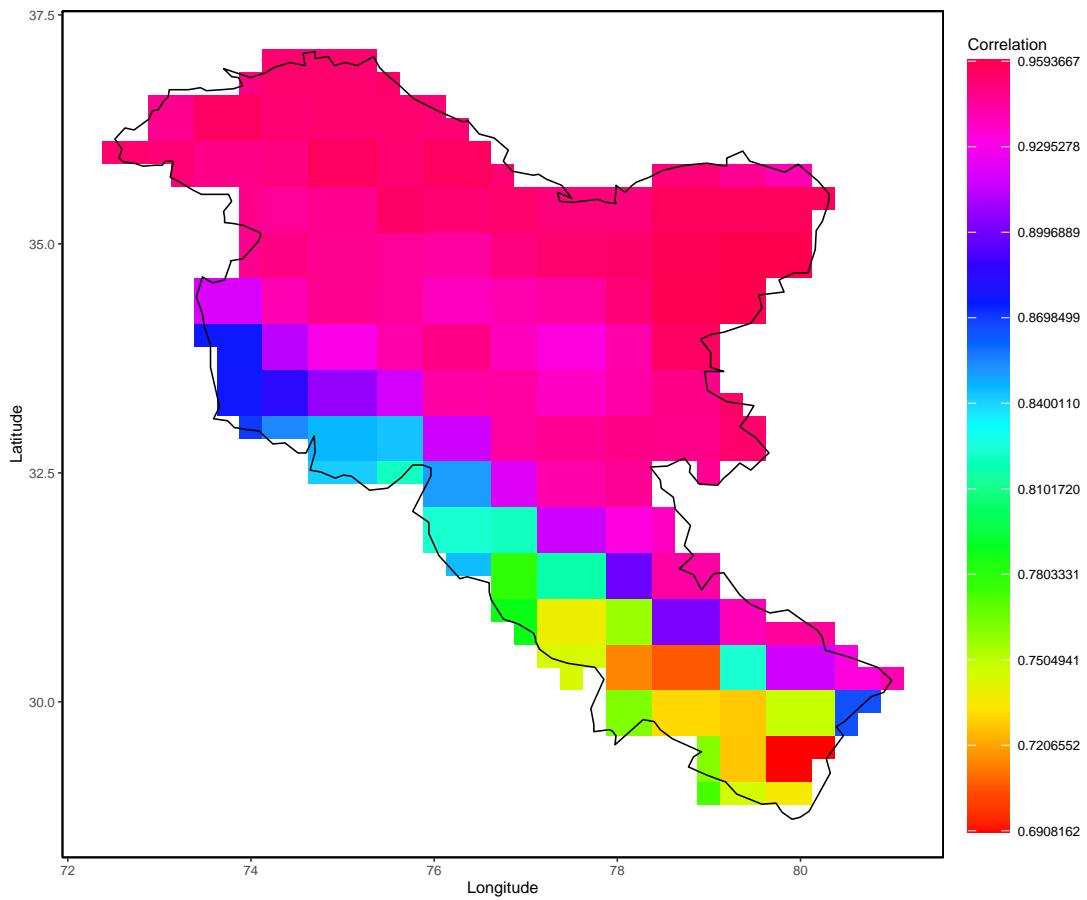


Fig. 3.1: Correlation between monthly maximum and minimum temperature from 1981-2021 over 525 grids of NWH.

From the Figure 3.1 it is clearly visible that as foothills of Himachal Pradesh and Uttrakhand have the low positive correlation in the region mean while the high elevated region of Ladakh has the highest positive correlation. As we are going from low to high elevation the correlation is getting increases.

Now we will convert this 41 year monthly maximum temperature data into annual maximum temperature data so that we can find the mean, standard deviation, and the best-fit probability distribution for each grid. The following code was used to find the mean temperature of annual maximum temperature from 1981 to 2021 and generated plot is given figure 3.2.

```

1 #plot of monthly maximum temperature from 1981-2020
2 library(moments)
3 library(ggplot2)
4 library(rgdal) #loading the rgdal package
5 max = read.csv(file.choose()) #read the temp data over each grid
6 library(data.table)
```

```

7 monthly.to.annual = function(tmax){ #this function will convert
8   df.tmax = data.frame()
9   for(i in 1:525){
10     x = numeric(41)
11     start = 1
12     end = 12
13     for(j in 1:41){
14       x[j] = max(as.numeric(tmax[i:i,start:end]))
15       start = start + 12
16       end = end + 12
17     }
18     df.tmax = rbind(df.tmax,x)
19   }
20   return(df.tmax = transpose(df.tmax))
21 }
22
23 max = monthly.to.annual(max) #read the temp data over each grid
24
25 mean.max = numeric(525)
26 for(i in 1:525){
27   mean.max[i] = mean(as.numeric(max[,i]))
28 }
29 #read dataset latlongNWH to find latitude and longitude
30 data = read.csv("C:\\\\Users\\\\Neeraj Poonia\\\\Desktop\\\\neeraj\\\\MA605\\\\NWH_temp_plot\\\\new_
     latlongNWH.csv")
31 #bind the lat long and calculated correlation in a single dataset
32 data=cbind(data,mean.max)
33
34 #now using the ggplot for grid wise data plot
35 obj = ggplot(data ,aes(x= long, y= lat, fill = mean.max)) +geom_tile() +theme_classic()
      +
36   theme(panel.border = element_rect(color = "black",fill = NA,size = 1)) +
37   xlab("Longitude") +ylab("Latitude") + labs(fill="Mean Temperature") +
38   scale_fill_gradientn(breaks = seq(min(data$mean.max), max(data$mean.max), length.out
      = 10),
      colors = rev(heat.colors(20))) + coord_fixed() +
39   theme(legend.key.height= unit(3.7, 'cm'),legend.key.width= unit(1, 'cm')); obj
40
41
42 shf = readOGR("C:\\\\Users\\\\Neeraj Poonia\\\\Desktop\\\\neeraj\\\\MA605\\\\NWH_temp_plot\\\\NWH ",
      "4-17-2018-899072")
43 obj.shf = geom_polygon(data=shf, aes(x= long, y= lat),
      colour ="black", fill="white", alpha=0)
44 obj + obj.shf

```

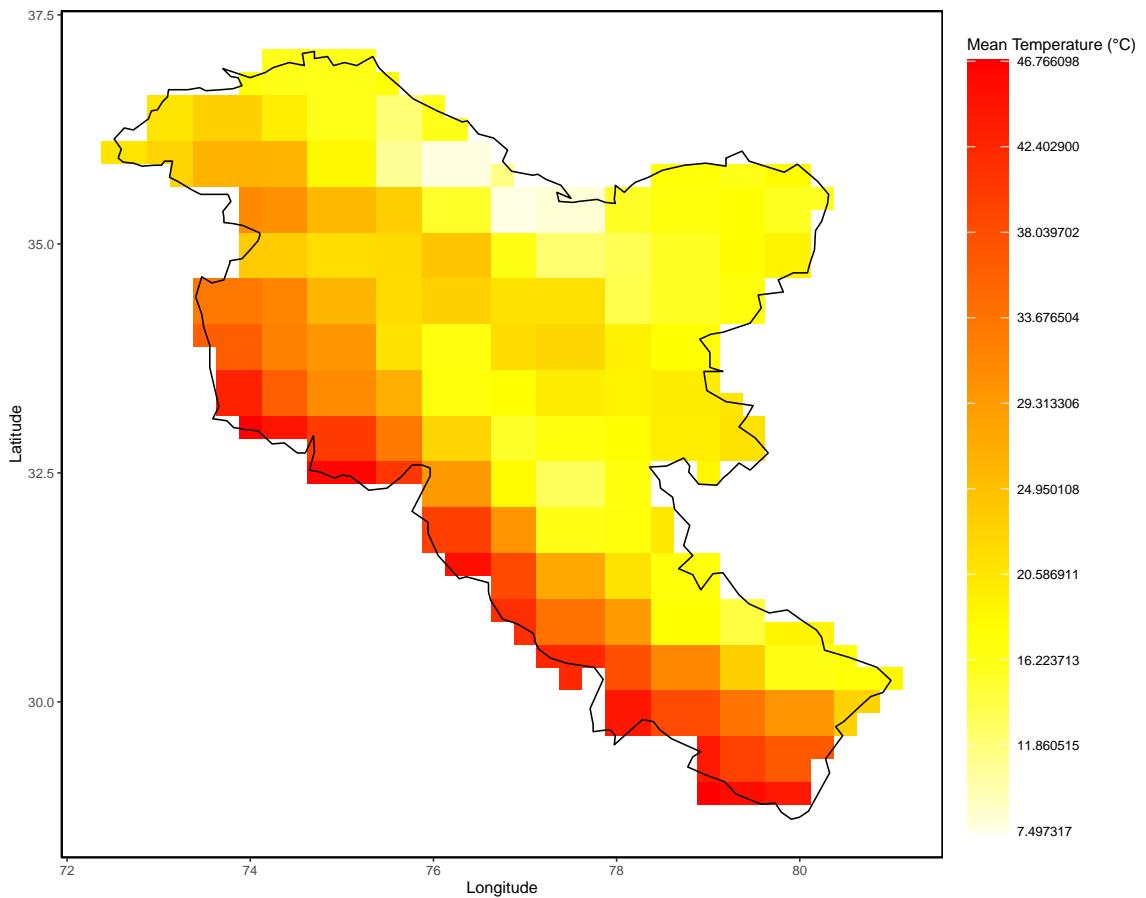


Fig. 3.2: Mean of annual maximum temperature from 1981-2021 over 525 grids of NWH.

From the above figure it is concluded that the low elevated area of NWH region has a high mean temperature and high elevated areas have a low mean temperature. Further, we plot the standard deviation at each grid using the below code and generated plot is given in figure 3.3.

```

1 rm(list=ls())
2 max = read.csv(file.choose())
3 library(data.table)
4 monthly.to.annual = function(tmax){ #this function will convert
5   df.tmax = data.frame()
6   for(i in 1:525){
7     x = numeric(41)
8     start = 1
9     end = 12
10    for(j in 1:41){
11      x[j] = max(as.numeric(tmax[i:i,start:end]))
12      start = start + 12
13      end = end + 12
14    }
15  }
16  df.tmax = as.data.table(df.tmax)
17  df.tmax[,lat] = rownames(df.tmax)
18  df.tmax[,lon] = colnames(df.tmax)
19  df.tmax[,year] = 1981:2021
20  df.tmax[,month] = 1:12
21  df.tmax[,tmax] = x
22  return(df.tmax)
23 }
24 tmax = monthly.to.annual(max)
25 
```

```

14 }
15 df.tmax = rbind(df.tmax,x)
16 }
17 return(df.tmax = transpose(df.tmax))
18 }
19
20
21 library(moments)
22 library(ggplot2)
23 library(rgdal) #loading the rgdal package
24 max = monthly.to.annual(max) #read the temp data over each grid
25
26 sd.max = numeric(525)
27 for(i in 1:525){
28   sd.max[i] = sd(as.numeric(max[,i]))
29 }
30 #read dataset latlongNWH to find latitude and longitude
31 data = read.csv("C:\\\\Users\\\\Neeraj Poonia\\\\Desktop\\\\neeraj\\\\MA605\\\\NWH_temp_plot\\\\new_
      latlongNWH.csv")
32 #bind the lat long and calculated correlation in a single dataset
33 data=cbind(data,sd.max)
34
35 #now using the ggplot for grid wise data plot
36 obj = ggplot(data ,aes(x= long, y= lat, fill = sd.max)) +geom_tile() +theme_classic() +
37   theme(panel.border = element_rect(color = "black",fill = NA,size = 1)) +
38   xlab("Longitude") +ylab("Latitude") + labs(fill="StDev") +
39   scale_fill_gradientn(breaks = seq(min(data$sd.max), max(data$sd.max), length.out =
      10),
                         colors = rainbow(20)) + coord_fixed() +
40   theme(legend.key.height= unit(3.7, 'cm'), legend.key.width= unit(1, 'cm')); obj
41
42
43 shf = readOGR("C:\\\\Users\\\\Neeraj Poonia\\\\Desktop\\\\neeraj\\\\MA605\\\\NWH_temp_plot\\\\NWH ",
      "4-17-2018-899072")
44 obj.shf = geom_polygon(data=shf, aes(x= long, y= lat),
                        colour ="black", fill="white", alpha=0)
45 obj + obj.shf

```

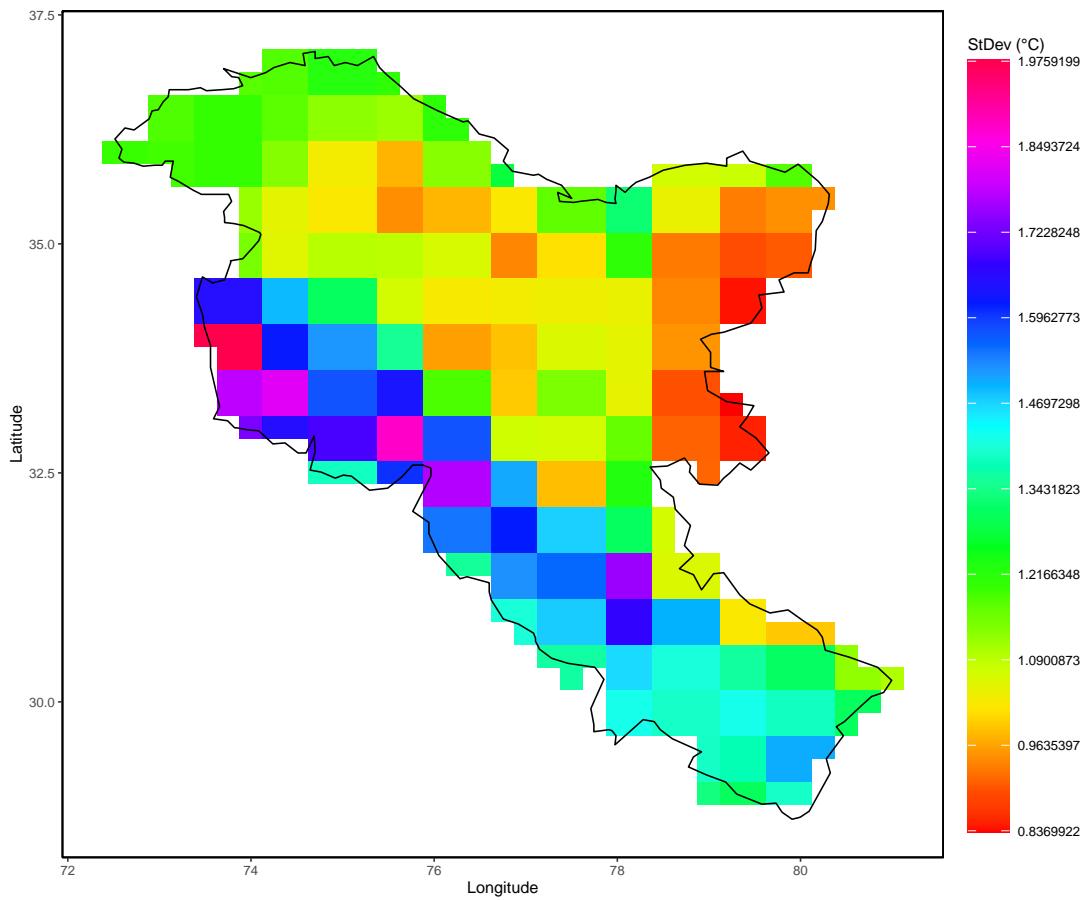


Fig. 3.3: Standard deviation of annual maximum temperature from 1981-2021 over 525 grids of NWH.

From figure 3.3 we observe that the Ladakh region has the smallest value of standard deviation which indicates this area has very less fluctuations in the annual maximum temperature data. Meanwhile, the southwest area of the NWH region has a high variability based on annual maximum temperature from 1981 to 2021. Further, the below code was used to find the best-fit probability distribution.

```

1 rm(list=ls())
2 library(data.table)
3 tmax = read.csv("C:\\\\Users\\\\Neeraj Poonia\\\\Desktop\\\\neeraj\\\\manuscripts\\\\copula_based_
    bivariate_model\\\\data\\\\Temperature (2M)_Maximum_NWH_1981-2021.csv")
4 df.tmax = data.frame()
5 for(i in 1:525){
6   x = numeric(41)
7   start = 1
8   end = 12
9   for(j in 1:41){
```

```

10   x[j] = max(as.numeric(tmax[i:i,start:end]))
11   start = start + 12
12   end = end + 12
13 }
14 df.tmax = rbind(df.tmax,x)
15 }
16 df.tmax = transpose(df.tmax)

```

The above code will convert the monthly to annual temperature. Now we have the annual maximum temperature at 525 grids of NWH region and five models, namely, Gamma, lognormal, Weibull, Gumbel, and normal distributions. Using the below code we will fit model to each of 525 grids.

```

1 library(fitdistrplus)
2 library(actuar)
3 data = read.csv("C:\\\\Users\\\\Neeraj Poonia\\\\Desktop\\\\neeraj\\\\manuscripts\\\\selection_of_
      best_fit_distribution\\\\data\\\\annual_max_temp_1901_2017_NWH_cru.csv")
4 data=df.tmax
5 dgumbel <- function(x, a, b) 1/b*exp((a-x)/b)*exp(-exp((a-x)/b)) #pdf
6 pgumbel <- function(q, a, b) exp(-exp((a-q)/b)) #cdf
7
8
9 aic.value = data.frame()
10 MLE.gamma = data.frame()
11 MLE.ln = data.frame()
12 MLE.wei = data.frame()
13 MLE.gum = data.frame()
14 MLE.norm = data.frame()
15 SE.value = data.frame()
16 for(i in 1:525){
17   print(i)
18   d = data[,i]
19   fg <- fitdist(d, "gamma")
20   fln <- fitdist(d, "lnorm")
21   fw <- fitdist(d, "weibull")
22   fgumbel = fitdist(d, "gumbel", start=list(a=10, b=10))
23   fnorm = fitdist(d,"norm")
24   MLE.gamma = rbind(MLE.gamma, fg$estimate)
25   MLE.ln = rbind(MLE.ln, fln$estimate)
26   MLE.wei = rbind(MLE.wei, fw$estimate)
27   MLE.gum = rbind(MLE.gum, fgumbel$estimate)
28   MLE.norm = rbind(MLE.norm, fnorm$estimate)
29   kk = gofstat(list(fg, fln, fw, fgumbel,fnorm),fitnames = c("Gamma", "lnorm", "wibull
      ", "Gumbel", "Norm"))[12]
30   aic.value = rbind(aic.value, as.numeric(unlist(kk)))
31 }
32

```

```

33 names(MLE.ln) = c("meanlog", "sdlog")
34 names(MLE.norm) = c("mean", "sd")
35 names(MLE.gamma) = c("shape", "rate")
36 names(MLE.wei) = c("shape", "scale")
37 names(aic.value) = c("Gamma", "Lognormal", "Weibull", "Gumbel", "Normal")
38
39latlong = read.csv("C:\\\\Users\\\\Neeraj Poonia\\\\Desktop\\\\neeraj\\\\MA605\\\\NWH_temp_plot\\\\
    new_latlongNWH.csv")
40
41 aic.value = data.frame(lat = latlong$lat, long = latlong$long, aic.value)
42
43 min.aic = numeric(525)
44 for(i in 1:525){
45   min.aic[i] = which(aic.value[i,3:7] == min(aic.value[i,3:7]))
46 }
47 min.aic
48 aic.value$model = min.aic

```

Lines 43 to 48 in the above code will assign a number from 1 to 5 to each grid based on the best distribution using the AIC value. The numbers 1,2,3,4, and 5 are representing the Gamma, log-normal, Weibull, Gumbel, and normal distributions respectively. Till now we have the best model corresponding to each grid. It is time to spatially plot this data. The just immediate code will help us in spatial plot. The plot is given in figure 3.4.

```

1 library(ggplot2)
2 library(rgdal)
3
4 obj.df = ggplot(data=aic.value, aes(x= long, y= lat, fill=as.factor(model))) + geom_
    tile() +
5   theme_classic() + theme(panel.border = element_rect(color = "black", fill = NA, size
      = 1) )+
6   guides(fill=guide_legend(title="Distribution\\n model")) + xlab("Longitude") +ylab("Latitude") +
7   coord_fixed() + scale_fill_manual(labels =c("Gamma", "lnorm", "wibull", "Gumbel", "Norm"),
      values=c("gold","gold1","gold2","gold3","gold4"))
8 plot(obj.df)
9
10 path = "C:\\\\Users\\\\Neeraj Poonia\\\\Desktop\\\\neeraj\\\\MA605\\\\NWH_temp_plot\\\\NWH"
11 shf = readOGR(path, "4-17-2018-899072")
12 obj.shf = geom_polygon(data=shf, aes(x= long, y= lat),
13                         colour ="black", fill="white", alpha=0)
14 obj.df+obj.shf

```

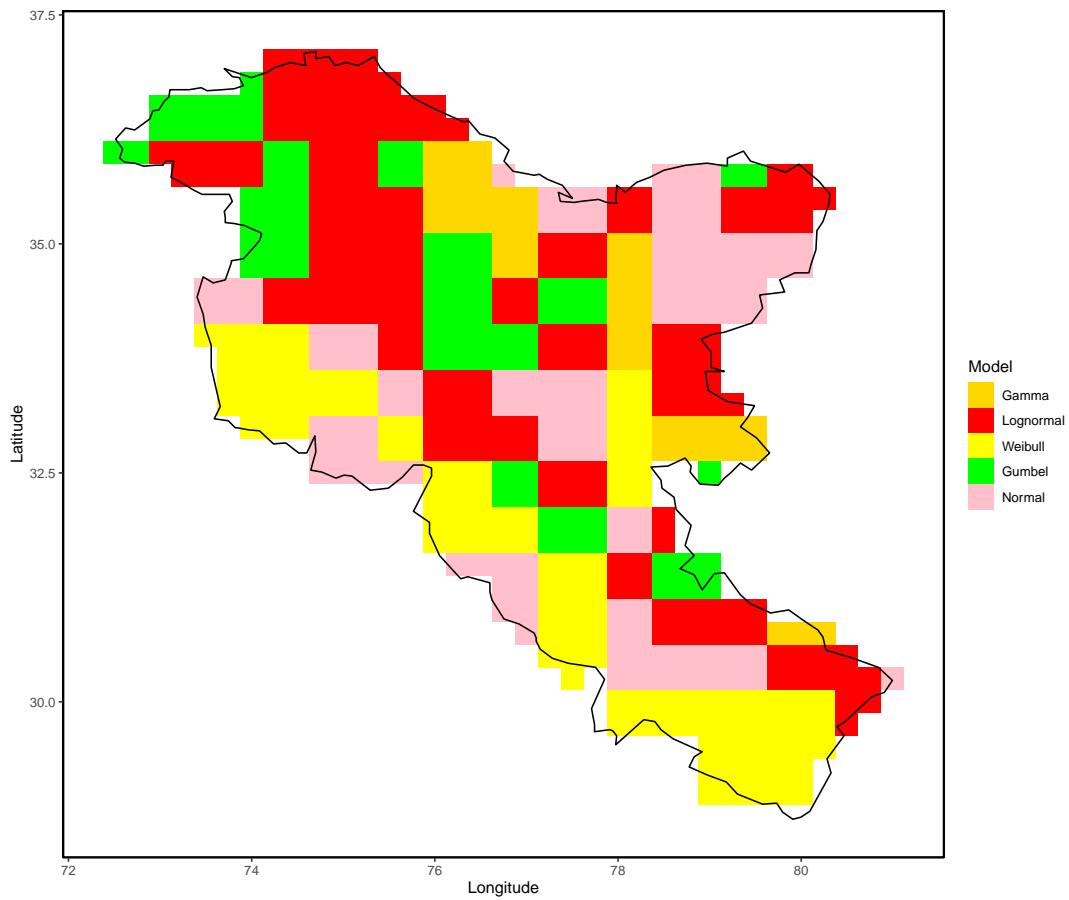


Fig. 3.4: The best-fitted distribution for maximum temperature from 1981-2021 over 525 grids of NWH.

```

1 from pyEOF import *
2 import xarray as xr
3 import pandas as pd
4 import numpy as np
5 import matplotlib.pyplot as plt
6 import geopandas as gpd
7 import statistics
8 from scipy import stats
9 from fitter import Fitter, get_common_distributions, get_distributions
10 df1 = pd.read_csv('E:\Downloads\\annual_max_temp_1901_2017_NWH_cru.csv')
11 df2 = df1.transpose()
12 latlon = pd.read_csv('E:\Downloads\\new_latlongNWH.csv')
13 latlon.head(3)
14 a=[]
15 for i in range(525):
16     ff = Fitter(df2.iloc[i], distributions = ['gamma', 'norm', 'lognorm', 'gumbel_r', 'weibull'])

```

```

    weibull_min'])
17 ff.fit()
18 ff.summary()
19 if (min(ff.summary()['aic'])==ff.summary()['aic']['gamma']):
20     a.append('Gamma')
21 elif (min(ff.summary()['aic'])==ff.summary()['aic']['lognorm']):
22     a.append('Lognormal')
23 elif (min(ff.summary()['aic'])==ff.summary()['aic']['weibull_min']):
24     a.append('Weibull')
25 elif (min(ff.summary()['aic'])==ff.summary()['aic']['gumbel_r']):
26     a.append('Gumbel')
27 elif (min(ff.summary()['aic'])==ff.summary()['aic']['norm']):
28     a.append('Normal')
29
30
31 gdf1 = gpd.read_file('E:\\PycharmProjects\\4-17-2018-899072.shp')
32 dft=pd.DataFrame(list(zip(latlon['lat'],latlon['long'],a)),columns=['Latitude','
   Longitude','distribution'])
33 gdft = gpd.GeoDataFrame(dft, geometry=gpd.points_from_xy(dft.Longitude, dft.Latitude))
34 ax=gdft.plot("distribution",legend=True)
35 fig = ax.figure
36 cb_ax.tick_params(labelsize=20)
37 gdf1.plot(ax=ax,linewidth=1.5,color='black')

```

3.4 Return level estimation

The primary goal of statistical modeling is to estimate the future return levels for a given return period. Here we have obtained the best-fitted distribution for each grid which also provides us with the best model for a grid and corresponding the estimated parameters of that model. We can combine the return period with model and its parameters to obtain the return level. The following code was used to obtain the return levels at 100 years return period for each grid and results are given in figure 3.5. We used the data of 525 annual maximum temperature times series from 1981 to 2021. And the return level obtained is for the next 100 years from 2021 which is for 2121.

```

1
2 rm(list=ls())
3 library(data.table)
4 tmax =read.csv("C:\\Users\\Neeraj Poonia\\Desktop\\neeraj\\manuscripts\\copula_based_
   bivariate_model\\data\\Temperature(2M)_Maximum_NWH_1981-2021.csv")
5 df.tmax = data.frame()
6 for(i in 1:525){
7     x = numeric(41)
8     start = 1
9     end = 12

```

```

10  for(j in 1:41){
11    x[j] = max(as.numeric(tmax[i:i,start:end]))
12    start = start + 12
13    end = end + 12
14  }
15  df.tmax = rbind(df.tmax,x)
16 }
17 df.tmax = transpose(df.tmax)
18
19 library(fitdistrplus)
20 library(actuar)
21 library(lmomco)
22 data=df.tmax
23 dgumbel <- function(x, a, b) 1/b*exp((a-x)/b)*exp(-exp((a-x)/b)) #pdf
24 pgumbel <- function(q, a, b) exp(-exp((a-q)/b)) #cdf
25
26
27 aic.value = data.frame()
28 MLE.gamma = data.frame()
29 MLE.ln = data.frame()
30 MLE.wei = data.frame()
31 MLE.gum = data.frame()
32 MLE.norm = data.frame()
33 SE.value = data.frame()
34 for(i in 1:525){
35   print(i)
36   d = data[,i]
37   fg <- fitdist(d, "gamma")
38   fln <- fitdist(d, "lnorm")
39   fw <- fitdist(d, "weibull")
40   fgumbel = fitdist(d, "gumbel", start=list(a=10, b=10))
41   fnorm = fitdist(d,"norm")
42   MLE.gamma = rbind(MLE.gamma, fg$estimate)
43   MLE.ln = rbind(MLE.ln, fln$estimate)
44   MLE.wei = rbind(MLE.wei, fw$estimate)
45   MLE.gum = rbind(MLE.gum, fgumbel$estimate)
46   MLE.norm = rbind(MLE.norm, fnorm$estimate)
47   kk = gofstat(list(fg, fln, fw, fgumbel,fnorm),fitnames = c("Gamma", "Inorm", "Weibull",
48   ", "Gumbel", "Norm"))[12]
49   aic.value = rbind(aic.value, as.numeric(unlist(kk)))
50 }
51 names(MLE.ln) = c("meanlog", "sdlog")
52 names(MLE.norm) = c("mean", "sd")
53 names(MLE.gamma) = c("shape", "rate")
54 names(MLE.wei) = c("shape", "scale")

```

```

55 names(aic.value) = c("Gamma", "Lognormal", "Weibull", "Gumbel", "Normal")
56
57 latlong = read.csv("C:\\\\Users\\\\Neeraj Poonia\\\\Desktop\\\\neeraj\\\\MA605\\\\NWH_temp_plot\\\\
58 new_latlongNWH.csv")
59
60 aic.value = data.frame(lat = latlong$lat, long = latlong$long, aic.value)
61
62 min.aic = numeric(525)
63 for(i in 1:525){
64   min.aic[i] = which(aic.value[i,3:7] == min(aic.value[i,3:7]))
65 }
66 min.aic
67 aic.value$model = min.aic
68 #=====
69
70 rp_fun = function(Time_period){
71 p=1/Time_period;p
72 RP = list()
73 for(i in 1:525){
74   #i=1
75   #p=1/20
76   d = data[,i]
77
78   if(aic.value$model[i]==1){
79     fg <- fitdist(d, "gamma")
80     x = qgamma(1-p, shape=as.numeric(fg$estimate[1]),
81                 rate = as.numeric(fg$estimate[2]))
82     RP = append(RP, x)
83     print("gamma")
84   } else if(aic.value$model[i]==2){
85     logx = log(d)
86     RP = append(RP, qlnorm(1-p, mean(logx), sd(logx)))
87     print("ln")
88   } else if(aic.value$model[i]==3){
89     fw = fitdist(d, "weibull")
90     x = qweibull(1-p, shape=as.numeric(fw$estimate[1]),
91                   scale = as.numeric(fw$estimate[2]))
92     RP = append(RP, x)
93     print("wei")
94   } else if(aic.value$model[i]==4){
95     fgum = fitdist(d, "gumbel", start=list(a=10, b=10))
96     x = fgum$estimate[1] - fgum$estimate[2]*log(-log(1-p))
97     RP = append(RP, x)
98     print("gum")
99   } else if(aic.value$model[i]==5){

```

```

100     RP = append(RP, qnorm(1-p,mean(d),sd(d)))
101     print("norm")
102   }
103 }
104 return(RP)
105 }

106
107
108 df = data.frame()
109 df = rbind(df, rp_fun(10)-data[41,], rp_fun(20)-data[41,], rp_fun(50)-data[41,], rp_fun
110   (100)-data[41,], rp_fun(200)-data[41,], rp_fun(500)-data[41,])
111 df = transpose(df)
112 names(df) = c("ten","twenty","fifty","hundred","twohundred","fivehundred")
113 tail(df,25)
114 df$long =latlong$long
115 df$lat =latlong$lat
116
117 #code 4
118 library(ggplot2)
119 library(rgdal)
120 obj.df = ggplot(df ,aes(x= long, y= lat, fill = hundred)) +geom_tile() +theme_classic()
121   +
122   theme(panel.border = element_rect(color = "black",fill = NA,size = 1)) +
123   xlab("Longitude") +ylab("Latitude") + labs(fill="Temperature") + ggtitle("(2121)")
124   +
125   scale_fill_gradientn(breaks = seq(min(df$hundred)+.1,max(df$hundred)-.1,length.out
126     =10) ,colors = rainbow(20)) +
127   coord_fixed() + theme(legend.key.height= unit(3, 'cm'), legend.key.width= unit(1, 'cm')
128   ))
129 plot(obj.df)
130
131 path = "C:\\\\Users\\\\Neeraj Poonia\\\\Desktop\\\\neeraj\\\\MA605\\\\NWH_temp_plot\\\\NWH"
132 shf = readOGR(path,"4-17-2018-899072")
133 obj.shf = geom_polygon(data=shf, aes(x= long, y= lat),
134   colour ="black", fill="white", alpha=0)
135 obj.df+obj.shf

```

```

1 time_period = 100
2 p= 1/time_period
3 rp=[]
4 for i in range(525):
5   if a[i]=='Weibull':
6     ff2 = Fitter(df2.iloc[i],distributions = ['weibull_min'])
7     ff2.fit()
8     weibull_params = ff2.fitted_param['weibull_min']

```

```

9      weibull_dist = stats.weibull_min.ppf(1-p, weibull_params[0], loc=
10     weibull_params[1], scale=weibull_params[2])
11     rp.append(weibull_dist)
12
13 elif a[i]=='Gamma':
14
15     ff2 = Fitter(df2.iloc[i],distributions = ['gamma'])
16     ff2.fit()
17
18     gamma_params = ff2.fitted_param['gamma']
19
20     gamma_dist = stats.gamma.ppf(1-p, gamma_params[0], loc=gamma_params[1], scale=
21     gamma_params[2])
22
23     rp.append(gamma_dist)
24
25 elif a[i]=='Lognormal':
26
27     ff2 = Fitter(df2.iloc[i],distributions = ['lognorm'])
28     ff2.fit()
29
30     lognorm_params = ff2.fitted_param['lognorm']
31
32     lognorm_dist = stats.lognorm.ppf(1-p, lognorm_params[0], loc=lognorm_params
33     [1], scale=lognorm_params[2])
34
35     rp.append(lognorm_dist)
36
37 elif a[i]=='Normal':
38
39     ff2 = Fitter(df2.iloc[i],distributions = ['norm'])
40
41     ff2.fit()
42
43     norm_params = ff2.fitted_param['norm']
44
45     norm_dist = stats.norm.ppf(1-p, norm_params[0], norm_params[1])
46
47     rp.append(norm_dist)
48
49 elif a[i]=='Gumbel':
50
51     ff2 = Fitter(df2.iloc[i],distributions = ['gumbel_r'])
52
53     ff2.fit()
54
55     gumbel_r_params = ff2.fitted_param['gumbel_r']
56
57     gumbel_r_dist = stats.gumbel_r.ppf(1-p, gumbel_r_params[0], gumbel_r_params
58     [1])
59
60     rp.append(gumbel_r_dist)
61
62
63 dft1=pd.DataFrame(list(zip(latlon['lat'],latlon['long'], rp)),columns=['Latitude','Longitude','return level'])
64
65 gdft1 = gpd.GeoDataFrame(dft1, geometry=gpd.points_from_xy(dft1.Longitude, dft1.
66     Latitude))
67
68 ax=gdft1.plot("return level",legend=True)
69
70 gdf1.plot(ax=ax,linewidth=1.5,color='black')
71
72
73 rp1 = []
74
75 for i in range(525):
76
77     rp1.append(rp[i]-df2.iloc[i].mean())
78
79
80 # After subtracting from mean of 41 years
81
82
83 dft1=pd.DataFrame(list(zip(latlon['lat'],latlon['long'], rp1)),columns=['Latitude','Longitude','return level'])

```

```

        Longitude','return_level'])
49 gdft1 = gpd.GeoDataFrame(dft1, geometry=gpd.points_from_xy(dft1.Longitude, dft1.
    Latitude))
50 ax=gdft1.plot("return_level",legend=True)
51 gdf1.plot(ax=ax,linewidth=1.5,color='black')

```

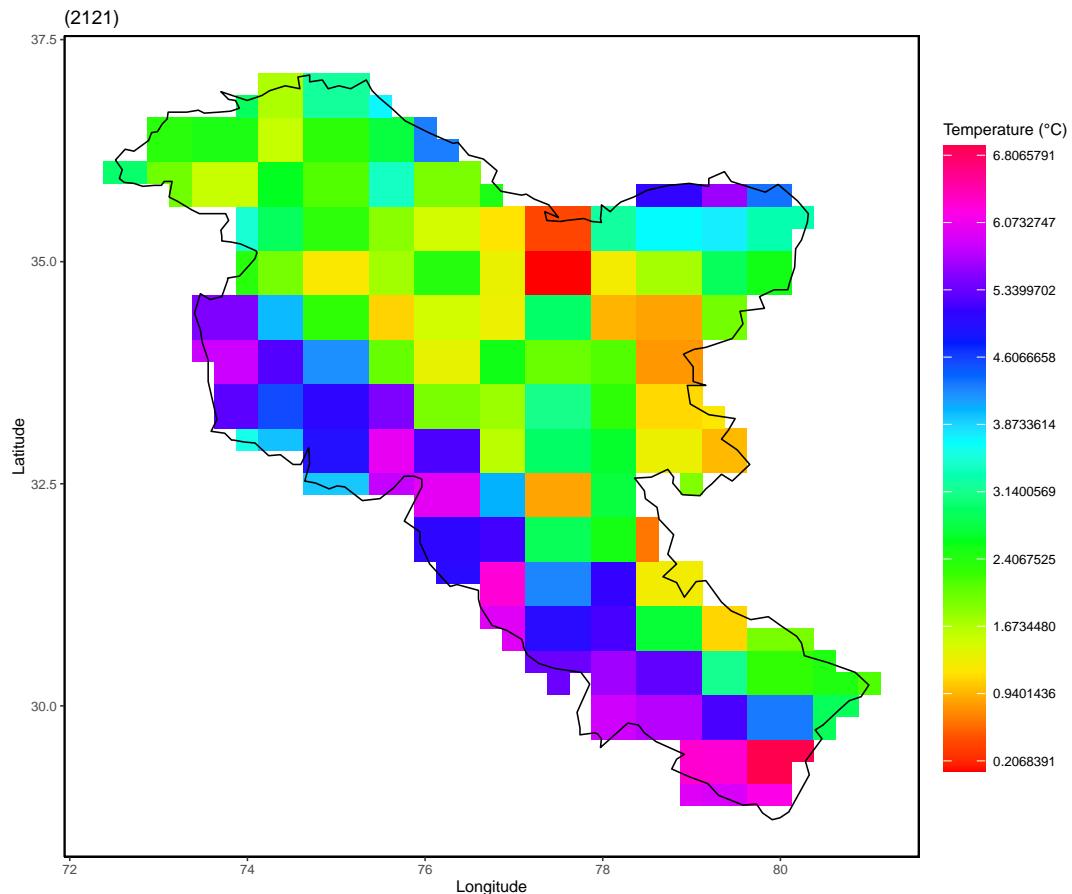


Fig. 3.5: Return level estimates for 100 years return period 525 grids of NWH.

3.5 Lab Assignment 1

About `numpy.random.choice()`:

`numpy.random.choice(a, size=None, replace=True, p=None)`

Parameters:

1. **a:** 1-D array of numpy having random samples.
2. **size:** Output shape of random samples of numpy array.
3. **replace:** Whether the sample is with or without replacement.

4. **p:** The probability attached with every sample in a.

Output: Return the numpy array of random samples.

Reference

numpy , numpy.random.choice

Question 1: Generate n random numbers in the range 0 to 100.

```
1 import random
2 n = 100
3 random_numbers = []
4 for i in range(n):
5     random_numbers.append(random.randrange(0, 101))
6
7 print(random_numbers)
```

Question 2: Display the frequencies of each random number generated in q-1 using a bar graph. Perform the experiment for n=1000, 2000, 5000.

```
1 import random
2 import matplotlib.pyplot as plt
3 n = 100
4 random_numbers = []
5 for i in range(n):
6     random_numbers.append(random.randrange(0, 101))
7
8 # print(random_numbers)
9
10 frequency = {}
11 for i in random_numbers:
12     if i in frequency:
13         frequency[i] += 1
14     else:
15         frequency[i] = 1
16
17 x_axis = list(frequency.keys())
18 y_axis = list(frequency.values())
19
20 plt.bar(range(len(frequency)), y_axis,
21         tick_label=x_axis, align='edge', width=0.3)
```

```

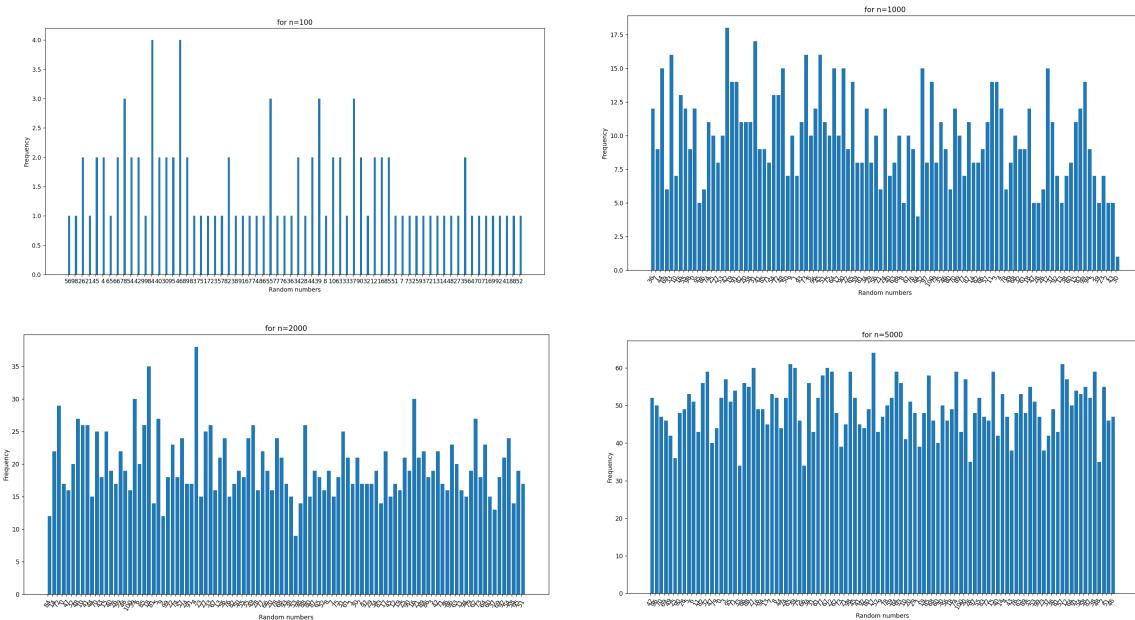
22 plt.xlabel('Random numbers')
23 plt.ylabel('Frequency')
24 plt.title('for n=100')
25 plt.show()
26
27 # for n=1000
28
29 n = 1000
30 random_numbers = []
31 for i in range(n):
32     random_numbers.append(random.randrange(0, 101))
33
34 # print(random_numbers)
35
36 frequency = {}
37 for i in random_numbers:
38     if i in frequency:
39         frequency[i] += 1
40     else:
41         frequency[i] = 1
42
43 x_axis = list(frequency.keys())
44 y_axis = list(frequency.values())
45
46 plt.bar(range(len(frequency)), y_axis, tick_label=x_axis,
47 align='edge')
48
49 plt.xticks(rotation=60)
50 plt.xlabel('Random numbers')
51 plt.ylabel('Frequency')
52 plt.title('for n=1000')
53 plt.show()
54
55 # for n=2000
56 n = 2000
57 random_numbers = []
58 for i in range(n):
59     random_numbers.append(random.randrange(0, 101))
60
61 # print(random_numbers)
62
63 frequency = {}
64 for i in random_numbers:
65     if i in frequency:
66         frequency[i] += 1
67     else:

```

```

68     frequency[i] = 1
69
70 x_axis = list(frequency.keys())
71 y_axis = list(frequency.values())
72
73 #plotting using matplotlib
74 plt.bar(range(len(frequency)), y_axis, tick_label=x_axis,
75 align='edge')
76 plt.xticks(rotation=60)
77 plt.xlabel('Random numbers')
78 plt.ylabel('Frequency')
79 plt.title('for n=2000')
80 plt.show()
81
82 # for n=5000
83 n = 5000
84 random_numbers = []
85 for i in range(n):
86     random_numbers.append(random.randrange(0, 101))
87
88 # print(random_numbers)
89
90 frequency = {}
91 for i in random_numbers:
92     if i in frequency:
93         frequency[i] += 1
94     else:
95         frequency[i] = 1
96
97 x_axis = list(frequency.keys())
98 y_axis = list(frequency.values())
99
100 #plotting using matplotlib
101 plt.bar(range(len(frequency)), y_axis, tick_label=x_axis,
102 align='edge')
103 plt.xticks(rotation=60)
104 plt.xlabel('Random numbers')
105 plt.ylabel('Frequency')
106 plt.title('for n=5000')
107 plt.show()

```



Question 3 Let the number of students in your class be n . Generate a random number from 1 to 365. We will thus have n birthdays. Find the probability that at least two people have the same birthday, denoted by p . For this case do the following by simulating the situation 1000 times:

- Find the probability for $n=23, 40, 80, 300$ and comment on the probabilities obtained.
- Plot p vs n where n varies from 1 to 300.
- Find the minimum value of n , for which the probability becomes 0.8 or greater.

```

1 #importing libraries
2 import random
3 import matplotlib.pyplot as plt
4
5 #function for calculating probability by generating random numbers
6 def prob(n):
7     c = 0 #count
8     #simulate for 1000 times
9     for i in range(1000):
10         l = [random.randrange(1, 366) for i in range(n)]
11         #generate random numbers from 1 to 365
12         for x in l:
13             if l.count(x) > 1:
14                 # condition that atleast two have same birthday
15                 c+=1 #increase the count

```

```

16         break
17     return c/1000
18
19 p1 = prob(23)
20 print(f'Probability that atleast 2 people
21 out of 23 people have same birthday:{p1}')
22 p2 = prob(40)
23 print(f'Probability that atleast 2 people
24 out of 40 people have same birthday:{p2}')
25 p3 = prob(80)
26 print(f'Probability that atleast 2 people
27 out of 80 people have same birthday:{p3}')
28 p4 = prob(300)
29 print(f'Probability that atleast 2 people
30 out of 300 people have same birthday:{p4}')
31
32 p = [] # list for storing probabilities (y-axis for plot)
33 np = [i for i in range(1, 301)]
34 # all numbers in range 1 to 300 (x-axis for plot)
35 # append probabilities by calling the prob function
36 for i in range(1, 301):
37     p.append(prob(i))
38
39 #plot using matplotlib
40 plt.plot(np, p)
41 plt.grid()
42 plt.xlabel('n')
43 plt.ylabel('p')
44 plt.title('p vs n')
45 plt.show()
46
47 #Finding n for p > = 0.8
48 for j in range(len(p)):
49     if p[j]>=0.8:
50         print(f'Probability is 0.8 for n = {j+1}')
51         break

```

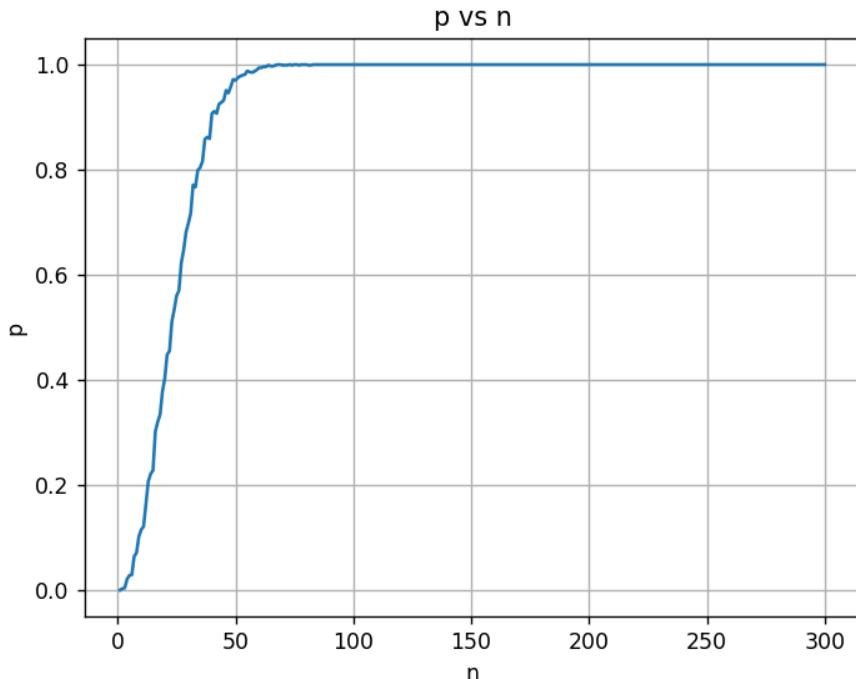


Fig. 3.6: Plot: Q3

Question 4 One Venus day lasts 243 earth days. If n number of people are born on a Venus day, find the probability that they share the same birthday according to the earth days. At some place on Venus there's sunlight for 122 days and night for 121 days. Assume that a person is twice as likely to be born in day time than at night. Compare the probabilities for the same values of $n=23, 40, 80$. Plot p vs n where n varies from 1 to 243. [Hint: use `numpy.random.choice` to generate random numbers with specified probabilities. Take care that the sum of probabilities should be 1]

```

1 #importing libraries
2 import random
3 import matplotlib.pyplot as plt
4 from numpy.random import choice
5
6 d = [i for i in range(1, 244)] #array for storing
7 numbers 1 to 244 (days)
8 l1 = [2/365 for i in range(1, 123)]
9 # probability array for 1-122 days (2/(122*2+121))
10
11 # appending probabilities for 123-243 days (1/(122*2+121))

```

```

12 for i in range(123, 244):
13     l1.append(1/365)
14
15 np, p = [], [] # declaring lists for n and p
16 for i in range(1, 244):
17     c=0 #count
18     np.append(i)
19     # simulate for 1000 times
20     for j in range(1000):
21
22         # generate random numbers from 1-243 with
23         # probabilities given by list l1
24         d1 = choice(d, size = i, p = l1) # numpy.random.choice()
25         # obtain list from numpy array d1
26         l = []
27         for k in range(i):
28             l.append(d1[k])
29         for li in l:
30             if l.count(li) > 1:
31                 # condition that atleast two have same birthday
32                 c+=1
33                 break
34         p.append(c/1000)
35 # Print probabilities for n = 23, 40, 80
36 print(f'P(23) = {p[22]}')
37 print(f'P(40) = {p[39]}')
38 print(f'P(80) = {p[79]}')

39
40 #plot using matplotlib
41 plt.plot(np, p)
42 plt.grid()
43 plt.xlabel('n')
44 plt.ylabel('p')
45 plt.title('p vs n')
46 plt.show()

```

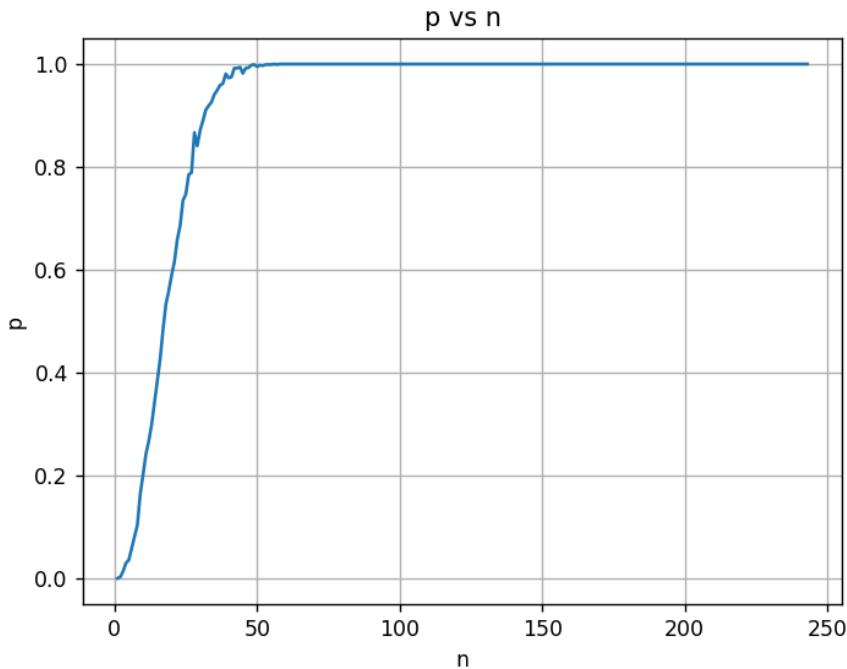


Fig. 3.7: Plot: Q4

3.6 Lab Assignment 2

Question 1 Suppose that a laboratory test to detect a certain disease has the following statistics.

A = event that the tested person has the disease

B = event that the test result is positive.

It is known that $P(B|A) = 0.99$ and $P(B|\bar{A}) = 0.005$, and 0.1 percent of the population actually has the disease. What is the probability that a person has the disease given that the test result is positive?

Solution:

$$P(A|B) = 0.99, P(B|\bar{A}) = 0.005, P(A) = 0.001, P(\bar{A}) = 0.999$$

$$\begin{aligned}
 P(A|B) &= \frac{P(B|A).P(A)}{P(B)} \\
 &= \frac{P(B|A).P(A)}{P(B|A).P(A) + P(B|\bar{A}).P(\bar{A})} \\
 &= \frac{0.99 \times 0.001}{0.99 \times 0.005 + 0.999 \times 0.001} \\
 &= 0.165
 \end{aligned} \tag{3.1}$$

Question 2 Let there be two unbiased N-sided dice that are thrown once, for instance, a 5-sided dice will have five faces, each having 1, 2, 3, 4, 5 number of dots respectively. Write a general program which takes N as input and give the following outputs.

1. Sample space $S = \dots$
2. Event E_1 that the sum of the dots on the dice equals N.
3. Event E_2 that the dots on the first dice is $\lfloor N/2 \rfloor$, where $\lfloor . \rfloor$ indicates the greatest integer function.
4. Event E_3 that the sum of the dots on the dice is greater than $\lfloor N \rfloor + \lfloor N/2 \rfloor$, where $\lfloor . \rfloor$ indicates the greatest integer function.
5. Event $E_4 = E_1 \cap E_3$, i.e., when the sum is N and greater than $\lfloor N \rfloor + \lfloor N/2 \rfloor$.
6. Probabilities of the events E_1 , E_2 , E_3 , and E_4 , i.e., $P(E_1)$, $P(E_2)$, $P(E_3)$, and $P(E_4)$.
7. Are events E_1 and E_2 independent? Also, output whether the events E_1 and E_3 independent.

```

1 # A
2 n = int(input("Enter the value of N"))
3 Sample_Space = set() # To handle the duplicate values
4 for i in range(1,n+1):
5     for j in range(1,n+1):
6         if(str(i)+str(j) not in Sample_Space):
7             Sample_Space.add(str(i)+str(j))
8             Sample_Space.add(str(j)+str(i))
9 print("Sample Space = ",Sample_Space)
10
11 # B
12
13 e1 = set()
14 for i in range(1,n):
15     e1.add(str(i)+str(n-i))
16 print("Sample Space of E1 = ",e1)
17
18 # C
19
20 import math
21 e2 = set()
```

```

22 element = int(math.floor(n/2))
23 print(element)
24 for i in range(1,n+1):
25     e2.add(str(element)+str(i))
26 print("Sample Space of E2 = ",e2)
27
28 # D
29
30 threshold = n + element
31 e3 = set()
32 for i in range(1,n+1):
33     for j in range(1,n+1):
34         if(i+j>threshold):
35             e3.add(str(i)+str(j))
36             e3.add(str(j)+str(i))
37 print("Sample Space of E3 = ",e3)
38
39 # E -
40
41 e4 = None
42 print("Sample Space of E4 = NULL")
43
44 # F
45
46 p_e1 = len(e1)/len(Sample_Space)
47 p_e2 = len(e2)/len(Sample_Space)
48 p_e3 = len(e3)/len(Sample_Space)
49 p_e4 = 0
50 print("P(E1) = ",p_e1,"nP(E2) = ",p_e2,"nP(E3) = ",p_e3,"nP(E4) = ",p_e4)
51
52 # G
53
54 e12 = set()
55 for i in range(1,n+1):
56     for j in range(1,n+1):
57         if(i==element and i+j==n):
58             e12.add(str(i)+str(j))
59 p_e12 = len(e12)/len(Sample_Space)
60 if(p_e12 == (p_e1*p_e2)):
61     print("E1 and E2 are : independent")
62 else:
63     print("E1 and E2 are : dependent")
64
65 print("E1 and E3 are : independent")
66 # Because p_e4 = 0 and p_e1 and p_e3 are non zero

```

Question 3 Repeat Question 2 to write a general program that takes N as input to output the parts 2f and 2g without using the probability formulas. Hence, run the simulation K times in a program and compute the probabilities by utilizing the counts of the desired outcomes. In your report, prepare the below table, Table 3, for a fixed N and increasing K to state your observations.

For a fixed value of $N =$	$K = 10$	$K = 50$	$K = 100$	$K = 1000$	$K = 5000$
$P(E_1)$	0.16	0.14	0.15	0.151	0.1624
$P(E_2)$	0.2	0.2	0.13	0.202	0.2038
$P(E_3)$	0.24	0.18	0.25	0.237	0.233
$P(E_4)$	0	0	0	0	0
Are E_1 & E_2 seem independent?	No	No	No	No	No
Are E_3 & E_4 seem independent?	Yes	Yes	Yes	Yes	Yes

Table 3.1: Table for Question-3

```

1 import random
2 k = int(input("Enter the value of K"))
3 dice1 = []
4 dice2 = []
5 for i in range(k):
6     dice1.append(random.randint(1,n))
7     dice2.append(random.randint(1,n))
8
9 e1 = 0
10 for i in range(k):
11     if((dice1[i]+dice2[i])==n):
12         e1+=1
13
14 e2 = dice1.count(element)
15
16 e3 = 0
17 for i in range(k):
18     if((dice1[i]+dice2[i])>(n+element)):
19         e3+=1
20
21 e12 = 0
22 for i in range(k):
23     if(dice1[i]==element and dice1[i]+dice2[i]==n):

```

```

24         e12+=1
25
26 p_e1 = e1/k
27 p_e2 = e2/k
28 p_e3 = e3/k
29 p_e4 = 0
30 p_e12 = e12/k
31 is_independent12 = "Dependent"
32 if(p_e12 == (p_e1*p_e2)):
33     is_independent12 = "Independent"
34 print("P(E1) = ",p_e1,"nP(E2) = ",p_e2,"\
35 P(E3) = ",p_e3,"nP(E4) = ",p_e4)
36 print("E1 and E2 are : ",is_independent12)
37 print("E3 and E4 are : independent")
38 # As p_e3e4 = 0 and p_e4 = 0

```

Note: The values in the table may vary due to randomness

3.7 Theory Assignment 1

Que. 1

We have four boxes. B1: 2000 items with 5% defective.

B2: 500 items with 40% defective.

B3: 1000 items with 10% defective.

B4: 1000 items with 10% defective.

We select one box at random and draw one component at random from the box. What is the probability that the selected item is defective?

Sol. 1

Box 1 => $2000 \times 5/100 = 100$ (defective)

Box 2 => $500 \times 40/100 = 200$ (defective)

Box 3 => $1000 \times 10/100 = 100$ (defective)

Box 4 => $1000 \times 10/100 = 100$ (defective)

$$P(B_1) = P(B_2) = P(B_3) = P(B_4) = 1/4$$

$$P(D | B_1) = 100/2000 = 0.05, P(D | B_2) = 200/500 = 0.4$$

$$P(D | B_3) = 100/1000 = 0.1, P(D | B_4) = 100/1000 = 0.1$$

$$P(D) = P(D | B_1) \cdot P(B_1) + P(D | B_2) \cdot P(B_2) + P(D | B_3) \cdot P(B_3) + P(D | B_4) \cdot P(B_4) = (0.05 + 0.4 + 0.1 + 0.1)/4 = 0.65/4 = 0.1625$$

Que. 2

How many committees of two chemists and one physicist can be formed from 4 chemists and 3

physicists?

Sol. 2

$$({}^4C_2)({}^3C_1) = 6 \times 3 = 18$$

Que. 3

Given that, by tossing an unfair coin we get 0.6 probability of getting heads. Then find.

- The probability of getting two heads in a row, by tossing the coin two times?
- The probability of getting tail, head and then tail in a row sequentially, by tossing the coin three times in a row.

Sol. 3

$$(a) P(H_1H_2) = P(H_1) \cdot P(H_2)$$

$$= 3/5 \times 3/5 = 9/25$$

$$(b) P(T_1H_2 T_3) = P(T_1) \cdot P(H_2) \cdot P(T_3)$$

$$= 2/5 \times 3/5 \times 2/5$$

$$= 12/125$$

Que. 4

If we randomly pick two television sets in succession from a shipman of 240 television sets of which 15 are defective, what is the probability that they will both be defective.

Sol. 4

Let A denote the event that the first television picked was defective.

Let B denote the event that the second television picked was defective.

Then $(A \cap B)$ denotes both were defective.

$$\begin{aligned} P(A \cap B) &= P(A)P(B | A) \\ \text{So,} \quad &= (15/240)(14/239) = 7/1912 \end{aligned}$$

Que. 5

A box of fuses contains 20 fuses of which 5 are defective. If 3 of them fuses are selected at random and removed from the box in succession without replacement. What is the probability that all fuses are defective?

Sol. 5

Let $A \rightarrow$ Event that the first fuse selected is defective.

Let $B \rightarrow$ Event that the second fuse selected is defective.

Let $C \rightarrow$ Event that the third fuse selected is defective.

Therefore,

$$\begin{aligned} P(A \cap B \cap C) &= P(A) \cdot P(B | A) \cdot P(C | (A \cap B)) \\ &= (5/20)(4/19)(3/18) \\ P(A \cap B \cap C) &= 1/114 \end{aligned}$$

Que. 6

Two boxes containing marbles are placed on a table. The boxes are labelled B_1 and B_2 . Box B_1 contains 7 green marbles and 4 white marbles.

Box B_2 contains 3 green marbles and 10 yellow marbles. The boxes are arranged so that the probability of B_1 is $1/3$ and the probability of selecting box B_2 is. Shyam is blindfolded and asked to select a marble. He will win a smartphone if selects a green marble.

- (a) What is the probability that Shyam will win the smartphone?
- (b) If Shyam wins the smartphone, what is the probability that the green marble was selected from first box?

Sol. 6

Let A be the event of drawing a green marble. The prior probabilities are.

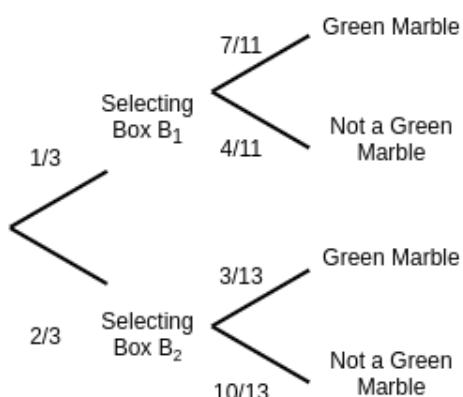
$$P(B_1) = 1/3, P(B_2) = 2/3$$

$$(a) P(A) = P(A | B_1)P(B_1) + P(A | B_2)P(B_2)$$

$$= (7/11)(1/3) + (3/13)(2/3)$$

$$P(A) = 157/429$$

(b)



Given that Shyam won the smartphone. The probability that the green marble was selected from B_1 , is:

$$\begin{aligned} P(B_1/A) &= P(A | B_1) \cdot P(B_1) / (P(A | B_1) \cdot P(B_1) + P(A | B_2) \cdot P(B_2)) \\ &= ((7/11) \times (1/3)) / ((7/11)(1/5) + (3/13) \cdot (2/3)) \\ P(B_1/A) &= 91/157 \end{aligned}$$

3.8 Theory Assignment 2

1. Eight different scented Dior perfumes, three different scented Chanel perfumes, four different scented Fragonard perfumes and six different scented Gucci perfumes to be arranged on a shelf. How many different arrangements are possible if

- (a) the perfumes of the same brand must all stand together?
- (b) only the Dior must stand together?
- (c) Chanel and Fragonard stand on shelf 1, and Dior and Gucci stand on shelf 2?
- (d) Chanel and Fragonard stand on one of the 2 horizontally together shelves, and Dior and Gucci stand on the other?

2. **Circular permutation**

Mayank is given a task by her mother to put 6 indistinguishable plates on a circular table separated by 6 distinguishable napkins. What is the number of possible permutations possible for arranging napkins and plates?

3. (a) Prove $(2n)! / (2^n \cdot n!) = (2n-1)(2n-3)\cdots 3 \cdot 1$.
- (b) Illustrate the equation using an example/ situation i.e STORY PROOF.

$$\binom{m+n}{k} = \sum_{j=0}^k \binom{m}{j} \binom{n}{k-j}$$

4. A goldsmith designs a hollow cylindrical gold disc (with height much smaller than the diameter). Say he fills it half with molten copper, seals it and hands it over to you. One side of the chip has Lotus engraved and the other plain. You toss the disc and say it falls on the engraved side. You repeat the toss (say infinite times). What can you say about the frequency of the plain side?
5. There is a single Amoeba in a pond. At every unit time interval, it either (a) dies (b) stays as it is (c) splits into two. If the probability of occurrence of (a) is twice the probability of (b) which is one-fourth the sum of (a) and (c), what is the probability that the amoeba population will die out?

6. Suppose that a box contains one green ball and four red balls, labeled A, B, C, and D. Now two of the five balls are selected at random, without replacement.
- If it is known that ball C has been selected, what is the probability that both balls are red?
 - If it is known that at least one red ball has been selected, what is the probability that both balls are red?

3.8.1 Theory Assignment-2-solutions

1. (a) Dior perfumes can be arranged among themselves in $8!$ ways, similarly

Chanel perfumes = $3!$

Fragonard perfumes = $4!$

Gucci perfumes = $6!$

And these 4 groups of perfumes can be arranged in $4!$ ways,

$$\text{Number of arrangements} = 4! * (8! * 3! * 4! * 6!) = 100,329,062,400$$

- (b) Consider Dior perfumes together as 1 unit i.e 1 perfume, hence,

we have $3+4+6+1= 14$ perfumes

In all these arrangements Dior perfumes will stay together, hence,

$$\text{Number of arrangements} = (8!) * 14! = 3.5 * 10^{15}$$

Another way, one of the interpretations could be only Dior perfumes are placed together, no other brands stand together. This can be solved using the inclusion-exclusion principle

- (c) No. of arrangements possible on shelf1 = $(3+4)!$

No. of arrangements possible on shelf2 = $(8+6)!$

Total no. = $7! * 14!$

- (d) This is similar to above question but since shelves aren't named, hence $2! * 7! * 14!$ is the total number of ways.

2. Since napkins will occupy alternate position on circular table with 1 plate between 2 napkins, the problem reduces to placing 6 plates on a circle with 6 napkins which don't make difference in no. of arrangements possible, hence,

the answer to this question is simply $(n-1)! = 5! = 120$

3. (a) $(2n)! = (1 . 3 . 5 . 7 \dots (2n)) * (2 . 4 . 6 . 8 \dots (2n))$

$$(2n)! = (1 . 3 . 5 . 7 \dots (2n)) * n! * 2^n$$

$$(2n)! / (2^n . n!) = (2n-1)(2n-3) \dots 3 . 1$$

Hence, proved.

- (b) Subjective to the creativity of students.

4. As the disk is tossed, assuming falls on X side means X side is facing ground, the engraved side is facing ground, The molten copper solidifies on the engraved side, and the disk is now biased. Eventually as the number of tosses increases and approaches infinity, the probability that we get the plain side approaches 1 and we can say the frequency of plain side approaches.
5. Let the probability of getting event a be $P(a)$, event b be $P(b)$, event c be $P(c)$

We know

$$P(a) + P(b) + P(c) = 1, P(a) = 2P(b), P(b) = (1/4) * (P(a) + P(c))$$

Solving for $P(a)$, $P(b)$ and $P(c)$

$$\begin{aligned} P(a) &= P(c) \\ &= \frac{2}{5} \\ P(b) &= 1 \end{aligned} \tag{3.2}$$

Let the probability amoeba population dies be $P(D)$,

Using LOTP,

$$P(D) = P(a) * P(D|a) + P(b) * P(D|b) + P(c) * P(D|c) \tag{3.3}$$

We can interpolate that

$$P(D|a) = 1, P(D|b) = P(D), P(D|c) = P(D)^2 \tag{3.4}$$

Substitute values in eq 2 and solve for variable $P(D)$

$$P(D)=1$$

6. (a) If a ball is selected, the other 4 are equally likely to be selected, thus, the required probability is $2/5$.
- (b)

$$\begin{aligned} P(R1 \cap R2) &= P(R1) * P(R2|R1) \\ &= \frac{4 * 3}{5 * 4} \\ &= \frac{3}{5} \end{aligned} \tag{3.5}$$

Chapter 4

Bayes' Theorem and Counting

4.1 Bayes' Theorem

Bayes' Theorem is a fundamental concept in probability theory that has important applications in a wide range of fields, including statistics, machine learning, and artificial intelligence. It was named after the Reverend Thomas Bayes, an 18th-century statistician and theologian who first discovered the principle.

Bayes' Theorem provides a way to calculate the probability of an event given information about related events. Essentially, it allows us to update our beliefs about the probability of an event, based on new evidence or information. The key idea is to start with a prior probability, which represents our initial belief or expectation about the probability of the event. Then, as we gather new evidence or data, we update this prior probability to obtain a new, posterior probability.

$$P(A \cap B) = P(A | B) \cdot P(B) - (1)$$

$$P(B \cap A) = P(B | A)P(A) - (2)$$

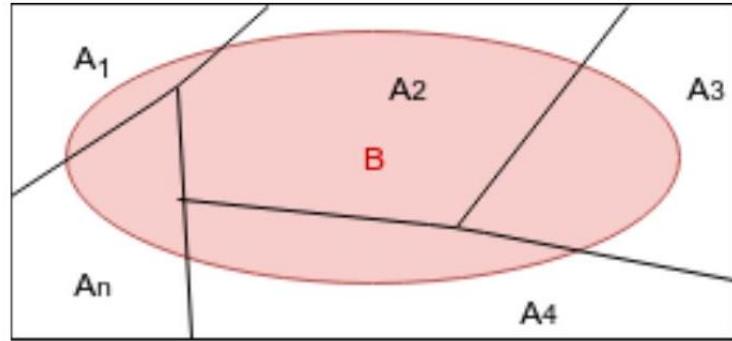
Equally (1) and (2)

Bayes' Theorem can be expressed mathematically as:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)} - (3)$$

$P(B)$, comes from total probability

Suppose that A_1, A_2, \dots, A_k form a partition of S : $A_i \cap A_j = \emptyset$; $\bigcup_i A_i = S$ $i = \dots n$



Let $B \subset S$, then

$$S = A_1 + A_2 + \dots + A_n$$

$$B = BS = B(A_1 + A_2 + \dots + A_n) = BA_1 + B_2 + \dots + A_n$$

$$\begin{aligned} P(B) &= P(BA_1) + P(BA_2) + \dots + P(BA_n) \\ &= P(B | A_1)P(A_1) + \dots + P(B | A_n)P(A_n) \end{aligned}$$

where $P(A | B)$ = posterior probability $P(A)$ = prior probability

$P(B | A)$ = Likelihood of our data being correct and $P(B)$ = Evidence

4.1.1 Bayes' Theorem Examples

Example-1: What is the probability that there is temperature increase, given that there is increment in GHG (green house gases)?

$$P(\text{temperature increase} | \text{GHG increase}) = ?$$

$P(\text{temperature increase})$ is the prior. $P(\text{GHG increase} | \text{temperature increase})$ is the likelihood $P(\text{GHG increase})$ is evidence

$$P(\text{temperature increase} | \text{GHG increase}) = \frac{P(\text{GHG increase} | \text{temperature increase})P(\text{temperature increase})}{P(\text{GHG increase})}$$

The same can be imagined with rain and clouds.

$$P(\text{rain} | \text{clouds})$$

Example-2: Coin flipping model

1st hypothesis: Coin is fair, 50% head or tail

$$P(A = \text{fair coin}) = 0.99$$

2nd hypothesis: Both sides of the coin are heads.

$$P(A = \text{unfair coin}) = 0.01$$

1st flip:

$$P(A = \text{fair} \mid B = \text{heads}) = \frac{P(B = \text{heads} \mid A = \text{fair})P(A = \text{fair})}{P(B = \text{heads})}$$

$$P(A = \text{fair}) = 0.99$$

$$P(B = \text{head} \mid A = \text{fair}) = 0.5$$

$$P(B = \text{heads}) = P(B = \text{heads} \mid A = \text{fair}) \times P(A = \text{fair})$$

$$+ P(B = \text{head} \mid A = \text{unfair})P(A = \text{unfair})$$

$$= 0.5 \times 0.99 + 1 \times 0.01 = 0.5050$$

$$P(A = \text{fair} \mid B = \text{heads}) = \frac{0.5 \times 0.99}{0.5050} = 0.9802$$

A coin is flipped a second time and it is head again.

The posterior in the previous time steps becomes the new prior.

$$P(A = \text{fair}) = 0.9802$$

$$P(B = \text{head} \mid A = \text{fair}) = 0.5$$

$$P(A = \text{fair} \mid B = \text{head})$$

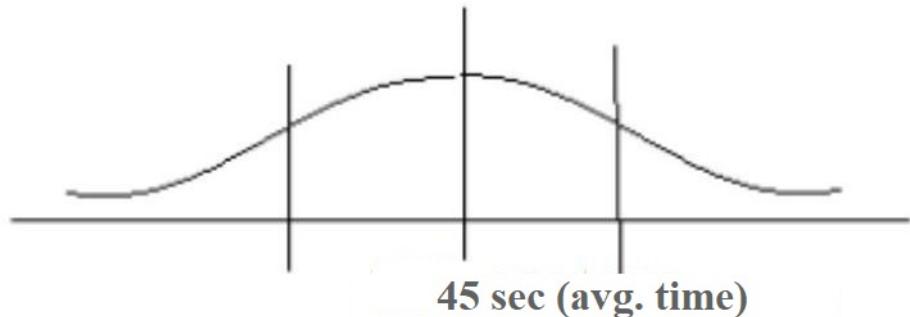
$$= \frac{P(B = \text{head} \mid A = \text{fair}) \times P(\text{fair})}{P(B = \text{head} \mid A = \text{fair})P(\text{fair}) + P(B = \text{head} \mid A = \text{unfair})P(A = \text{unfair})}$$

$$= \frac{0.5 \times 0.9802}{0.5 \times 0.9802 + 1 \times 0.0198}$$

$$= \frac{0.5 \times 0.9802}{0.5099} = 0.96$$

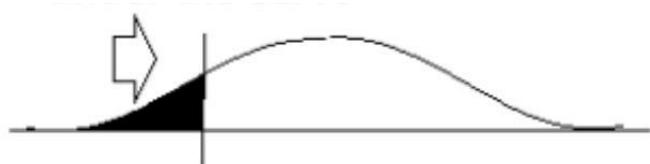
So, in this example we have shown the probability of the coin being fair will decrease if in every flip head appears. As more and more flips are made and new data is observed, our beliefs get updated.

Example-3: Suppose there is a car racing game in Delhi and Sania is master in it. 10 racers follow the normal distribution $T \sim N(45, 25)$, i.e. Time taken to ride 1 km has a normal distribution.



What is the probability that Sania rides 1 km in less than 36.5 secs.

$$P(T < 36.5) = 0.045 \text{ (calculated from area under the curve)}$$



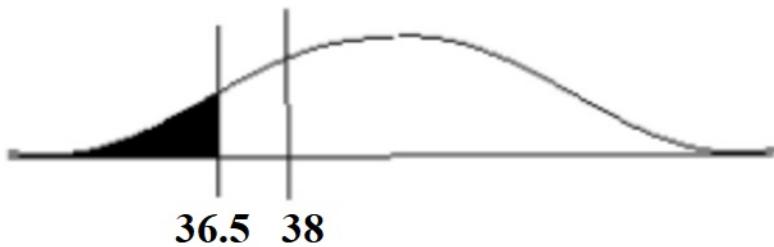
Suppose we have more information, to get into the team of car racers of Delhi, you need to be able to ride 1 km in less than 38sec. Suppose we also know that Sania is in car racers team of Delhi.

So, now can we change the probability $P(T < 36.5) = 0.045$, if we know that she is in car racers team of Delhi.

Hypothesis H: Sania rides 1 km in less than 36.5sec

Evidence E = Sania is in car racers team of Delhi

$$\begin{aligned} P(H | E) &= \frac{P(H \cap E)}{P(E)} \\ &= \frac{P(T < 36.5 \cap T < 38)}{P(T < 38)} \\ &= \frac{P(T < 36.5)}{P(T < 38)} \\ &= \frac{0.045}{0.081} \approx 56\% \end{aligned}$$

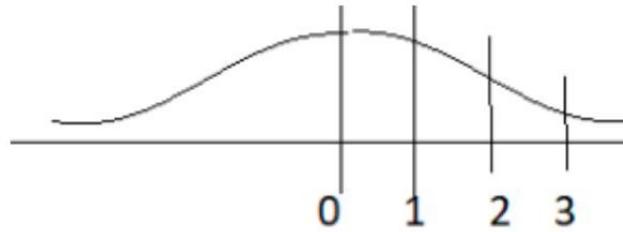


Evidence made us significantly update our thoughts about whether Sania can ride 1 km in 36.5 secs.

Difference between Probability and likelihood

$$P((1 \leq x \leq 3) | \mu = 0, \text{sd} = 1)$$

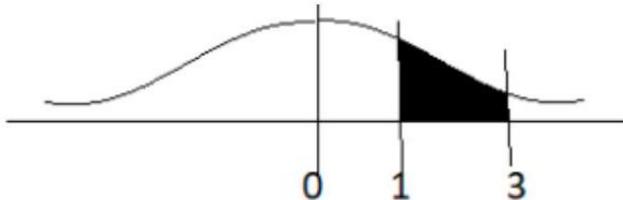
Probability = chance of occurrence of an event



What is the event, this year total rainfall in regions of India will increase from 1% to 3%.

Likelihood: We are trying to estimate the chance of a region where rainfall has been increased between 1% to 3% of this distribution.

$$L_A(\mu = 0, \text{sd} = 1 | (1 \leq x \leq 3))$$



Example 4. Consider weather data is coming from two sources Acuwheather and Wheather online. Let A_1 be the event that data received from Accuwheather and A_2 be the event that data received from Wheather online. We get 45% data from Accuwheather and 55% from Wheather online.

Thus:

$$P(A_1) = .45 \text{ and } P(A_2) = .55$$

Quality of data differs between two sources

	Percentage Correct data	Percentage Wrong data
Source 1	93	7
Source 2	97	3

Let G denote an instance when the correct data is processed and B denote the event that the wrong

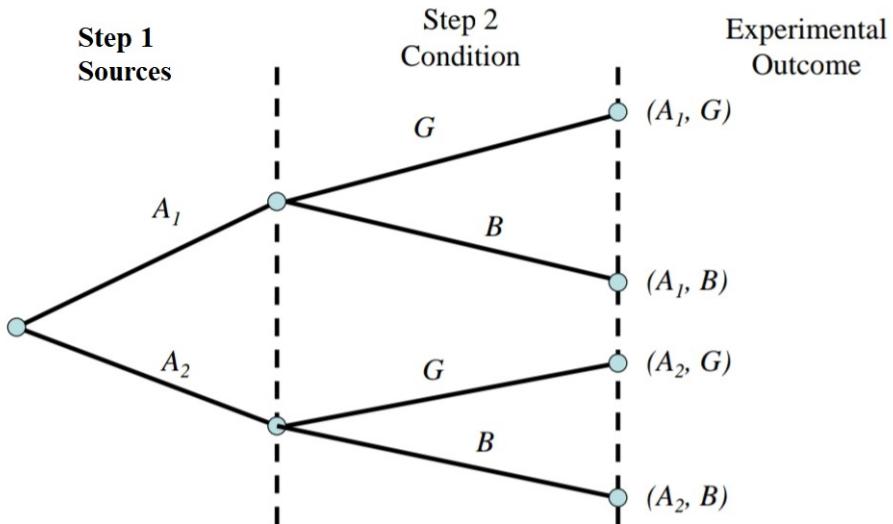
data is processed. Thus we have the following conditional probabilities:

$$P(G | A_1) = .93 \text{ and } P(B | A_1) = .07$$

$$P(G | A_2) = .97 \text{ and } P(B | A_2) = .03$$

Wrong data is processed and the reputation of the company is questioned. What is the probability the data came from Accuwheather?

Tree Diagram for Two-Source Example



We know from the law of conditional probability that:

$$P(A_1 | B) = \frac{P(A_1 \cap B)}{P(B)}$$

Observe from the probability tree that:

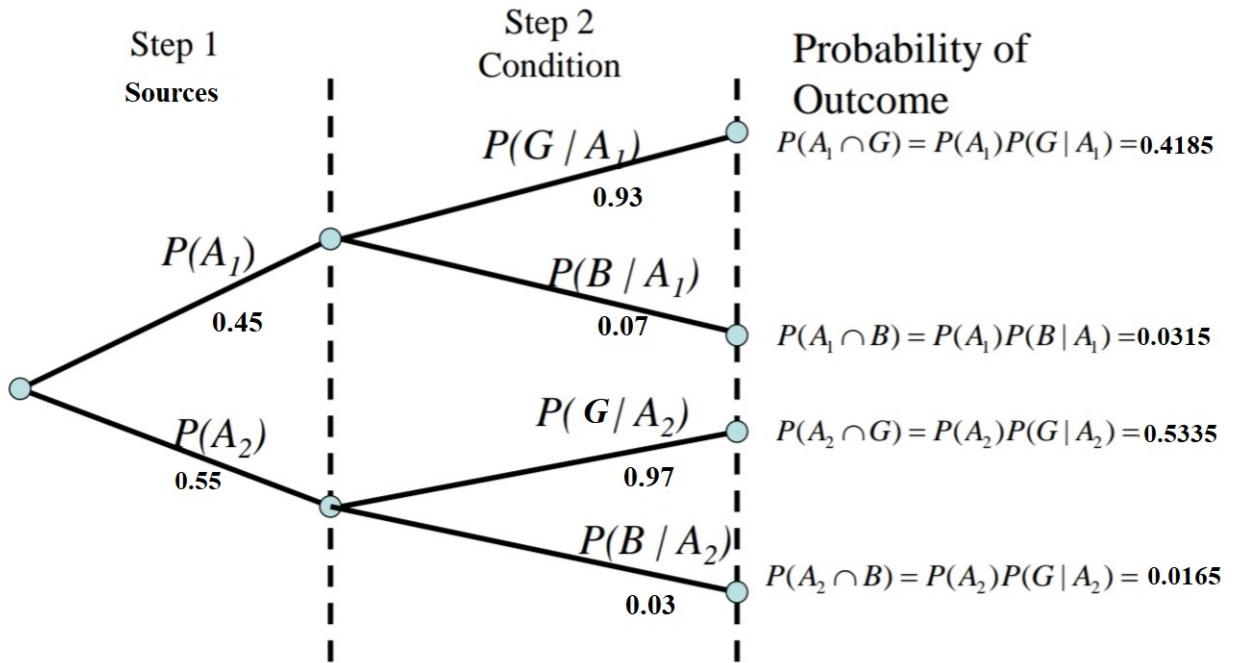
$$P(A_1 \cap B) = P(A_1) P(B | A_1)$$

The probability of getting a wrong data is found by adding together the probability of getting a wrong data from source 1 and the probability of getting a wrong data from source 2.

That is:

$$\begin{aligned}
P(B) &= P(A_1 \cap B) + P(A_2 \cap B) \\
&= P(A_1) P(B | A_1) + P(A_2) P(B | A_2)
\end{aligned}$$

Probability Tree for Two-Source Example



$$\begin{aligned}
P(A_1 | B) &= \frac{P(A_1) P(B | A_1)}{P(A_1) P(B | A_1) + P(A_2) P(B | A_2)} \\
&= \frac{(0.45)(0.07)}{(0.45)(0.07) + (0.55)(0.03)} = \frac{0.0315}{0.048} = 0.65625
\end{aligned}$$

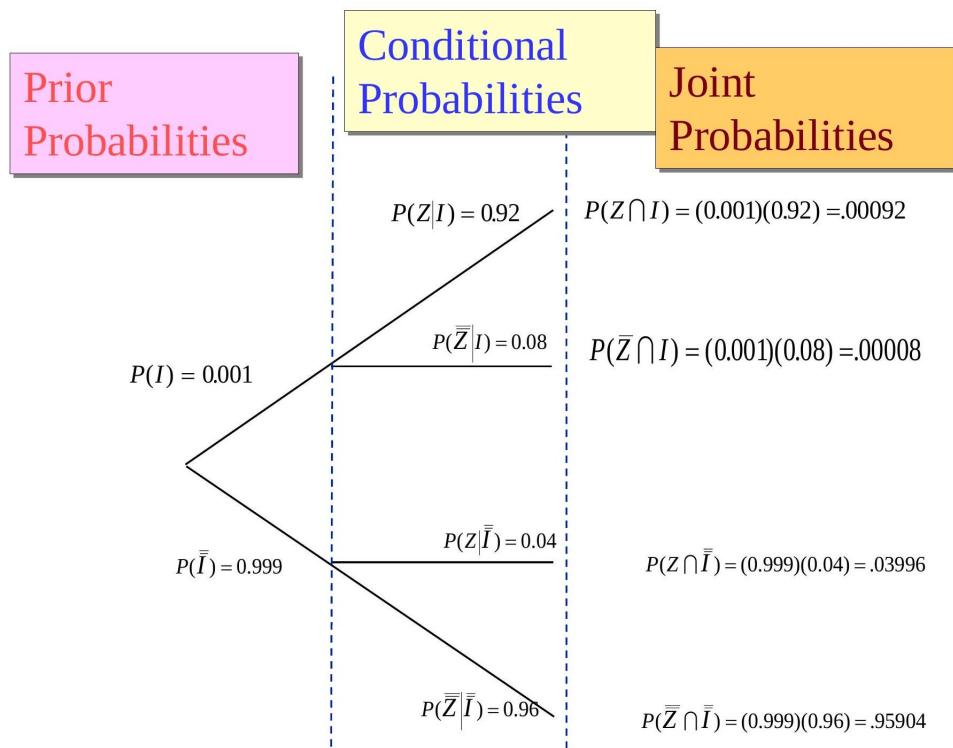
$$\begin{aligned}
P(A_2 | B) &= \frac{P(A_2) P(B | A_2)}{P(A_1) P(B | A_1) + P(A_2) P(B | A_2)} \\
&= \frac{(0.55)(0.03)}{(0.45)(0.07) + (0.55)(0.03)} = \frac{0.0165}{0.048} = 0.34375
\end{aligned}$$

Tabular Approach to Bayes' Theorem- 2-Source Problem

Events A_i	Prior Probabilities $P(A_i)$	Conditional Probabilities $P(B A_i)$	Joint Probabilities $P(A_i \cap B)$	Posterior Probabilities $P(A_i B)$
A_1	.45	.07	0.0315	$\frac{0.0315}{0.048} = 0.65625$
A_2	.55	.03	.0165	$\frac{0.0165}{0.048} = 0.34375$
	1.00		$P(B) = .048$	1.000

Example 5. 0.1% of the population has X disease. A screening test accurately detect the disease for 92% of people with it. The test also indicates the disease for 4% of the people without it (false positive). Suppose a person is Screened for the disease test positive. What is the probability that he actually has the disease?

- A medical test for a rare disease (affecting 0.1% of the population [$P(I) = 0.001$]) is imperfect:
 - ✓ When administered to an ill person, the test will indicate so with probability 0.92 [$P(Z | I) = .92 \Rightarrow P(\bar{Z} | I) = .08$]
 - The event ($\bar{Z} | I$) is a false negative
 - ✓ When administered to a person who is not ill, the test will erroneously give a positive result (false positive) with probability 0.04 [$P(Z|\bar{I}) = 0.04 \Rightarrow P(\bar{Z} | \bar{I}) = 0.96$]
 - The event ($Z | \bar{I}$) is a false positive.



Applying Bayes' Theorem

$$\begin{aligned}
P(I) &= 0.001 \\
P(\bar{I}) &= 0.999 \\
P(Z|I) &= 0.92 \\
P(Z|\bar{I}) &= 0.04
\end{aligned}$$

$$\begin{aligned}
P(I|Z) &= \frac{P(I \cap Z)}{P(Z)} \\
&= \frac{P(I \cap Z)}{P(I \cap Z) + P(I \cap \bar{Z})} \\
&= \frac{P(Z|I)P(I)}{P(Z|I)P(I) + P(Z|\bar{I})P(\bar{I})} \\
&= \frac{(0.92)(0.001)}{(0.92)(0.001) + (0.04)(0.999)} \\
&= \frac{0.00092}{0.00092 + 0.03996} = \frac{0.00092}{0.04088} \\
&= 0.0225
\end{aligned}$$

$$\begin{aligned}
P(I) &= 0.001 \\
P(\bar{I}) &= 0.999 \\
P(Z | I) &= 0.92 \\
P(Z | \bar{I}) &= 0.04P(I | Z) && = \frac{P(I \cap Z)}{P(Z)} \\
&= \frac{P(I \cap Z)}{P(I \cap Z) + P(I \cap \bar{Z})} \\
&= \frac{P(Z | I)P(I)}{P(Z | I)P(I) + P(Z | \bar{I})P(\bar{I})} \\
&= \frac{(0.92)(0.001)}{(0.92)(0.001) + (0.04)(0.999)} \\
&= \frac{0.00092}{0.00092 + 0.03996} = \frac{0.00092}{0.04088} \\
&= 0.0225
\end{aligned}$$

4.2 Counting techniques

Counting can be used in a wide range of fields, including mathematics, computer science, engineering, physics, and many others. It is used to analyze and solve problems that involve questions of probability, optimization, and decision-making.

Counting is the process of determining the number of ways in which a certain event can happen or a certain object can be arranged or selected. It involves identifying the relevant objects and the criteria by which they can be combined, arranged, or selected, and then using mathematical formulas or methods to determine the total number of possibilities.

There are various techniques and methods used for counting, including:

- **Multiplication principle:** This principle states that if there are n ways to perform one task and m ways to perform another task, then there are - possible ways to perform both tasks in sequence.

- **Addition principle:** This principle states that if there are n ways to perform one task and m ways to perform another task, and these tasks are mutually exclusive, then there are $n + m$ possible ways to perform one of the tasks.
- **Permutations:** A permutation is an arrangement of objects in which order matters. The number of permutations of n objects taken k at a time is denoted by $P(n, k)$ and is given by $P(n, k) = n(n - 1)(n - 2) \cdots (n - k + 1)$.
- **Combinations:** A combination is a selection of objects in which order does not matter. The number of combinations of n objects taken k at a time is denoted by $C(n, k)$ and is given by $C(n, k) = \frac{n!}{k!(n-k)!}$.

4.2.1 Finite uniform probability space

The “uniform probability measure” on a finite sample space S assigns the same probability, $1/|S|$, to each outcome. By additivity, $P(A) = |A|/|S|$ where $|A|$ denotes the number of elements in A . Many examples fall into this category

1. Finite number of outcomes
2. All outcomes are equally likely
3. $P[E] = \frac{n(E)}{n(S)} = \frac{n(E)}{N} = \frac{\text{no. of outcomes in } E}{\text{total no. of outcomes}}$

Note: $n(A) = \text{no. of elements of } A$

To handle problems in case we have to be able to count. Count $n(E)$ and $n(S)$.

4.2.2 Techniques for counting

4.2.3 Rule 1

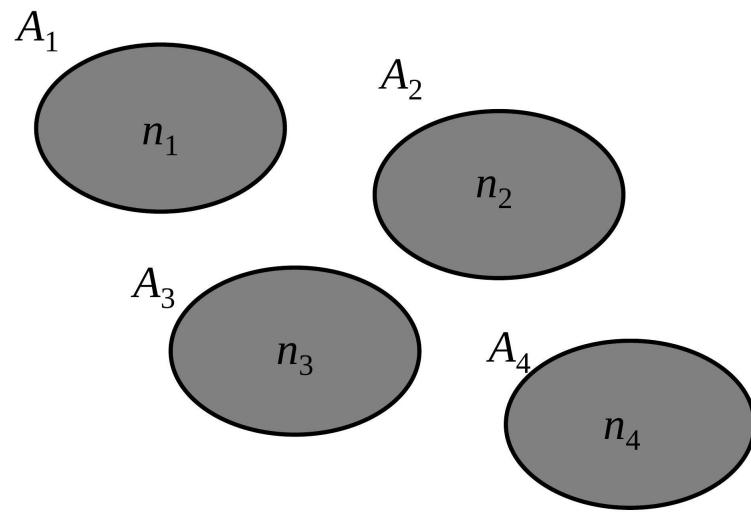
Suppose we carry out have a sets A_1, A_2, A_3, \dots and that any pair are mutually exclusive (i.e. $A_1 \cap A_2 = \emptyset$) Let

$n_i = n(A_i) = \text{the number of elements in } A_i$.

Let $A = A_1 \cup A_2 \cup A_3 \cup \dots$

Then,

$$\begin{aligned} N = n(A) &= \text{the number of elements in } A \\ &= n_1 + n_2 + n_3 + \dots \end{aligned}$$



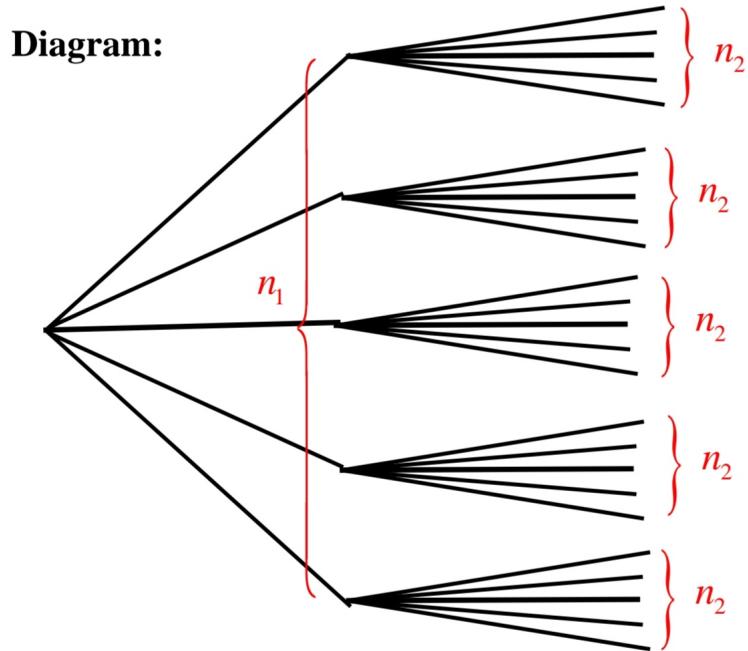
4.2.4 Rule 2

Suppose we carry out two operations in sequence

Let n_1 = the number of ways the first operation can be performed.

n_2 = the number of ways the second operation can be performed once the first operation has been completed.

Then, $N = n_1 n_2$ = the number of ways the two operations can be performed in sequence.



Examples

1. We have a team of 11 cricket players. We want to choose one captain and one vice-captain. How many ways we can do this?

Solution: Let n_1 = the number of ways the captain can be chosen = 11.

Let n_2 = the number of ways the vice-captain can be chosen once the chair has been chosen = 10.

$$\text{Then } N = n_1 n_2 = (11)(10) = 110$$

2. One card has been drawn from a well-shuffled pack. What is the probability that card being either a queen or a red? **Solution:** Let event E is the card drawn being either red or a queen.

The total number of outcomes = 52

There are 26 red cards, and 4 cards which are queen. However, 2 of the red cards are queen. If we add 26 and 4, we will be counting these two cards twice.

Thus, the correct number of outcomes which are favorable to E is

$$26 + 4 - 2 = 28$$

Hence, the probability of event occurring is

$$P(E) = 28/52 = 7/13$$

4.2.5 The Multiplicative Rule of Counting

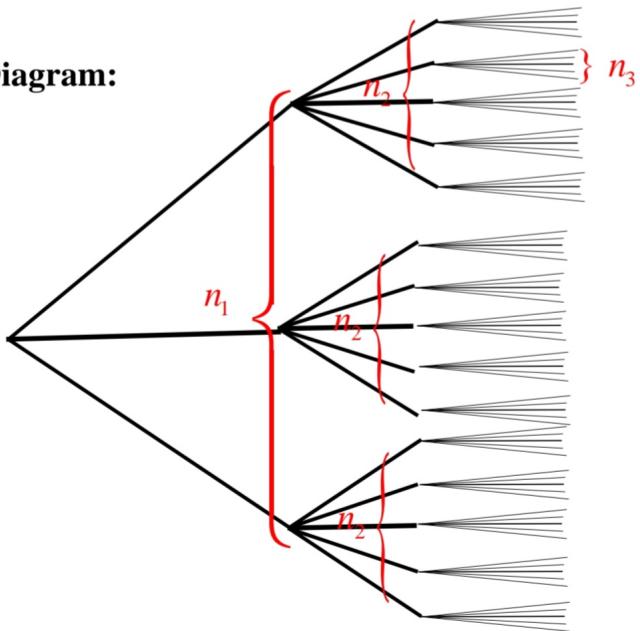
Suppose we carry out k operations in sequence

Let n_1 = the number of ways the first operation can be performed

n_i = the number of ways the i^{th} operation can be performed once the first $(i - 1)$ operations have been completed. $i = 2, 3, \dots, k$

Then $N = n_1 n_2 \dots n_k$ = the number of ways the k operations can be performed in sequence.

Diagram:



4.2.6 Permutations:

How many ways can you order n objects

Solution:

Ordering n objects is equivalent to performing n operations in sequence.

1. Choosing the first object in the sequence ($n_1 = n$)
2. Choosing the 2nd object in the sequence ($n_2 = n - 1$).
- k. Choosing the k^{th} object in the sequence ($n_k = n - k + 1$)
- n. Choosing the n^{th} object in the sequence ($n_n = 1$)

The total number of ways this can be done is:

$$N = n(n - 1) \dots (n - k + 1) \dots (3)(2)(1) = n !$$

Example How many ways can you order the 4 objects

Solution:

$$N = 4! = 4(3)(2)(1) = 24$$

Let objects are numbered by 1,2,3,4 respectively. Then, orderings can be given by:

1234	1243	1324	1342	1423	1432
2134	2143	2314	2341	2413	2431
3124	3142	3214	3241	3412	3421
4123	4132	4213	4231	4312	4321

4.2.7 Permutations of size $k (< n)$:

How many ways can you choose k objects from n objects in a specific order

Solution: This operation is equivalent to performing k operations in sequence.

1. Choosing the first object in the sequence ($n_1 = n$)
2. Choosing the 2^{nd} object in the sequence ($n_2 = n - 1$).
- k. Choosing the k^{th} object in the sequence ($n_k = n - k + 1$)

The total number of ways this can be done is: $N = n(n - 1) \dots (n - k + 1) = n!/(n - k)!$! This number is denoted by the symbol $_nP_k = n(n - 1) \dots (n - k + 1) = \frac{n!}{(n - k)!}$

Definition: $0! = 1$

This definition is consistent with $_nP_k = n(n - 1) \dots (n - k + 1) = \frac{n!}{(n - k)!}$ for $k=n$
 $_nP_n = \frac{n!}{0!} = \frac{n!}{1} = n!$

Example How many permutations of size 3 can be found in the group of 5 objects $\{A, B, C, D, E\}$

Solution: $_5P_3 = \frac{5!}{(5-3)!} = 5(4)(3) = 60$

ABC	ABD	ABE	ACD	ACE	ADE	BCD	BCE	BDE	CDE
ACB	ADB	AEB	ADC	AEC	AED	BDC	BEC	BED	CED
BAC	BAD	BAE	CAD	CAE	DAE	CBD	CBE	DBE	DCE
BCA	BDA	BEA	CDA	CEA	DEA	CDB	CEB	DEB	DEC
CAB	DAB	EAB	DAC	EAC	EAD	DBC	EBC	EBD	ECD
CAB	DBA	EBA	DCA	ECA	EDA	DCB	ECB	EDB	EDC

Example We have a committee of $n = 11$ cricket players and we want to choose a captain, a vice-captain and an umpire.

Solution: We want to choose 3 players out of 11 in a specific order. (Permutations of size 3 from a group of 10).

$${}_{11}P_3 = \frac{11!}{(11-3)!} = \frac{11!}{8!} = (11)(10)(9) = 990$$

Example: In a class, there are 26 boys and 14 girls. The teacher wants to select a monitor, sub-monitor, and checker. 1 boy and 1 girl to represent the class for a function. What is the probability that the three executives selected are all girls?

Solution: We want to choose 3 people out of 40 in a specific order. (Permutations of size 3 from a group of 40).

$${}_{40}P_3 = \frac{40!}{(40-3)!} = \frac{40!}{37!} = (40)(39)(38) = 59280$$

This is the size, $N = n(S)$, of the sample space S . Assume all outcomes in the sample space are equally likely.

Let E be the event that all three executives are girls

$$n(E) = {}_{14}P_3 = \frac{14!}{(14-3)!} = \frac{14!}{11!} = (14)(13)(12) = 2184$$

Hence

$$P[E] = \frac{n(E)}{n(S)} = \frac{2184}{59280} = 0.036$$

Thus if all students are equally likely to be selected to any position on the executive then the probability of selecting all girl executive is 0.036.

4.2.8 Combinations of size ($k < n$) :

A combination of size k chosen from n objects is a subset of size k where the order of selection is irrelevant. How many ways can you choose a combination of size k objects from n objects (order of selection is irrelevant).

Here are the combinations of size 3 selected from the 5 objects $\{A, B, C, D, E\}$

$\{A, B, C\}$	$\{A, B, D\}$	$\{A, B, E\}$	$\{A, C, D\}$	$\{A, C, E\}$
$\{A, D, E\}$	$\{B, C, D\}$	$\{B, C, E\}$	$\{B, D, E\}$	$\{C, D, E\}$

Example How many ways can you choose a combination of size k objects from n objects (order of selection is irrelevant).

Solution: Let n_1 denote the number of combinations of size k . One can construct a permutation of size k by:

1. Choosing a combination of size k (n_1 = unknown)
2. Ordering the elements of the combination to form a permutation ($n_2 = k!$)

Thus ${}_nP_k = \frac{n!}{(n-k)!} = n_1k!$

and $n_1 = \frac{{}_nP_k}{k!} = \frac{n!}{(n-k)!k!}$ = the # of combinations of size k .

The number:

$$n_1 = \frac{{}_nP_k}{k!} = \frac{n!}{(n-k)!k!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots(1)}$$

is denoted by the symbol

$${}_nC_k \text{ or } \binom{n}{k} \text{ read 'n choose } k'$$

It is the number of ways of choosing k objects from n objects (order of selection irrelevant). ${}_nC_k$ is also called a **binomial coefficient**. It arises when we expand $(x+y)^n$ (**the binomial theorem**)

The Binomial theorem:

$$\begin{aligned}
 (x+y)^n &= {}_n C_0 x^0 y^n + {}_n C_1 x^1 y^{n-1} + {}_n C_2 x^2 y^{n-2} + \\
 &\quad \dots + {}_n C_k x^k y^{n-k} + \dots + {}_n C_n x^n y^0 \\
 &= \binom{n}{0} x^0 y^n + \binom{n}{1} x^1 y^{n-1} + \binom{n}{2} x^2 y^{n-2} + \\
 &\quad \dots + \binom{n}{k} x^k y^{n-k} + \dots + \binom{n}{n} x^n y^0
 \end{aligned}$$

Proof: The term $x^k y^{n-k}$ will arise when we select x from k of the factors of $(x+y)^n$ and select y from the remaining $n - k$ factors. The no. of ways that this can be done is:

$$\binom{n}{k}$$

Hence there will be $\binom{n}{k}$ terms equal to $x^k y^{n-k}$ and

$$\begin{aligned}
 (x+y)^n &= \binom{n}{0} x^0 y^n + \binom{n}{1} x^1 y^{n-1} + \binom{n}{2} x^2 y^{n-2} + \\
 &\quad \dots + \binom{n}{k} x^k y^{n-k} + \dots + \binom{n}{n} x^n y^0
 \end{aligned}$$

Pascal's triangle - a procedure for calculating binomial coefficients

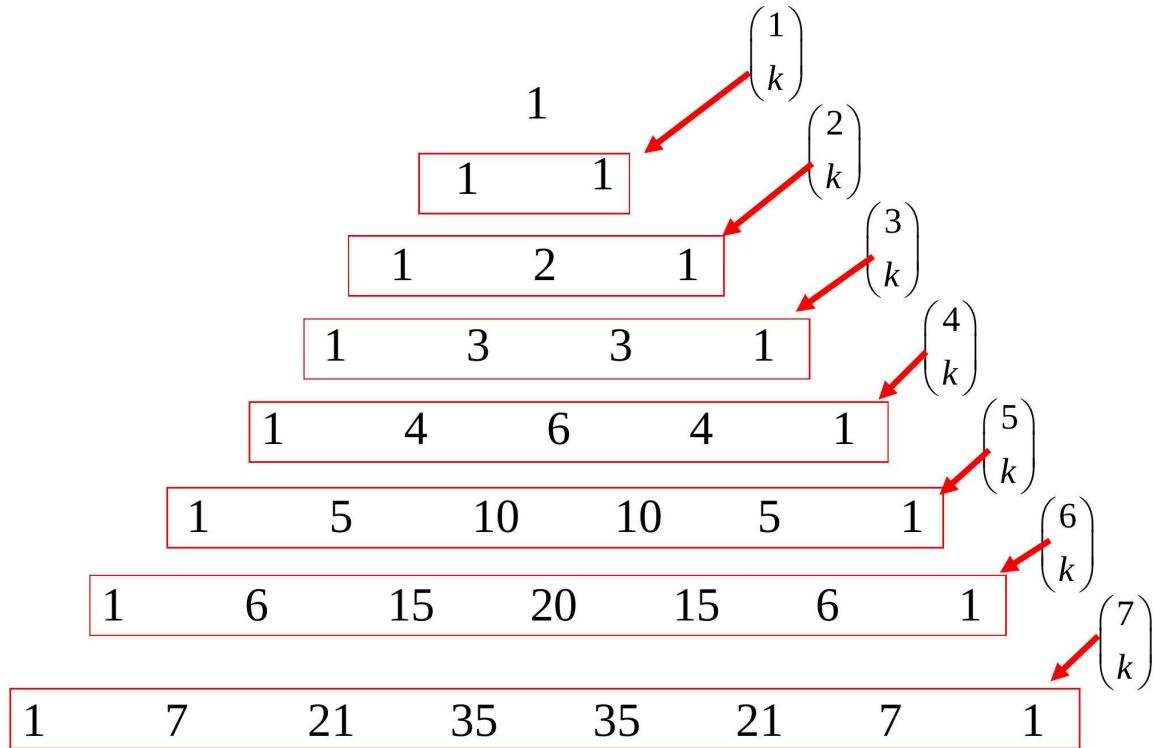
Pascal's triangle is a triangular array of numbers in which each number is the sum of the two numbers directly above it. The first and second rows are always 1, and each subsequent row is generated by adding the two adjacent numbers in the previous row. Pascal's triangle is used to find coefficients for binomial expansions.

			1							
			1	1						
			1	2	1					
			1	3	3	1				
			1	4	6	4	1			
			1	5	10	10	5	1		
			1	7	21	35	35	21	7	1

- The two edges of Pascal's triangle contain 1's
- The interior entries are the sum of the two nearest entries in the row above
- The entries in the n^{th} row of Pascals triangle are the values of the binomial coefficients

$$\binom{n}{0} \binom{n}{1} \binom{n}{2} \binom{n}{3} \binom{n}{4} \dots \binom{n}{k} \dots \binom{n}{n-1} \binom{n}{n}$$

Pascal's triangle



The Binomial Theorem

$$(x + y)^1 = x + y$$

$$(x + y)^2 = x^2 + 2xy + y^2$$

$$(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

$$(x + y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$$

$$(x + y)^5 = x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5$$

$$(x + y)^6 = x^6 + 6x^5y + 15x^4y^2 + 20x^3y^3 + 15x^2y^4 + 6xy^5 + y^6$$

$$(x + y)^7 = x^7 + 7x^6y + 21x^5y^2 + 35x^4y^3 + 35x^3y^4 + 21x^2y^5 + 7xy^6 + y^7$$

Summary of counting rules

Rule 1

$$n(A_1 \cup A_2 \cup A_3 \cup \dots) = n(A_1) + n(A_2) + n(A_3) + \dots$$

if the sets A_1, A_2, A_3, \dots are pairwise mutually exclusive (i.e. $A_i \cap A_j = \emptyset$)

Rule 2

$N = n_1 n_2$ = the number of ways that two operations can be performed in sequence if

n_1 = the number of ways the first operation can be
performed

n_2 = the number of ways the second operation can be performed once the first operation has been completed.

Rule 3

$N = n_1 n_2 \dots n_k$ = the number of ways the k operations can be performed in sequence if n_1 = the number of ways the first operation can be performed n_i = the number of ways the i^{th} operation can be performed once the first $(i - 1)$ operations have been completed. $i = 2, 3, \dots, k$

Basic counting formulae

1. Ordering

$n!$ = the number of ways you can order n objects

2. Permutations

The number of ways that you can
 $n P_k = \frac{n!}{(n - k)!}$ = choose k objects from n in a
specific order

3. Combinations

$\binom{n}{k} = {}^n C_k = \frac{n!}{k!(n - k)!}$ = The number of ways that you
can choose k objects from n
(order of selection irrelevant)

Application to Lotto 6/49

Here you choose 6 numbers from the integers 1, 2, 3, ..., 47, 48, 49. Six winning numbers are chosen together with a bonus number.

How many choices for the 6 winning numbers.

$$\begin{pmatrix} 49 \\ 6 \end{pmatrix} = {}^{49}C_6 = \frac{49!}{6!43!} = \frac{49(48)(47)(46)(45)(44)}{6(5)(4)(3)(2)(1)} = 13,983,816$$

You can lose and win in several ways

1. No winning numbers - lose
2. One winning number - lose
3. Two winning numbers - lose
4. Two + bonus - win \$5.00
5. Three winning numbers - win \$10.00
6. Four winning numbers - win approx. \$80.00
7. 5 winning numbers - win approx. \$2,500.00
8. 5 winning numbers + bonus - win approx. \$100,000.00
9. 6 winning numbers - win approx. \$4,000,000.00

Counting the possibilities

1. No winning numbers - lose

All six of your numbers have to be chosen from the losing numbers and the bonus.

$$\begin{pmatrix} 43 \\ 6 \end{pmatrix} = 6,096,454$$

2. One winning numbers - lose

One number is chosen from the six winning numbers and the remaining five have to be chosen from the losing numbers and the bonus.

$$\begin{pmatrix} 6 \\ 1 \end{pmatrix} \begin{pmatrix} 43 \\ 5 \end{pmatrix} = 6(962,598) = 5,775,588$$

3. Two winning numbers - lose

Two numbers are chosen from the six winning numbers and the remaining four have to be chosen

from the losing numbers (bonus not included)

$$\binom{6}{2} \binom{42}{4} = 15(111,930) = 1,678,950$$

4. Two winning numbers + the bonus - win \$5.00

Two numbers are chosen from the six winning numbers, the bonus number is chose and the remaining three have to be chosen from the losing numbers.

$$\binom{6}{2} \binom{1}{1} \binom{42}{3} = 15(1)(11,480) = 172,200$$

5. Three winning numbers - win \$10.00

Three numbers are chosen from the six winning numbers and the remaining three have to be chosen from the losing numbers + the bonus number

$$\binom{6}{3} \binom{43}{3} = 20(12,341) = 246,820$$

6. four winning numbers - win approx. \$80.00

Four numbers are chosen from the six winning numbers and the remaining two have to be chosen from the losing numbers + the bonus number

$$\binom{6}{4} \binom{43}{2} = 15(903) = 13,545$$

7. five winning numbers (no bonus) - win approx. \$2,500.00

Five numbers are chosen from the six winning numbers and the remaining number has to be chosen from the losing numbers (excluding the bonus number)

$$\binom{6}{5} \binom{42}{1} = 6(42) = 252$$

8. five winning numbers + bonus - win approx. \$100,000.00 Five numbers are chosen from the six winning numbers and the remaining number is chosen to be the bonus number.

$$\binom{6}{5} \binom{1}{1} = 6(1) = 6$$

9. six winning numbers (no bonus) - win approx. \$4,000,000.00

Six numbers are chosen from the six winning numbers,

$$\binom{6}{6} = 1$$

Summary

0 winning	6,096,454	nil	0.4359649755
1 winning	5,775,588	nil	0.4130194505
2 winning	1,678,950	nil	0.1200637937
2 + bonus	172,200	\$ 5.00	0.0123142353
3 winning	246,820	\$ 10.00	0.0176504039
4 winning	13,545	\$ 80.00	0.0009686197
5 winning	252	\$ 2,500.00	0.0000180208
5 + bonus	6	\$ 100,000.00	0.0000004291
6 winning	1	\$ 4,000,000.00	0.0000000715
Total	13,983,816		

4.3 Birthday Paradox

The birthday paradox, also known as the birthday problem, states that in a random group of 23 people, there is about a 50 percent chance that two people have the same birthday. There are multiple reasons why this seems like a paradox. One is that when in a room with 22 other people, if a person compares his or her birthday with the birthdays of the other people it would make for only 22 comparisons—only 22 chances for people to share the same birthday. But when all 23 birthdays are compared against each other, it makes for much more than 22 comparisons. How much more? Well, the first person has 22 comparisons to make, but the second person was already compared to the first person, so there are only 21 comparisons to make. The third person then has 20 comparisons, the fourth person has 19 and so on. If you add up all possible comparisons ($22 + 21 + 20 + 19 + \dots + 1$) the sum is 253 comparisons, or combinations. Consequently, each group of 23 people involves 253 comparisons, or 253 chances for matching birthdays.

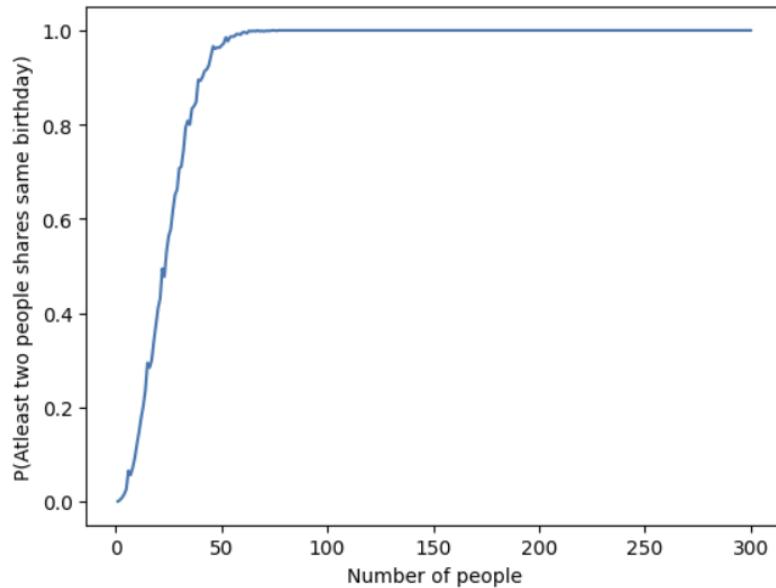
4.3.1 Experiment, Observations and results

Take 10 to 12 groups of 23 or more people so you have enough different groups to compare. Did about 50 percent of the groups of 23 or more people include at least two people with the same birthdays? When comparing probabilities with birthdays, it can be easier to look at the probability that people

do not share a birthday. A person's birthday is one out of 365 possibilities (excluding February 29 birthdays). The probability that a person does not have the same birthday as another person is 364 divided by 365 because there are 364 days that are not a person's birthday. This means that any two people have a $364/365$, or 99.726027 percent, chance of not matching birthdays. As mentioned before, in a group of 23 people, there are 253 comparisons, or combinations, that can be made. So, we're not looking at just one comparison, but at 253 comparisons. Every one of the 253 combinations has the same odds, 99.726027 percent, of not being a match. If you multiply 99.726027 percent by 99.726027 253 times, or calculate $(364/365)^{253}$, you'll find there's a 49.952 percent chance that all 253 comparisons contain no matches. Consequently, the odds that there is a birthday match in those 253 comparisons is $1 - 49.952\% = 50.048\%$, or just over half! The more trials you run, the closer the actual probability should approach 50 percent.

4.3.2 Results of above experiment by simulations

Let n be the number of people in group and the probability that at least two people have the same birthday, denoted by p . For this case, plot p vs n where n varies from 1 to 300 by simulating the situation 1000 times is given below.



4.4 Theory Assignment-3

4.4.1 Questions

- Q1.** Let H = event that person has disease, S = event that test result is positive R_1 = event that positive test result came from Machine 1 R_2 = event that positive test result came from Machine 2

$$P(H)=0.001, P(R1)=0.2, P(R2)=0.8, P(R1|H)=0.89, P(R2|H)=0.99, P(R1|\bar{H})=0.025, P(R2|\bar{H})=0.005$$

Given that person has a disease and test result is positive. What is the probability that the result came from Machine 2 i.e $P(R2|H, S)$?

Q2. We believe there are three types of managers: underperformers, in-line performers, and outperformers. The underperformers (MU) beat the market only 25% of the time, the in-line performers (MI) beat the market 50% of the time, and the outperformers (MO) beat the market 75% of the time.

Initially we believe a given manager is most likely to be an in-line performer, and is less likely to be an underperformer or an outperformer. Specifically, our prior belief is that a manager has a 60% chance of being an in-line performer, a 20% chance of being an underperformer, and a 20% chance of being an outperformer. We can summarize this as:

$$P[\text{MU}]=P[p=0.25]=20\%, P[\text{MI}]=P[p=0.50]=60\%, P[\text{MO}]=P[p=0.75]=20\% \dots \dots (1)$$

a.) By taking the values given in (1), suppose the manager beats the market two years in a row. What should our updated beliefs be?

b.) After updating the beliefs from part (a), what if the manager again beats the market next year, what should be the updates now?

c.) By taking the values given in (1), suppose the manager beats the market three years in a row. What should our updated beliefs be? How is it different from part (b)?

4.4.2 Answers

Q1. Let H = event that person has disease, S = event that test result is positive $R1$ = event that positive test result came from Machine 1 $R2$ = event that positive test result came from Machine 2

$$P(H)=0.001, P(R1)=0.2, P(R2)=0.8, P(R1|H)=0.89, P(R2|H)=0.99, P(R1|\bar{H})=0.025, P(R2|\bar{H})=0.005$$

Given that person has a disease and test result is positive. What is the probability that the result came from Machine 2 i.e $P(R2|H, S)$?

Solution: We want to find the probability that the result came from Machine 2 given that person has a disease and test result is positive which means

$$P(R2|H, S) = P(H|R2, S)P(R2|S)P(H|S) \text{ (Bayes formula for 3 variable)}$$

$$\begin{aligned} P(S) &= P(\text{Test result is positive}) = P(R1) * (P(H) * P(R1|H) + P(\bar{H}) * P(R1|\bar{H})) + P(R2) * \\ &(P(H) * P(R2|H) + P(\bar{H}) * P(R2|\bar{H})) = 0.2 * (0.001 * 0.89 + 0.999 * 0.025) + 0.8 * (0.001 * 0.99 + \\ &0.999 * 0.005) = 0.0096 \end{aligned}$$

$$\begin{aligned} P(H|S) &= P(\text{Person has disease given that test result is positive}) = \frac{P(H)*(P(R1)*P(R1|H)+P(R2)*P(R2|H))}{P(S)} \\ &= 0.001 * (0.2 * 0.89 + 0.8 * 0.99) / 0.0096 = 0.00097 / 0.0096 = 0.097 \end{aligned}$$

$$\begin{aligned} P(R2|S) &= P(\text{Positive test result came from Machine 1 given that test result is positive}) = \\ &\frac{P(R2)*(P(H)*P(R2|H)+P(\bar{H})*P(R2|\bar{H}))}{P(S)} = 0.8 * (0.001 * 0.99 + 0.999 * 0.005) / 0.0096 = 0.00478 / 0.0096 = \\ &0.48 \end{aligned}$$

$$P(H|R2, S) = P(\text{Person has disease given that test result is positive and came from Machine 1}) =$$

$$\frac{P(H)*P(R2)*P(R2|H)}{P(H)*P(R2)*P(R2|H)+P(\bar{H})*P(R2)*P(R2|\bar{H})} = 0.001*0.8*0.99/(0.001*0.8*0.99+0.999*0.8*0.005) = 0.000792/0.00478 = 0.165$$

$P(R2|H, S) = P(H|R2, S)P(R2|S)P(H|S) = 0.165*0.48/0.097 = 0.816$ Now, for the second run,

$$P(R2) = 0.8164, P(R1) = 1 - 0.8164 = 0.1836 \quad P(R2|H, S) = 0.8319$$

For third run,

$$P(R2) = 0.8319, P(R1) = 1 - 0.8319 = 0.1684 \quad P(R2|H, S) = 0.8463 \text{ and so on...}$$

So, the probability that the result came from Machine 2 will increase with increase in simulations because of the fact that it is giving better results than Machine 1.

Q2. We believe there are three types of managers: underperformers, in-line performers, and outperformers. The underperformers (MU) beat the market only 25% of the time, the in-line performers (MI) beat the market 50% of the time, and the outperformers (MO) beat the market 75% of the time.

Initially we believe a given manager is most likely to be an in-line performer, and is less likely to be an underperformer or an outperformer. Specifically, our prior belief is that a manager has a 60% chance of being an in-line performer, a 20% chance of being an underperformer, and a 20% chance of being an outperformer. We can summarize this as:

$$P[MU] = P[p=0.25] = 20\%, P[MI] = P[p=0.50] = 60\%, P[MO] = P[p=0.75] = 20\% \dots \dots (1)$$

a.) By taking the values given in (1), suppose the manager beats the market two years in a row. What should our updated beliefs be?

b.) After updating the beliefs from part (a), what if the manager again beats the market next year, what should be the updates now?

c.) By taking the values given in (1), suppose the manager beats the market three years in a row. What should our updated beliefs be? How is it different from part (b)?

Solution: Given, $P[MU] = 20\%$, $P[MI] = 60\%$, $P[MO] = 20\%$ $P[2B] = P[\text{Manager beats market two times in a row}] = P(\text{Manager is underperformer}) (P(\text{underperformer beats the market}))^2 + P(\text{Manager is in-line performer}) (P(\text{in-line performer beats the market}))^2 + P(\text{Manager is outperformer}) (P(\text{outperformer beats the market}))^2 = 0.2 * (0.25)^2 + 0.6 * (0.50)^2 + 0.2 * (0.75)^2 = 0.275$ $P[MU|2B] = 0.2 * (0.25)^2 / 0.275 = 0.0125 / 0.275 = 0.0454 = 4.54\%$ $P[MI|2B] = 0.6 * (0.50)^2 / 0.275 = 0.15 / 0.275 = 0.5454 = 54.54\%$ $P[MO|2B] = 0.2 * (0.75)^2 / 0.275 = 0.1125 / 0.275 = 0.4091 = 40.91\%$

We should update our belief that a manager has a 54.54% chance of being an in-line performer, a 4.54% chance of being an underperformer, and a 40.91% chance of being an outperformer.

b.) For the next run,

$$P[MU] = 4.54\%, P[MI] = 54.54\%, P[MO] = 50.91\%$$

$$P[B] = 0.045 * (0.25) + 0.5454 * (0.50) + 0.4091 * (0.75) = 0.5908$$

$$P[MU|B] = 0.045 * (0.25) / 0.5908 = 0.01125 / 0.5908 = 0.01904 = 1.90\%$$

$$P[MI|B] = 0.5454 * (0.50) / 0.5908 = 0.2727 / 0.5908 = 0.4616 = 46.16\%$$

$$P[MO|B] = 0.4091 * (0.75) / 0.5908 = 0.3068 / 0.5908 = 0.5193 = 51.93\%$$

We should update our belief that a manager has a 46.16% chance of being an in-line performer, a 1.90% chance of being an underperformer, and a 51.93% chance of being an outperformer.

So, with an increase in the number of simulations, the chance of being outperformer will increase while others decrease because when the manager beats the market again and again then, there is a high chance that he's an out performer.

c.) Given, $P[\text{MU}] = 20\%$, $P[\text{MI}] = 60\%$, $P[\text{MO}] = 20\%$

$$P[3B] = P[\text{Manager beats market three times in a row}] = 0.2 * (0.25)^3 + 0.6 * (0.50)^3 + 0.2 * (0.75)^3 \\ = 0.1625$$

$$P[\text{MU}|3B] = 0.2 * (0.25)^3 / 0.1625 = 0.003125 / 0.1625 = 0.0192 = 1.92\% \quad P[\text{MI}|3B] = 0.6 * (0.50)^3 / 0.1625 = 0.075 / 0.1625 = 0.4615 = 46.15\% \quad P[\text{MO}|3B] = 0.2 * (0.75)^3 / 0.1625 = 0.084375 / 0.1625 = 0.5192 = 51.92\%$$

We should update our belief that a manager has a 46.15% chance of being an in-line performer, a 1.92% chance of being an underperformer, and a 51.92% chance of being an outperformer. We can see there is a slight difference in probabilities when we have done it for 2 years, updated the information, then done it for 1 year and when we have done it for 3 years directly.

4.5 Lab Assignment 3

Q1. The given dataset ‘data.csv’ consists of some weather data over the past 100 days for a region in Himachal Pradesh. It consists of columns - Day, Cloudy, Rain and Snow. The column ‘Day’ has days numbered from 1-100. If a given day was cloudy it is denoted by ‘1’ else ‘0’. If it rained greater than 0.01 inches on a particular day, it is represented as ‘1’ else ‘0’. Similarly, if it snows, it is represented by ‘1’ else ‘0’. With the given information, calculate the following: Probability that a day is: (i) cloudy (ii) raining (iii) snowing Probability that it will rain given that it is a cloudy day and also the Probability that it is cloudy given it is raining. Using the values obtained, verify the formula: $P(A|B) = P(A)P(B|A)P(B)$ Probability of a sun shower i.e. it is raining given it is not cloudy. Probability that it will either rain or snow, given it is a cloudy day. Probability that it will both rain and snow, given it is a cloudy day.

```

1 import pandas as pd
2 df = pd.read\_csv( data . c s v )
3 cloudy = df[ Cloudy ].to\_list()
4 rain = df[ Rain ].to\_list()
5 snow = df[ Snow ].to\_list()
6
7 p\_cloudy = cloudy.count(1)/len(cloudy)
8 p\_rain = rain.count(1)/len(rain)
9 p\_snow = snow.count(1)/len(snow)
10
11 print( A )

```

```

12 print(p\_cloudy, p\_rain, p\_snow)
13
14 rain\_and\_cloudy = 0
15 for i in range(len(rain)):
16     if rain[i] and cloudy[i]:
17         rain\_and\_cloudy+=1
18
19 ric = rain\_and\_cloudy/len(rain)
20
21 p\_rc = ric/p\_cloudy
22 p\_cr = ric/p\_rain
23
24 print(    B    )
25 print(p\_rc)
26 print(p\_cr)
27
28 if p\_rc == (p\_rain * p\_cr)/p\_cloudy :
29     print(    Verified    )
30 else:
31     print( There is some problem )
32
33
34 rainy\_and\_not\_cloudy = 0
35 for i in range(len(rain)):
36     if rain[i] and cloudy[i]==0 :
37         rainy\_and\_not\_cloudy+=1
38
39 rinc = rainy\_and\_not\_cloudy/len(rain)
40 p\_rnc = rinc/(1-p\_cloudy)
41
42 print(    C    )
43 print(p\_rnc)
44
45 snow\_or\_rain\_and\_cloudy = 0
46 for i in range(len(rain)):
47     if (snow[i] or rain[i]) and cloudy[i]:
48         snow\_or\_rain\_and\_cloudy+=1
49 sur\_ic = snow\_or\_rain\_and\_cloudy/len(rain)
50 p\_s\_rc = sur\_ic/p\_cloudy
51 print(    D    )
52 print(p\_s\_rc)
53
54 snow\_and\_rain\_and\_cloudy = 0
55 for i in range(len(rain)):
56     if snow[i] and rain[i] and cloudy[i]:
57         snow\_and\_rain\_and\_cloudy+=1

```

```

58
59 sric = snow\_and\_rain\_and\_cloudy/len(rain)
60
61 p\_src = sric/p\_cloudy
62 print( E )
63 print(p\_src)
64
65 Code Answer: A) 0.43 0.31 0.06
66 B) 0.5813953488372093~~0.8064516129032259~~Verified
67 C) 0.10526315789473682
68 D) 0.627906976744186
69 E) 0.06976744186046512

```

Q2. Suppose that a laboratory test to detect a certain disease has the following statistics. A = event that the person has the disease B = event that the test result is positive It is known that $P(B|A) = 0.95$ and $P(B|\bar{A}) = 0.05$, and 1% of the population actually has the disease. (Note: The question is similar q1 of the previous assignment with minor modifications) a) Theoretically calculate the probability that a person tested positive actually has the disease. b) Simulate the experiment in python. Take the size of Population(N) as 1000, 10000, 100000. Calculate the probability(P) experimentally for the given values of N. c) Plot the Graph of P v/s N (domain of N should from 1 to 100000) [Hint: Make two lists, one of population(healthy/diseased) and second of test_results(positive/negative). Use random.random() for sampling]

```

1 import random
2 import matplotlib.pyplot as plt
3 from sklearn.metrics import confusion_matrix
4 N = int(input("Enter the value of N: "))
5 def Simulation(N):
6     p_A = 0.01
7
8     p_B_given_A = 0.95
9     p_B_given_not_A = 0.05
10
11
12
13
14 population = []
15 for i in range(N):
16     if random.random() < p_A:
17         population.append(1)
18     else:
19         population.append(0)
20
21 test_results = []
22 for patient in population:

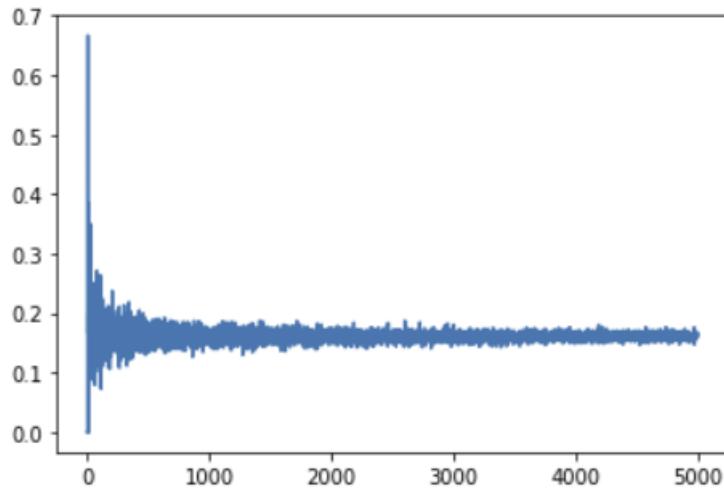
```

```

23     ~~~~ if patient == 1:
24         ~~~~ if random.random() < p\_B\_given\_A:
25             ~~~~~ test\_results.append(1)
26         ~~~~ else:
27             ~~~~~ test\_results.append(0)
28     ~~~~ else:
29         ~~~~ if random.random() < (p\_B\_given\_not\_A):
30             ~~~~~ test\_results.append(1)
31         ~~~~ else:
32             ~~~~~ test\_results.append(0)
33
34
35 ~`positive\_tests = [population[i] for i in range(N) if test\_results[i] == 1]
36 ~`positive\_tests = []
37 ~`for i in range (N):
38     ~~~~ if(population[i]==1 and test\_results[i]==1):
39         ~~~~~ positive\_tests.append(1)
40     ~~~~ elif(population[i]==0 and test\_results[i]==1):
41         ~~~~~ positive\_tests.append(0)
42 ~`p\_A\_given\_B = sum(positive\_tests) / max(1, len(positive\_tests))
43 ~`prevalance = sum(population) / N * 100
44 ~`accuracy = (sum(positive\_tests) + (N - max(1, len(positive\_tests)))) / N * 100
45
46 ~`return p\_A\_given\_B, prevalance, accuracy
47
48 ans = Simulation(N)
49
50 print("Simulated prevalence of the disease: {:.2f}%.format(ans[1]))
51 print("Simulated accuracy of the test: {:.2f}%.format(ans[2]))
52 print("Simulated posterior probability of having the disease given a positive test
      result: {:.2f}%.format(ans[0] * 100))
53 p = []
54 for N in range (2, 100000, 20):
55     ~`p.append(Simulation(N)[0])
56
57 plt.plot(p)
58 plt.show()
59
60 \textbf{Code Answer} Enter the value of N: 100000
61 Simulated prevalence of the disease: 0.99%
62 Simulated accuracy of the test: 95.02%
63 Simulated posterior probability of having the disease given a positive test result:
      15.94%

```

Q3. Let H = event that person has disease, S = event that test result is positive R1 = event that positive test result came from Machine 1 R2 = event that positive test result came from Machine 2



$$P(H)=0.001, P(R1)=0.2, P(R2)=0.8, P(R1|H)=0.89, P(R2|H)=0.99, P(R1|\bar{H})=0.025, P(R2|\bar{H})=0.005$$

a) Given that person has a disease and test result is positive. What is the probability that the result came from Machine 2 i.e $P(R2|H, S)$? b) Now This would be the updated prior Probability for R2 i.e $P(R2) = P(R2|H, S)$, similarly the prior probability of R1 would be updated $P(R1) = 1 - P(R2|H, S)$: Using the new beliefs: Calculate the updated probability that the result came from Machine 2 if the test result for a person having disease comes positive 10 times in a row.

```

1 def funct(p\_r1, p\_r2, simulations):
2     for i in range(simulations):
3         P\_s = 0.001*(p\_r1*0.89+p\_r2*0.99)+0.999*(p\_r1*0.025+p\_r2*0.005)
4         P\_h\_given\_s = 0.001*(p\_r1*0.89+p\_r2*0.99)/P\_s
5         P\_r2\_given\_s = p\_r2*(0.001*0.99+0.999*0.005)/P\_s
6         P\_h\_given\_r2s = 0.001*p\_r2*0.99/(p\_r2*(0.001*0.99+0.999*0.005))
7         P\_r2\_given\_hs = P\_h\_given\_r2s*P\_r2\_given\_s/P\_h\_given\_s
8         p\_r1 = 1 - P\_r2\_given\_hs
9         p\_r2 = P\_r2\_given\_hs
10        print("probability that the result came from Machine 2 for", i+1, "times is",
11             P\_r2\_given\_hs)
12
13
14 Code Answer:
15 probability that the result came from Machine 2 for 1 times is 0.8164948453608247
16 probability that the result came from Machine 2 for 2 times is 0.8319151193633951
17 probability that the result came from Machine 2 for 3 times is 0.8462835506354438
18 probability that the result came from Machine 2 for 4 times is 0.8596309667949942
19 probability that the result came from Machine 2 for 5 times is 0.8719947096591083
20 probability that the result came from Machine 2 for 6 times is 0.8834171407289191
21 probability that the result came from Machine 2 for 7 times is 0.8939442699225116
22 probability that the result came from Machine 2 for 8 times is 0.9036245284151371

```

```

23 probability that the result came from Machine 2 for 9 times is 0.9125076960445421
24 probability that the result came from Machine 2 for 10 times is 0.9206439852763159

```

Q4. We believe there are three types of managers: underperformers, in-line performers, and outperformers. The underperformers (MU) beat the market only 25% of the time, the in-line performers (MI) beat the market 50% of the time, and the outperformers (MO) beat the market 75% of the time.

Initially we believe a given manager is most likely to be an in-line performer, and is less likely to be an underperformer or an outperformer. Specifically, our prior belief is that a manager has a 60% chance of being an in-line performer, a 20% chance of being an underperformer, and a 20% chance of being an outperformer. We can summarize this as:

$$P[\text{MU}] = P[p=0.25] = 20\%, \quad P[\text{MI}] = P[p=0.50] = 60\%, \quad P[\text{MO}] = P[p=0.75] = 20\% \dots \dots (1)$$

a.) By taking the values given in (1), suppose the manager beats the market two years in a row. What should our updated beliefs be?

b.) After updating the beliefs from part (a), what if the manager again beats the market next year, what should be the updates now?

c.) By taking the values given in (1), suppose the manager beats the market three years in a row.

What should our updated beliefs be? How is it different from part (b)?

For the lab- Simulate part (b) for the next 10 years.

```

1 def performer(n):
2     p_beats_n_times = 0.2 * (0.25)**n + 0.6 * (0.50)**n + 0.2 * (0.75)**n
3     p_mu_given_beats_n_times = 0.2 * (0.25)**n / p_beats_n_times
4     p_mi_given_beats_n_times = 0.6 * (0.50)**n / p_beats_n_times
5     p_mo_given_beats_n_times = 0.2 * (0.75)**n / p_beats_n_times
6     return [p_mu_given_beats_n_times, p_mi_given_beats_n_times, p_mo_given_beats_n_times]
7 print("Updated information of probability when beats 2 times in a row for MU, MI, MO
      are", performer(2), "respectively.") \#A4 (a)
8 print("Updated information of probability when beats 3 times in a row for MU, MI, MO
      are", performer(3), "respectively.") \#A4 (c)
9 Code Answer:
10 Updated information of probability when beats 2 times in a row for MU, MI, MO are
      [0.045454545454545456, 0.5454545454545454, 0.40909090909090906] respectively.
11 Updated information of probability when beats 3 times in a row for MU, MI, MO are
      [0.019230769230769232, 0.4615384615384615, 0.5192307692307693] respectively.

```

```

1 #4(b)
2 def perf(p_mu, p_mi, p_mo, simulations):
3     for i in range(simulations):
4         p_beats = p_mu * (0.25) + p_mi * (0.50) + p_mo * (0.75)
5         p_mu_given_beats = p_mu * (0.25) / p_beats
6         p_mi_given_beats = p_mi * (0.50) / p_beats
7         p_mo_given_beats = p_mo * (0.75) / p_beats
8         p_mu = p_mu_given_beats

```

```

9     p\_mi = p\_mi\_given\_beats
10    p\_mo = p\_mo\_given\_beats
11    print("Updated information of probability when beats", i+1, " times for MU, MI, MO
12      are", p\_mu\_given\_beats, p\_mi\_given\_beats, p\_mo\_given\_beats)
13 perf(0.0454, 0.5454, 0.409, 10)    \#After updated probabilities of part A4 (a)
14 Code Answer:
15 Updated information of probability when beats 1 times for MU, MI, MO are
16   0.01921123899796886 0.46157752200406227 0.5192112389979688
17 Updated information of probability when beats 2 times for MU, MI, MO are
18   0.007684495599187544 0.36926201760324984 0.6230534867975626
19 Updated information of probability when beats 3 times for MU, MI, MO are
20   0.002938207046519455 0.2823785239069093 0.7146832690465712
21 Updated information of probability when beats 4 times for MU, MI, MO are
22   0.0010835115319993893 0.20826332671452574 0.7906531617534748
23 Updated information of probability when beats 5 times for MU, MI, MO are
24   0.00038841529979843486 0.14931573886172075 0.8502958458384808
25 Updated information of probability when beats 6 times for MU, MI, MO are
26   0.000136290496889935 0.10478637815509502 0.8950773313480151
27 Updated information of probability when beats 7 times for MU, MI, MO are
28   4.707885064556099e-05 0.07239275458562525 0.9275601665637291
29 Updated information of probability when beats 8 times for MU, MI, MO are
30   1.6081516712308903e-05 0.04945682729102815 0.9505270911922595
31 Updated information of probability when beats 9 times for MU, MI, MO are
32   5.450417456356927e-06 0.03352424520962331 0.9664703043729204
33 Updated information of probability when beats 10 times for MU, MI, MO are
34   1.837344355888585e-06 0.02260215229917342 0.9773960103564707

```

Chapter 5

Discrete and Continuous Random Variables

5.1 Introduction

Till now, we have gone through the basics of set theory and probability theory. Now, we will be heading towards the most important and fundamental entity of the probability theory and statistical inference. In this chapter, we would extend the so far developed theory for outcomes of numerical nature. We would also discuss about the probability density (and also the mass) functions and the cumulative distribution functions which are used to describe these random variables.

5.2 Definition

Consider a random experiment with sample space S . Then, a random variable X is defined as a measurable function $X : S \rightarrow E$, from the sample space S to a measurable space E . For the most practical cases, we can define a random variable as a function $X : S \rightarrow \mathbb{R}$.

In simple terms, we can define a random variable as a function that assigns some numerical value, called as the *value* of X to each sample point in the sample space S .

The sample space S of the random experiment becomes the *domain* of the random variable X and the values taken by it constitutes the *range* of the random variable X . It is important to note that two or more different sample points might take on the same value of the random variable X , but two different numbers in the range can't be assigned to the same sample point.

In practical applications, which we will discuss later, sometimes we incline on numerical aspects of the random experiment rather than the outcome of the random experiment itself. This laid the foundation to the concepts of *random variables*. A point to be noted here is that the concerned variable, random variable, is different from the conventional variable, which we have studied in algebra and calculus. The conventional variable have definite values and could be described by a single or group of numbers. However, random variables don't have any particular number and should be described by using probabilities.

Most of the confusion arises from the fact that both of these variables represent similar things though they aren't identical. To clear the confusion, let us consider a game in which we have a set of finite numbers and the player have to pick a number, which would then denote that player's score. So in this case, the conventional variable denotes what we actually got, which has zero variance. On the other hand, the random variable comes into play before the game starts, i.e., it would be used to describe the outcome to be expected.

For the representation part, we would denote, as per the convention, the random variable by upper case, X , while the lower case, x , to denote a quantity whose value remains constant.

In the next subsection, we would discuss some real-life examples of random variables.

5.3 Examples

Here, we would go through some of the common, yet relatable examples in the context of random variables. The real-life examples are as follows:

1. You are typing a letter in MS word and you have the option of selecting the font size of your text. The font size can be represented as a random variable.
2. Suppose you roll two dice and you are interested in the product of the two numbers obtained on each die rather than the outcome itself. So, here the product of the numbers on the two faces of the die is our random variable.
3. If you flip a coin, then random variable may takes on the values representing the number of heads (or tails) so obtained.

5.4 Types

There are mainly three types of random variables, namely,

1. Discrete Random Variable
2. Continuous Random Variable
3. Mixed Random Variable

The above classification is based on the whether the concerned random variable is countable (both finite and infinite) or uncountable. We would discuss about them in detail in the subsequent sections.

5.5 Discrete Random Variables

A random variable X is said to be *discrete random variable* if X takes on only a finite number of different values or at most, an infinite sequence of different values. More formally, we can define a discrete random variable as follows:

The random variable X , $X : S \rightarrow E$, is a discrete random variable if the range or image of the random variable is countable.

Some of the examples for discrete random variables are as follows:

1. Consider your class of this course IC-252. Suppose the course has a strength of 320 and on a particular day, the logistic TAs mark the attendance of students present in that class. The random variable taking the number of students present would be a discrete random variable type.
2. Consider your own family. Then, the random variable taking on the number of people in the family would be a discrete random variable.
3. Consider the parking lot of our college. Then the random variable taking on the number of vehicles parked there would be a discrete random variable.
4. Consider the number of trees (or even the number of species) on the Griffon peak. The random variable taking on the number of trees (or even the number of species) there, would be a discrete random variable.

In the next sub-section, we would discuss about the probability mass function, which is used to characterize the probabilities that the random variable would take on.

5.5.1 Probability Mass Function

If X is a discrete random variable, then it's range, R_X , is a countable set. So, we can write its range as R_X , given by

$$R_X = \{x_1, x_2, \dots\}$$

where x_1, x_2, \dots are the possible values of the random variable X .

The random variable X , here would assume each of those values with certain probability. Mathematically, it would be useful if we represent all the probabilities of the values or elements in its *range* by a single formula. The entity used to represent all the probabilities is known as the *probability mass function* or *pmf*.

Let us now define the *probability mass function* or *pmf* as follows:

Let X be a discrete random variable with range $R_X = \{x_1, x_2, \dots\}$. Then, the *probability mass function* or *pmf* of X is the set of probability values assigned to each of the values x_i taken by the discrete random variable X .

$$p_X(x_i) = P(X = x_i), \quad \text{for } i = 1, 2, 3, \dots$$

Thus, the *pmf* is a probability measure that gives us the probabilities of the possible values for a random variable. The subscript X , here, indicates that this *pmf* is for the random variable X .

Even though the *pmf* is usually defined for only the values in its range, R_X , it is customary to extend the *pmf* of X to all the real numbers. If $x \notin R_X$, we can simply write

$$p_X(x) = P(X = x) = 0$$

. Thus, in general we can write

$$p_X(x) = \begin{cases} P(X = x), & \text{if } x \in R_X \\ 0, & \text{otherwise} \end{cases}$$

To better visualize the *pmf*, we can plot it.

Now let us see the properties of the *pmf*. The *pmf* must satisfy the following properties:

1. $0 \leq p_X(x) \leq 1, \forall x \in R_X$
2. $p_X(x) = 0, \quad \text{if } x \notin R_X$
3. $\sum_k p_X(x_k) = 1, \quad k = 1, 2, \dots$

Let us consider two examples to get more insights into discrete random variable.

Example 1: Consider tossing a fair coin twice, and the random variable X denote the number of heads observed while tossing the coin twice. Find the range of X , R_X and the probability mass function of X .

Solution: The sample space of tossing a coin twice is S , given by

$$S = \{HH, HT, TH, TT\}$$

If we toss a coin twice, then the possible values for the number of heads is either 0, 1 or 2. Thus, the range of X , R_X , is given by

$$R_X = \{0, 1, 2\}.$$

Since, R_X is a finite set, the random variable X is a discrete random variable. It's *pmf* is given by

$$p_X(x) = P(X = x), \text{ for } x = 0, 1, 2$$

Now,

$$p_X(0) = P(X = 0) = P(TT) = \frac{1}{4}$$

$$p_X(1) = P(X = 1) = P(\{HT, TH\}) = \frac{2}{4}$$

$$p_X(2) = P(X = 2) = P(HH) = \frac{1}{4}$$

So, the *pmf* of X is given by

$$p_X(x) = \begin{cases} \frac{1}{4} & x = 0 \\ \frac{2}{4} & x = 1 \\ \frac{1}{4} & x = 2 \end{cases}$$

For visualization, the *pmf* is as shown in the figure below:

Example 2: Consider an unfair coin for which $P(\text{Head})=p$, where $0 < p < 1$. If we toss the coin repeatedly until we observe a head for the first time and Y be the total number of coin tosses. Find the distribution of Y .

Solution: Since, Y is total number of coin tosses required to get a head for the first time, it can possibly take any positive number. So, the range of Y , R_Y would be given as

$$R_Y = \{1, 2, 3, \dots\}.$$

The *pmf* of Y is given by

$$p_Y(y) = P(Y = y)$$

So now,

$$\begin{aligned} p_Y(1) &= p(Y = 1) = P(H) = p \\ p_Y(2) &= p(Y = 2) = P(TH) = (1 - p)p \\ p_Y(3) &= p(Y = 3) = P(TTH) = (1 - p)^2 p \\ &\vdots \\ p_Y(k) &= p(Y = k) = P(T \dots TH) = (1 - p)^{k-1} p \end{aligned}$$

Hence, the *pmf* of the random variable Y is given by

$$p_Y(y) = \begin{cases} (1 - p)^{y-1} p & y = \{1, 2, 3, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

5.5.2 Cumulative Distribution Function

In the previous sub-section, we have discussed about the *probability mass function*, which characterizes the distribution of a discrete random variable. Now, in this sub-section, we would discuss about the *cumulative distribution function* or *cdf*, which is also used to describe the distribution of a random variable. It can be described for both the discrete and continuous random variables.

The *cumulative distribution function* or *cdf* of a random variable X is the function defined as

$$F(x) = P(X \leq x)$$

Like *pmf*, *cdf* too describes the probabilistic characteristics of the random variable X . Knowing either of these functions, we can get knowledge of the other distribution.

For a discrete random variable, the *cumulative distribution function* or *cdf* is defined as follows:

$$F(x) = \sum_{y:y \leq x} P(X = y)$$

In simple words, the *cumulative distribution function* or *cdf*, $F(x)$, is the sum of probabilities $P(y)$ such that y isn't greater than x .

The *cumulative distribution function* or *cdf* of a random variable X is an increasing step function with steps at the values taken by the random variable. The heights of the steps are the probabilities

of taking these values. The *probability mass function* is related to the *cumulative distribution function* as

$$P(X = x) = F(x) - F(x^-)$$

where $F(x^-)$ is the limiting value from below of the cumulative distribution function.

Now, let us see the properties of the *cumulative distribution function* or *cdf*, which are as follows:

1. $0 \leq F(x) \leq 1$
2. $F(x_1) \leq F(x_2)$ if $x_1 < x_2$
3. $\lim_{x \rightarrow \infty} F(x) = F(\infty) = 1$
4. $\lim_{x \rightarrow -\infty} F(x) = F(-\infty) = 0$
5. $\lim_{x \rightarrow a^+} F(x) = F(a^+) = F(a)$, i.e., it is right continuous

5.5.3 Functions of Random Variables

In this subsection, we would discuss about the functions of random variables.

Let us consider a random variable X , then its function, $Y = g(X)$, is a random variable itself. Hence, we can comment about its *pmf*, *cdf* and other related properties.

Note that the range of the random variable Y is given as

$$R_Y = \{g(x) | x \in R_X\}.$$

Now, to know the *pmf* of the random variable Y , we first need to know the *pmf* of the random variable X . Let the *pmf* of the random variable X is denoted as $P_X(x)$. Then for $Y = g(X)$, the *pmf* of the random variable Y is given as

$$P_Y(y) = P(Y = y) = P(g(X) = y) = \sum_{x:g(x)=y} P_X(x)$$

Let us go through some examples to get better understanding of it.

Example 1: Let X be a discrete random variable with $P_X(k) = \frac{1}{5}$, for $k = -1, 0, 1, 2, 3$. Let $Y = 2|X|$. Find the range and *pmf* of Y .

Solution: Given, $Y = 2|X|$, where the *pmf* of X is given as

$$P_X(k) = \frac{1}{5}, \text{ for } k = -1, 0, 1, 2, 3$$

Then, the range of Y would be

$$R_Y = 2|x| \quad \forall x \in R_X$$

$$\Rightarrow R_Y = \{0, 2, 4, 6\}.$$

Now, we would find the *pmf* of Y .

$$P_Y(y) = P(Y = y) \quad \forall y \in R_Y$$

$$P_Y(0) = P(Y = 0) = P(2|X| = 0) = P(X = 0) = \frac{1}{5}$$

$$P_Y(2) = P(Y = 2) = P(2|X| = 2) = P(X = -1) + P(X = 1) = \frac{2}{5}$$

$$P_Y(4) = P(Y = 4) = P(2|X| = 4) = P(X = 2) = \frac{1}{5}$$

$$P_Y(6) = P(Y = 6) = P(2|X| = 6) = P(X = 3) = \frac{1}{5}$$

We can summarize the *pmf* of Y as follows:

$$P_Y(y) = \begin{cases} \frac{1}{5} & y = 0 \\ \frac{2}{5} & y = 2 \\ \frac{1}{5} & y = 4 \\ \frac{1}{5} & y = 6 \end{cases}$$

The random variable Y satisfies all the required properties.

Example 2: Let X be a random variable with given *pmf*, $P_X(x)$ and *cdf*, $F_X(x)$. Let Y be another random variable such that $Y = a + bx$, where a and b are some constants and $b > 0$. Find the *cdf* of Y in terms of $F_X(x)$.

Solution Given, a random variable X with *pmf*, $P_X(x)$ and *cdf*, $F_X(x)$. And also Y is another random variable such that $Y = a + bx$, where a and b are some constants and $b > 0$.

The *pmf* of Y is given as

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(a + bx \leq y) = P(bx \leq y - a) = P(x \leq \frac{y-a}{b}) \\ &\Rightarrow F_Y(y) = F_X\left(\frac{y-a}{b}\right) \end{aligned}$$

where $b > 0$.

5.5.4 Special Distributions

In this sub-section, we would discuss about some of the distributions which we encounter frequently in our day-to-day life. So these are given special names. Each of these distributions describe a random experiment, which models many of our day-to-day activities.

5.5.4.1 Bernoulli Distribution

It is the simplest discrete distribution. A Bernoulli random variable takes only two possible values, usually 0 and 1. This describes the random experiments that have only two possible outcomes, usually referred to as “success” and “failure.”

Formally, it is defined as follows:

A random variable X is said to be a *Bernoulli random variable* with parameter p , $X \sim \text{Bernoulli}(p)$, if its *pmf* is given by

$$P_X(x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $0 \leq p \leq 1$, is known as the success probability of the experiment.

Some of the cases, where this could be used are

1. In an exam, either we pass ($X = 1$) or fail ($X = 0$) with certain probabilities.
2. Tossing a coin once, we get either a head or tail. We can assign 1 to either of them.
3. The pen we are using to write the exam would either work ($X = 1$) or not ($X = 0$).

The *expectation* of this random variable is given by

$$E[X] = 0 * P(X = 0) + 1 * P(X = 1) = 0 * (1 - p) + 1 * p = p.$$

Also, $E[X^2] = 0^2 * P(X = 0) + 1^2 * P(X = 1) = p$. So, the variance becomes

$$Var(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1 - p)$$

5.5.4.2 Binomial Distribution

In the previous subsection, we have discussed the Bernoulli distribution, which consists of only a single trial known as *Bernoulli trial*. However, in practical applications, it is useful if we have distribution consisting of a sequence of Bernoulli trials. In such cases, the random variable of interest would be the number of *successes*, within a finite number of trials, with success being well defined. Such a distribution is known as the Binomial distribution. It is the most important discrete random variable.

Consider that there are n independent Bernoulli trials X_1, X_2, \dots, X_n , each of having a probability p of registering a 1, then the random variable

$$X = X_1 + X_2 + \dots + X_n$$

is known as the *Binomial distribution*, denoted as $X \sim Bin(n, p)$. The random variable X can take on values ranging from 0 to n and counts the number of 1's (success) in the n trials.

The *probability mass function* of a Binomial random variable X is given as

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad \text{for } x=0, 1, \dots, n.$$

The corresponding *cdf* is given as

$$F_X(x) = \sum_{k=0}^n \binom{n}{k} p^k (1 - p)^{n-k} \quad n \leq x < n + 1$$

There are some properties which have to be satisfied by the random variable in order to say that it follows a Binomial distribution and they are as follows:

1. The experiment must consist of n trials.
2. these trials must be independent.
3. each of them must have a constant probability p of success.

Some of the practical applications are as follows:

1. In an exam, for say, IC-252 mid-semester, the number of students, if each of them has the same probability of clearing the exam, can be modelled as a Binomial random variable.
2. Tossing a coin n times and the number of heads we get can be modelled using this.

The expectation of a Binomial random variable, $X \sim Bin(n, p)$ is np and the variance is npq .

5.5.4.3 Hypergeometric distribution

In the previous subsection, we have studied the Binomial distribution. One of the criteria to be met for Binomial distribution is that the success probability p must remain constant. However, in practice, we often deal with situations where the success probability has to be changed (one of the possibility is that replacement isn't allowed). In this case, the hypergeometric distribution comes into picture.

Now, we would define it formally with greater details. Consider a collection of N items of which r are of a special kind. If one of the items is chosen at random, then the probability that it is of the special kind is $\frac{r}{N}$. If n of these items are chosen *with replacement*, then the distribution of the random variable X , taking on the success count, follows Binomial distribution, $X \sim Bin(n, \frac{r}{N})$. However, if the replacement isn't allowed then X won't follow the Binomial distribution as the success probability changes. The appropriate distribution here is the hypergeometric distribution. The *probability mass function* of X is given as

$$P(X = x) = \frac{\binom{r}{x} \binom{N-r}{n-x}}{\binom{N}{n}}$$

for $\max\{0, n + r - N\} \leq x \leq \min\{n, r\}$.

The expected value and variance of the random variable X are as follows:

$$E[X] = \frac{nr}{N} \quad \text{and,}$$

$$Var(X) = \left(\frac{N-n}{N-1}\right) * n * \frac{r}{N} * \left(1 - \frac{r}{N}\right)$$

One of the practical application is that if we have N number of computer chips, out of which r are defective and we choose n chips at random without replacement. Then, to find the number of defective chips can be modelled by using hypergeometric distribution.

5.5.4.4 Poisson Distribution

Till now, no distribution had considered the time constraint. However, we encounter events, on a daily basis, which occur within a specified time period. For say, consider, we might be interested in finding the number of telephone calls received by someone within a specified time period or number of balls thrown by a player within a specified period. So an appropriate distribution for these kinds of cases is the Poisson distribution.

A

5.6 Practice Problem Set-1

1. A floor in charge supervises the operation of three machines, namely X, Y, and Z. At any given time, each of them can be classified as either working, denoted by 1 or being idle, denoted by 0. The notation (0, 1, 0) is used to represent the situation where machine Y is working but machines X and Z are both idle. Then give the sample space for the status of the three machines at a particular point in time.
2. Two fair dice are thrown, one red and one blue. What is the probability that the red die has a score that is strictly greater than the score of the blue die?
3. A die is loaded in such a way that an even number is twice as likely to occur as an odd number. If E is the event that a number less than 4 occurs on a single toss of the die, find the probability of the event E, $P(E)$.
4. If the probabilities are, respectively, 0.09, 0.15, 0.21, and 0.23 that a person purchasing a new automobile will choose the colour green, white, red, or blue, what is the probability that a given buyer will purchase a new automobile that comes in one of those colours?
5. If the probabilities that an automobile mechanic will service 3, 4, 5, 6, 7, or 8 or more cars on any given workday are, respectively, 0.12, 0.19, 0.28, 0.24, 0.10, and 0.07, what is the probability that he will service at least 5 cars on his next day at work?
6. When rolling a fair die, an event A is defined as the event of getting an even number and event B is defined as the event of getting a high score and they are given as $A = \{2,4,6\}$ and, $B = \{4,5,6\}$. Then find:
 - (a) Probability of their intersection, $P(A \cap B)$.
 - (b) Probability of their union, $P(A \cup B)$.

5.7 Practice Problem Set-1 Solutions

1. Given, there are three machines, namely X, Y, and Z. Each of them can be either in working or idle state. The notation (0,1,0) is used to represent the situation where machine Y is working but machines X and Z are both idle.

The sample space of an experiment contains all the possible outcomes of it. Since, there are only machines and each of them can be in either two states represented by 1 and 0, the sample space would be

$$S = \{(0, 0, 0) (0, 0, 1) (0, 1, 0) (0, 1, 1) (1, 0, 0) (1, 0, 1) (1, 1, 0) (1, 1, 1)\}.$$

2. Without loss of generality, we can consider that the first die is a red die and second one is a blue die.

Since there are two fair dice, the total possible outcomes of it are 36, denoted by $n(S) = 6 \times 6 = 36$.

Let E be the event of getting a greater number on the red die and $\{(a,b)\}$ represent an event, where ' a ' represents the number that appeared on the red die and ' b ' represents the number that appeared on the blue die. Then, the possible outcomes of E are given as

$$E = \{(2,1), (3,1), (3,2), (4,1), (4,2), (4,3), (5,1), (5,2), (5,3), (5,4), (6,1), (6,2), (6,3), (6,4), (6,5)\}$$

Total number of possible outcomes of E , $n(E) = 15$.

$$\text{Then, probability of the event } E, P(E) = \frac{n(E)}{n(S)} = \frac{15}{36} = \frac{5}{12}.$$

Therefore, the probability that the red die has a score that is strictly greater than the score of the blue die is $\frac{5}{12}$.

3. Given, that an even number is twice as likely to occur as an odd number and E is the event that a number less than 4 occurring on a single toss of the die.

Let us consider that if the probability of an odd number is x , then the probability of an even number would be $2x$. Thus, we get

$$P(1) = P(3) = P(5) = x \quad \text{and,}$$

$$P(2) = P(4) = P(6) = 2x$$

But we know that the total probability would be 1. So we get,

$$P(1) + P(3) + P(5) + P(2) + P(4) + P(6) = 1$$

$$\Rightarrow x + x + x + 2x + 2x + 2x = 1$$

$$\Rightarrow 9x = 1$$

$$\Rightarrow x = \frac{1}{9}.$$

$$\Rightarrow P(1) = P(3) = P(5) = \frac{1}{9} \text{ and } P(2) = P(4) = P(6) = 2x = \frac{2}{9}.$$

The possible outcomes of the event $E = \{1,2,3\}$. Then,

$$P(E) = P(1) + P(2) + P(3) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9}.$$

Thus, the probability of the event E is $\frac{4}{9}$.

4. Given, the probabilities that a person purchasing a new automobile will choose the colour green, white, red, or blue are 0.09, 0.15, 0.21, and 0.23, respectively.

Let E be an event that a given buyer will purchase a new automobile that comes in one of these colours. Then,

$$P(E) = P(\text{green}) + P(\text{white}) + P(\text{red}) + P(\text{blue}) = 0.09 + 0.15 + 0.21 + 0.23 = 0.68.$$

Therefore, the probability that a given buyer will purchase a new automobile that comes in one of those colours is 0.68.

5. Given, the probabilities that an automobile mechanic will service 3, 4, 5, 6, 7, or 8 or more cars on any given workday are, respectively, 0.12, 0.19, 0.28, 0.24, 0.10, and 0.07.

Let E be an event that he will service at least 5 cars on his next day at work. Then,

$$P(E) = P(5) + P(6) + P(7) + P(8 \text{ or more}) = 0.28 + 0.24 + 0.10 + 0.07 = 0.69.$$

Therefore, the probability that he will service at least 5 cars on his next day at work is 0.69.

6. Given, an event A, defined as the event of getting an even number, when a fair die is rolled and event B defined as the event of getting a high score and they are given as $A = \{2,4,6\}$ and $B = \{4,5,6\}$.

Then, the events $A \cap B$ and $A \cup B$ are given as $\{4, 6\}$ and $\{2, 4, 5, 6\}$, respectively. Then,

$$(a) P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{2}{6} = \frac{1}{3}$$

$$(b) P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{2}{3}.$$

5.8 Practice Problem Set-2

1. Define a discrete random variable and suppose that the probability of having x accidents is given by the following probability mass function (pmf)

$$P(X = x) = \frac{1}{2^{x+1}}$$

Then, comment on the validity of the pmf.

2. If there is a probability of 0.261 that a milk container is underweight and these milk containers are shipped to retail outlets in boxes of 20 containers. Then,

(a) What is the distribution of the number of underweight containers in a box, if the underweight containers are independent of each other?

(b) What is the probability that exactly 7 of them would be underweight?

3. Telephone ticket sales for an event are handled by a bank of telephone salespersons who start accepting calls at a specified time. In order to get through to an operator, a caller has to be lucky enough to place a call at just the time when a salesperson has become free from a previous client. Suppose that the chance of this is 0.1. Then,

(a) What is the distribution of the number of calls that a person needs to make until a salesperson is reached?

- (b) What is the probability that 15 or more calls are needed?
4. Suppose that a plane's engines start successfully at a given attempt with a probability of 0.75. Any time that the mechanics are unsuccessful in starting the engines, they must wait five minutes before trying again. Then, what is the probability that the plane is launched within 10 minutes of the first attempt?
5. Suppose that the number of errors in a piece of software has a Poisson distribution with parameter $\lambda = 3$. Then, what is the probability that there are three or more errors in a piece of software?
6. A quality inspector at a glass manufacturing company inspects sheets of glass to check for any slight imperfections. Suppose that the number of these flaws in a glass sheet has a Poisson distribution with parameter $\lambda = 0.5$. Then, find the probability that there are
- (a) no flaws in a sheet
 - (b) at least 2 flaws

5.9 Practice Problem Set-2 Solutions

1. A random variable is a function that assigns a numerical value to the outcomes of an experiment. A discrete random variable is a random variable that takes on a finite set of values.

The probability of having x accidents is given by the following probability mass function (pmf)

$$P(X = x) = \frac{1}{2^{x+1}}$$

We know that the total sum of probabilities must sum to 1. Hence, for the above pmf to be valid,

$$\sum_{x=0}^{\infty} P(X = x) = 1$$

$$\Rightarrow \sum_{x=0}^{\infty} \frac{1}{2^{x+1}} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1$$

Hence, the given pmf is valid.

2. Given, there is a probability of 0.261 that a milk container is underweight and 0.739 for its complementary event.

In each box, there are a total of 20 containers.

- (a) Given that the underweight containers are independent of each other and there are only two possible outcomes for each trial i.e., being underweight and not being underweight, the distribution of the number of underweight containers in a box would be a binomial distribution with parameters $n = 20$ and $p = 0.261$.
- (b) If we consider that X is a random variable that takes on the number of underweight containers, then the probability that exactly 7 of them would be underweight is given by $P(X=7) = \binom{20}{7} (0.261)^7 (0.739)^{13} = 0.125$.
3. Given, the probability that the caller is lucky enough to place a call at just the time when a salesperson has become free from a previous client is 0.1.
- (a) Here, placing a call represents a Bernoulli trial with a success probability of $p = 0.1$ and the quantity of interest is the number of calls made until the first success. So, the distribution of the required number of calls would be a geometric distribution.
- (b) The probability that 15 or more calls are needed is
- $$P(X \geq 15) = 1 - P(X \leq 14) = 1 - (1 - 0.9^{14}) = 0.9^{14} = 0.229$$
4. Here, the quantity of interest is at the number of trials until the first success. Hence, the geometric distribution is the appropriate distribution for the number of trials required to start a plane's engine.
- The probability that the plane is launched within 10 minutes of the first attempt to start the engine is the probability that no more than three attempts are required, is
- $$P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.75 + 0.25 * 0.75 + 0.25^2 * 0.75 = 0.9845$$
5. Given that the number of errors in a piece of software has a Poisson distribution with parameter $\lambda = 3$. The probability that there are three or more errors in a piece of software is
- $$P(X \geq 3) = 1 - P(X = 0) - P(X = 1) - P(X = 2)$$
- $$= 1 - \frac{e^{-3}*3^0}{0!} - \frac{e^{-3}*3^1}{1!} - \frac{e^{-3}*3^2}{2!} = 0.577$$

6. Given, the number of flaws in a glass sheet has a Poisson distribution with parameter $\lambda = 0.5$.
- (a) The probability that there are no flaws is
- $$P(X = 0) = \frac{e^{-0.5}*0.5^0}{0!} = e^{-0.5} = 0.607$$
- (b) The probability that there are at least 2 errors is
- $$P(X \geq 2) = 1 - P(X = 0) - P(X = 1)$$
- $$= 1 - \frac{e^{-0.5}*0.5^0}{0!} - \frac{e^{-0.5}*0.5^1}{1!} = 1 - 0.910 = 0.090.$$

5.10 Practice Problem Set-3

1. Consider an exam consisting of 20 multiple choice questions. Each of them have four possible options and only one of them is correct. Suppose you know the answers to only 10 questions and have no idea about the other 10 questions and you choose to answer randomly. Let X be a random variable taking the total number of correct answers. Then find the probability mass function (pmf) of the random variable X and also the probability $P(X > 15)$.
2. Let X be a discrete random variable with the following probability mass function (pmf)

$$f_X(x) = \begin{cases} 0.1, & x = 0.2 \\ 0.2, & x = 0.4 \\ 0.2, & x = 0.5 \\ 0.3, & x = 0.8 \\ 0.2, & x = 1.0 \\ 0, & \text{otherwise} \end{cases}$$

- (a) Find the range of the random variable X .
- (b) Find $P(0.25 < X < 0.75)$.
- (c) Find $P(X = 0.2 - X < 0.6)$.
3. Consider an experiment of rolling two dice, one red and one blue. Let X be a random variable taking twice the value of the blue die if the red die has an even value and to be the value of the red die minus the value of the blue die if the red die has an odd value. Construct its probability mass function of the random variable X .
4. Suppose that the random variable X is the time taken by a garage to service a car. These times are distributed between 0 and 10 hours with a cumulative distribution function (cdf)

$$F(x) = A + B * \ln(3x + 2), \quad 0 \leq x \leq 10.$$

Now, find the

- (a) values of A and B such that the given cdf is valid.
- (b) probability that a repair job takes longer than two hours.
5. Suppose that two persons A and B can transmit messages between them and another person C has the ability to eavesdrop them, with a probability of 0.8, independently of the other messages.

On a certain day, A and B sent each other a total of 8 messages and let X be the number of those 8 messages that were eavesdropped by C. Then,

- (a) Find the pmf of X and $E[X]$.
 - (b) What is the most probable number of messages eavesdropped by C?
 - (c) Find the probability that exactly two messages were eavesdropped, given that at least one message was eavesdropped.
6. Consider a game in which a fair coin is tossed ten times and a player would win if at least four coin tosses show tails. Let X be the random variable denoting the number of tails observed in the game, and let W denote the event that the player wins.
- (a) Find $P(W)$
 - (b) Find $P(X \leq 5 | W)$.
7. Suppose vehicles arrive at a road signal at an average rate of 360 per hour and the cycle of the traffic lights is set at 40 seconds. In what percentage of cycles will the number of vehicles arriving will be
- (a) exactly 5
 - (b) less than 5
8. A committee has been formed to decide whether to base a recovery vehicle on a stretch of road to clear accidents as quickly as possible. The concerned road carries over 5000 vehicles during peak rush hour period. On average, the number of incidents during the peak hour is 5. The committee wont base a vehicle on the road if the percentage of having more than 5 incidents is less than 30 %. Comment whether the committee would approve of the vehicle or not.
9. Suppose that n students are selected at random without replacement from a class containing T students, of whom A are boys and $T - A$ are girls. Let X denote the number of boys that are obtained. Find the sample size n such that the variance of X is maximum.
10. A bus arrives every 10 minutes at a bus stop. It is assumed that the waiting time for a particular individual is a random variable with a continuous uniform distribution.
- (a) What is the probability that the individual waits more than 7 minutes?
 - (b) What is the probability that the individual waits between 2 and 7 minutes?

11. A new battery supposedly with a charge of 1.5 volts actually has a voltage with a uniform distribution between 1.43 and 1.60 volts. Then,
- What is the expectation and the standard deviation of the voltage?
 - What is its cumulative distribution function?
 - If a box contains 50 batteries, what is the expectation and variance of the number of batteries in the box with a voltage less than 1.5 volts?
12. Suppose that components have failure times that are independent and can be modeled with an exponential distribution with parameter value of 0.0065 per day. If a box contains ten components, what is the probability that the box has at least eight components that last longer than 150 days?
13. Suppose a certain mechanical component produced by a company has a width that is normally distributed with mean of 2600 and a standard deviation of 0.6. Find
- the proportion of the components having a width in the range 2599 to 2601.
 - If the company needs to be able to guarantee to its purchaser that no more than 1 in 1000 of the components have a width outside the range 2599 to 2601, by how much does the standard deviation need to be reduced?
14. Find the probability density function if its cumulative density function is given as
- $$F(x) = (1 - e^{-x})U(x - c), \alpha > 0$$
- and $U(x) = 1, \forall x \geq 0$ and 0 otherwise. Consider the derivative of $U(x)$ as $\delta(x)$.
15. Let X and Y be independent Poisson variables with parameters λ and μ respectively. Show that:
- $X + Y$ is Poisson distributed with parameter $\lambda + \mu$.
 - the conditional distribution of X , given $X + Y = n$, is binomial and also find its parameters.
16. Suppose we observe a random variable X and then based on it, we make an attempt to predict the value of another random variable Y . Let $g(X)$ be the predictor of Y . Then, find the best possible predictor of Y .
17. Let X be a binomial random variable with parameters (n, p) . Show that as k goes from 0 to n , the probability mass function of X first increases monotonically and then decreases monotonically, reaching its largest value when k is the largest integer less than or equal to $(n + 1)p$.

5.11 Practice Problem Set-3 Solutions

1. Given that there are 20 multiple-choice questions, and each of them have 4 possible options.

We know answers of only 10 questions and for the remaining 10 questions we randomly choose an option. The random variable X is the total number of correct answers.

Let us consider that Y is the random variable taking the number of correct answers to the 10 questions that we randomly answer. Then, the total score becomes $X = Y + 10$.

Since, there are 4 options for every question, the success probability for randomly answering the question is $\frac{1}{4}$. So, we need to perform 10 independent Bernoulli trials with parameter $\frac{1}{4}$ and Y is the number of successes. Thus, we can say that $Y \sim \text{Binomial}(10, \frac{1}{4})$.

Now, as $X = Y + 10$, the range of X is $(10, 11, \dots, 20)$. Since, we know the pmf of Y , we can derive the pmf of X as, where k is in the range of X ,

$$P_X(k) = P(X = k) = P(Y + 10 = k) = P(Y = k - 10)$$

$$\Rightarrow P_X(k) = \binom{10}{k-10} \left(\frac{1}{4}\right)^{k-10} \left(\frac{3}{4}\right)^{10-(k-10)} = \binom{10}{k-10} \left(\frac{1}{4}\right)^{k-10} \left(\frac{3}{4}\right)^{20-k}$$

So the pmf of the random variable X would be given as

$$P_X(k) = \begin{cases} \binom{10}{k-10} \left(\frac{1}{4}\right)^{k-10} \left(\frac{3}{4}\right)^{20-k}, & k = 10, 11, \dots, 20 \\ 0, & \text{otherwise} \end{cases}$$

The probability of getting a score of more than 15 is given as

$$P(X > 15) = P_X(16) + P_X(17) + P_X(18) + P_X(19) + P_X(20)$$

$$= \binom{10}{6} \left(\frac{1}{4}\right)^6 \left(\frac{3}{4}\right)^4 + \binom{10}{7} \left(\frac{1}{4}\right)^7 \left(\frac{3}{4}\right)^3 + \binom{10}{8} \left(\frac{1}{4}\right)^8 \left(\frac{3}{4}\right)^2 + \binom{10}{9} \left(\frac{1}{4}\right)^9 \left(\frac{3}{4}\right)^1 + \binom{10}{10} \left(\frac{1}{4}\right)^{10} \left(\frac{3}{4}\right)^0$$

$$= 210 * 0.772 * 10^{-4} + 120 * 0.257 * 10^{-4} + 45 * 0.0853 * 10^{-4} + 10 * 0.0286 * 10^{-4} + 0.0095 * 10^{-4}$$

$$= 0.016212 + 0.003089 + 0.000386 + 0.0000286 + 0.00000095 = 0.0197$$

Hence, the required probability is 0.0197.

2. Given a discrete random variable X .

(a) The range of the random variable X consists of all possible values of X . So, from the given

pmf, the range of X is obtained as

$$R_X = \{0.2, 0.4, 0.5, 0.8, 1\}.$$

(b) $P(0.25 < X < 0.75) = P(X = 0.4) + P(X = 0.5) = 0.2 + 0.2 = 0.4$

(c) The conditional probability, $P(X = 0.2|X < 0.6)$ is given as

$$P(X = 0.2|X < 0.6) = \frac{P(\{X=0.2 \cap X < 0.6\})}{P(X < 0.6)} = \frac{P(X=0.2)}{P(X < 0.6)}$$

$$= \frac{P(X=0.2)}{P(X=0.2)+P(X=0.4)+P(X=0.5)} = \frac{0.1}{0.1+0.2+0.2} = 0.2$$

3. Given an experiment of rolling two dice, one red and one blue and X is a random variable taking twice the value of the blue die if the red die has an even value and to be the value of the red die minus the value of the blue die if the red die has an odd value.

Without loss of generality, we can assume that the first die is the red one and the second one is the blue die and the output of this experiment is denoted as (r,b), where 'r' takes the value of the red die and 'b' takes the value of blue die.

The sample space of rolling two dice is $\{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$.

The value of the random variable X is given by

$$X = \begin{cases} 2b, & r = \text{even} \\ r - b, & r = \text{odd} \end{cases}$$

So, for an even number on the red die, X would take on the following values $\{2, 4, 6, 8, 10, 12\}$ and for odd values of red die, X would take on the following values $\{-5, -4, -3, -2, -1, 0, 1, 3\}$. So the range of X is $\{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 6, 8, 10, 12\}$.

The pmf of the random variable X is given as follows:

$$P(X = -5) = P((r,b) = (1,6)) = \frac{1}{36}$$

$$P(X = -4) = P((r,b) = (1,5)) = \frac{1}{36}$$

$$P(X = -3) = P((r,b) = (1,4)) + P((r,b) = (3,6)) = \frac{2}{36}$$

$$P(X = -2) = P((r,b) = (1,3)) + P((r,b) = (3,5)) = \frac{2}{36}$$

$$P(X = -1) = P((r,b) = (1,2)) + P((r,b) = (3,4)) + P((r,b) = (5,6)) = \frac{3}{36}$$

$$P(X = 0) = P((r,b) = (1,1)) + P((r,b) = (3,3)) + P((r,b) = (5,5)) = \frac{3}{36}$$

$$\begin{aligned}
P(X = 1) &= P((r,b) = (3,2)) + P((r,b) = (5,4)) = \frac{2}{36} \\
P(X = 2) &= P((r,b) = (2,1)) + P((r,b) = (3,1)) + P((r,b) = (4,1)) + P((r,b) = (5,3)) + P((r,b) \\
&= (6,1)) = \frac{5}{36} \\
P(X = 3) &= P((r,b) = (5,2)) = \frac{1}{36} \\
P(X = 4) &= P((r,b) = (2,2)) + P((r,b) = (4,2)) + P((r,b) = (5,1)) + P((r,b) = (6,2)) = \frac{4}{36} \\
P(X = 6) &= P((r,b) = (2,3)) + P((r,b) = (4,3)) + P((r,b) = (6,3)) = \frac{3}{36} \\
P(X = 8) &= P((r,b) = (2,4)) + P((r,b) = (4,4)) + P((r,b) = (6,4)) = \frac{3}{36} \\
P(X = 10) &= P((r,b) = (2,5)) + P((r,b) = (4,5)) + P((r,b) = (6,5)) = \frac{3}{36} \\
P(X = 12) &= P((r,b) = (2,6)) + P((r,b) = (4,6)) + P((r,b) = (6,6)) = \frac{3}{36}
\end{aligned}$$

So the pmf of the random variable X is given as

$$f_X(x) = \begin{cases} \frac{1}{36}, & X = -5 \\ \frac{1}{36}, & X = -4 \\ \frac{2}{36}, & X = -3 \\ \frac{2}{36}, & X = -2 \\ \frac{3}{36}, & X = -1 \\ \frac{3}{36}, & X = 0 \\ \frac{2}{36}, & X = 1 \\ \frac{5}{36}, & X = 2 \\ \frac{1}{36}, & X = 3 \\ \frac{4}{36}, & X = 4 \\ \frac{3}{36}, & X = 6 \\ \frac{3}{36}, & X = 8 \\ \frac{3}{36}, & X = 10 \\ \frac{3}{36}, & X = 12 \end{cases}$$

4. Given a random variable X, which is the time taken by a garage to service a car distributed between 0 and 10 hours with a cumulative distribution function (cdf)

$$F(x) = A + B * \ln(3x + 2), 0 \leq x \leq 10$$

- (a) From the properties of the cdf, we know that $F(0) = 0$ (lower limit)
 $\Rightarrow A + B \ln(3*0 + 2) = 0$

$$= A + B \ln(2) = 0 \quad (1)$$

$$F(10) = 1 \text{ (upper limit)}$$

$$\Rightarrow A + B \ln(3*10 + 2) = 1$$

$$= A + B \ln(32) = 1 \quad (2)$$

Subtracting (1) from (2) gives us,

$$B \ln(32) - B \ln(2) = 1$$

$$= B \ln(16) = 1$$

$$\Rightarrow B = \frac{1}{\ln(16)}$$

This gives us $A = -0.25$. Hence, the values of A and B are -0.25 and 0.361 respectively.

So, the cdf becomes

$$F(x) = -0.25 + 0.361 \ln(3x + 2), 0 \leq x \leq 10.$$

(b) The probability that a repair job takes longer than two hours is

$$\begin{aligned} P(X > 2) &= 1 - P(X \leq 2) = 1 - F(2) \\ &= 1 - (-0.25 + 0.361 * \ln(3 * 2 + 2)) = 0.4993 \end{aligned}$$

5. Given that two persons A and B transmit messages between them and another person C eavesdrops on them, with a probability of 0.8, independently of the other messages. On a certain day, A and B sent each other a total of 8 messages and X is the number of those 8 messages that were eavesdropped by C.

(a) Since C can either eavesdrop the messages or can't, with a success probability of 0.8 and also C eavesdrops each message independently. Hence, the distribution of the random variable X is Binomial distribution, Bin (8, 0.8). Thus, the pmf of X is

$$P_X(k) = \binom{8}{k} \left(\frac{8}{10}\right)^k \left(\frac{2}{10}\right)^{8-k}$$

Since, it is a Binomial distribution, its mean would be given as

$$E[X] = np = 8 * 0.8 = 6.4$$

(b) The most probable number of messages eavesdropped by C is the mode of the Binomial distribution, which is given by

$$\lfloor (n+1)p \rfloor = \lfloor (8+1)0.8 \rfloor = \lfloor 7.2 \rfloor = 7$$

- (c) The probability that exactly two messages were eavesdropped, given that at least one message was eavesdropped is given as

$$\begin{aligned}
P(X = 2|X \geq 1) &= \frac{P(X = 2 \cap X \geq 1)}{P(X \geq 1)} = \frac{P(X = 2)}{P(X \geq 1)} \\
&= \frac{P(X = 2)}{1 - P(X = 0)} \\
&= \frac{\binom{8}{2} \left(\frac{8}{10}\right)^2 \left(\frac{2}{10}\right)^{8-2}}{1 - \binom{8}{0} \left(\frac{8}{10}\right)^0 \left(\frac{2}{10}\right)^{8-0}} \\
&= 4.096 * 10^{-5}
\end{aligned}$$

6. Given a game in which a fair coin is tossed ten times and a player wins if at least four coin tosses show tails. The random variable X denotes the number of tails observed in the game, and W denotes the event that the player wins.

- (a) The probability that the player wins is $P(W)$ given as

$$P(W) = P(X \geq 4) = 1 - P(X < 4)$$

Since each trial is independent of other trials and also each trial has only two outcomes, the distribution of the random variable X is Binomial distribution, $\text{Bin}(10, 0.5)$. Hence, the required probability is given as

$$\begin{aligned}
P(W) &= 1 - \left[\binom{10}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{10} + \binom{10}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^9 + \binom{10}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^8 + \binom{10}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^7 \right] \\
&= 0.828
\end{aligned}$$

$$(b) P(X \leq 5|W) = \frac{P((X \leq 5) \cap (X \geq 4))}{P(X \geq 4)} = \frac{P(X=5)+P(X=4)}{P(X \geq 4)}$$

$$= \frac{\binom{10}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^6 + \binom{10}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^5}{P(X \geq 4)} = 231424 = 0.5448$$

7. Given that vehicles arrive at a road signal at an average rate of 360 per hour and the cycle of the traffic lights is set at 40 seconds. So the average rate for 40 seconds would be 4, since the rate of 360 per hour converges to 0.1 per second.

Since here the problem of interest is to find the probability of vehicles arriving in an interval of time, the random variable X , denoting the number of vehicles arriving, takes on the Poisson distribution with parameter, $\lambda = 4$ per second.

(a) The probability that there would be exactly 5 vehicles arriving is

$$P(X = 5) = \frac{\lambda^5}{5!} e^{-\lambda} = \frac{4^5}{5!} e^{-4} = 8.533 * e^{-4} = 0.15629.$$

Hence, the required percentage is 15.62.

(b) The probability that there would be less than 5 vehicles arriving is

$$\begin{aligned} P(X < 5) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \\ &= e^{-4} \left[1 + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \frac{\lambda^4}{4!} \right] \\ &= e^{-4} \left[1 + 4 + 8 + \frac{32}{3} + \frac{32}{3} \right] \\ &= 0.6288. \end{aligned}$$

Hence, the required percentage is 62.88.

8. Given, a committee has been formed to decide whether to base a recovery vehicle on a stretch of road to clear accidents as quickly as possible. The concerned road carries over 5000 vehicles during peak rush hour period. On average, the number of incidents during the peak hour is 5. The committee won't base a vehicle on the road if the percentage of having more than 5 incidents is less than 30 %.

Since here the problem of interest is to find the probability of accidents in an interval of time, the random variable X, denoting the number of accidents, takes on the Poisson distribution with parameter, $\lambda = 5$ per hour.

The probability of having more than 5 incidents is given as

$$\begin{aligned} P(X > 5) &= 1 - P(X \leq 5) \\ &= 1 - [e^{-5} (1 + \frac{5^1}{1!} + \frac{5^2}{2!} + \frac{5^3}{3!} + \frac{5^4}{4!} + \frac{5^5}{5!})] \\ &= 1 - [0.00674 + 0.03369 + 0.08422 + 0.14037 + 0.17547 + 0.17547] \\ &= 1 - 0.61596 = 0.384. \end{aligned}$$

So, the percentage of having more than 5 accidents is 38.4, which is more than 30 %. Hence, the committee should place the recovery vehicle.

9. Given that n students are selected at random without replacement from a class of T students, of whom A are boys and T - A are girls. The random variable X denotes the number of boys obtained. Now, we have to find a sample size n such that the variance is maximum.

Here, the core is that n students are drawn without replacement out of T students, of which A are boys and $T - A$ are girls. The number of boys is our success rate. Hence, the random variable X takes on the hyper-geometric distribution with parameters $A, T - A, n$.

The variance of X is given as

$$\begin{aligned} Var(X) &= n * \frac{A}{T} * (1 - \frac{A}{T}) * \frac{T - n}{T - 1} \\ &= \frac{nA(T - A)(T - n)}{T^2(T - 1)} \\ &= \frac{A(T - A)(nT - n^2)}{T^2(T - 1)} \end{aligned}$$

To get the value of n such that the variance is maximum is obtained by differentiating $\text{Var}(X)$ wrt n and equating it to 0,

$$\begin{aligned} \frac{A(T - A)(T - 2n)}{T^2(T - 1)} &= 0 \\ \Rightarrow n &= \frac{T}{2} \end{aligned}$$

Since, n can be integer only, the required n value will be $\frac{T}{2}$ if T is an even number and if T is odd, then the required n value will be $\frac{T-1}{2}$ and $\frac{T+1}{2}$.

10. Given that a bus arrives every 10 minutes at a bus stop and the waiting time for a particular individual is a random variable, X , with a continuous uniform distribution. So, we can say that $X \sim \text{Uni}(0,10)$.

- (a) The probability that the individual waits more than 7 minutes is given as

$$\begin{aligned} P(X > 7) &= \frac{1}{10 - 0} \int_{x=7}^{10} dx \\ &= \frac{1}{10} * \frac{10 - 7}{1} = 0.3 \end{aligned}$$

- (b) The probability that the individual waits between 2 and 7 minutes is given as

$$\begin{aligned} P(2 \leq X \leq 7) &= \frac{1}{10 - 0} \int_{x=2}^{7} dx \\ &= \frac{1}{10} * \frac{7 - 2}{1} = 0.5 \end{aligned}$$

11. Given that a new battery supposedly with a charge of 1.5 volts actually has a voltage with a uniform distribution between 1.43 and 1.60 volts. So the random variable, X , denoting the

voltage of the battery is $X \sim \text{Uni}(1.43, 1.60)$.

(a) The expectation of the voltage is given as

$$E[X] = \frac{\text{sum of limits}}{2} = 1.43 + 1.602 = 1.515$$

and the standard deviation of the voltage

$$\begin{aligned}\sigma &= \sqrt{\frac{(\text{difference between the limits})^2}{12}} = \frac{1.60 - 1.43}{\sqrt{12}} \\ &= 0.0491\end{aligned}$$

(b) Its cumulative distribution function is given as

$$\begin{aligned}F(x) &= \int_{1.43}^x \frac{1}{1.60 - 1.43} dx = \frac{x - 1.43}{1.60 - 1.43} \\ &= \frac{x - 1.43}{0.17}, \quad 1.43 \leq x \leq 1.60\end{aligned}$$

(c) Given that a box contains 50 batteries. Let X be the random variable denoting the number of batteries with a voltage less than 1.5 volts. Then, X would take on the Binomial distribution with parameters $n = 50$ and a success probability equals the probability that a battery will have a voltage less than 1.5 volts.

The probability that a battery will have a voltage less than 1.5 volts is given as $F(1.5) = \frac{1.5 - 1.43}{0.17} = \frac{0.07}{0.17} = 0.412$. Hence, the random variable $X \sim \text{Bin}(50, 0.412)$. So, the expected value is $E[X] = np = 50 * 0.412 = 20.6$ and the variance of X is $\text{Var}(X) = np(1-p) = 50 * 0.412 * 0.588 = 12.11$.

12. Given that components have failure times that are independent are modeled with an exponential distribution with parameter value of 0.0065 per day. A box contains ten components. Let T be the random variable denoting the failure time.

Let X be the random variable denoting the number of components that last longer than 150 days. Hence, the random variable X takes on the Binomial distribution with parameters $n = 10$ and a success probability, p , equals the probability that it will last longer than 150 days. Now, since failure times are modeled by exponential distribution, we can get p as follows,

$$\begin{aligned}p &= P(T > 150) = 1 - P(T \leq 150) \\ &= 1 - (1 - e^{-0.0065 * 150}) = e^{-0.0065 * 150} = 0.377\end{aligned}$$

Then, the required probability is

$$\begin{aligned} P(X \geq 8) &= P(X = 8) + P(X = 9) + P(X = 10) \\ &= \binom{10}{8}(0.377)^8(0.623)^2 + \binom{10}{9}(0.377)^9(0.623)^1 + \binom{10}{10}(0.377)^{10}(0.623)^0 \\ &= 0.00713 + 0.00096 + 0.00006 = 0.00815. \end{aligned}$$

13. Given that a certain mechanical component produced by a company has a width that is normally distributed with mean of 2600 and a standard deviation of 0.6. Let X be the random variable taking the width value, then $X \sim N(2600, 0.36)$.

- (a) The probability of the components having a width in the range 2599 to 2601 is given by

$$\begin{aligned} P(2599 \leq X \leq 2601) &= \phi\left(\frac{2601 - 2600}{0.6}\right) - \phi\left(\frac{2599 - 2600}{0.6}\right) \\ &= 0.9522 - 0.0478 = 0.9044. \end{aligned}$$

- (b) Given that the company needs to be able to guarantee to its purchaser that no more than 1 in 1000 of the components have a width outside the range 2599 to 2601 and we need to find a suitable standard deviation.

$$\begin{aligned} P(2599 \leq X \leq 2601) &= 1 - 0.001 = 0.999 \\ \Rightarrow \phi\left(\frac{2601 - 2600}{\sigma}\right) - \phi\left(\frac{2599 - 2600}{\sigma}\right) &= 0.999 \\ &= \phi\left(\frac{1}{\sigma}\right) - \phi\left(-\frac{1}{\sigma}\right) = 0.999 \\ &= \phi\left(\frac{1}{\sigma}\right) - (1 - \phi\left(\frac{1}{\sigma}\right)) = 0.999 \\ &= 2\phi\left(\frac{1}{\sigma}\right) - 1 = 0.999 \\ \Rightarrow \phi\left(\frac{1}{\sigma}\right) &= \frac{1 + 0.999}{2} = 0.9995 \\ \frac{1}{\sigma} &= \phi^{-1}(0.9995) = 3.2905 \\ \Rightarrow \sigma &= \frac{1}{3.2905} = 0.304. \end{aligned}$$

So the required standard deviation is 0.304.

14. Given the cumulative density function is

$$F(x) = (1 - e^{-x})U(x - c), \alpha > 0$$

and $U(x) = 1, \forall x \geq 0$ and 0 otherwise.

Its pdf can be obtained as

$$\frac{d(F(x))}{dx} = \frac{d((1 - e^{-x})U(x - c))}{dx}$$

$$\begin{aligned} &= U(x - c) * \frac{d(1 - e^{-x})}{dx} + (1 - e^{-x}) * \frac{d(U(x - c))}{dx} \\ &= U(x - c) * (0 - e^{-\alpha x} * (-\alpha)) + (1 - e^{-\alpha x}) * \delta(x - c), \\ &= \alpha e^{-\alpha x} * U(x - c) + (1 - e^{-\alpha x}) * \delta(x - c), \end{aligned}$$

Hence, the required pdf is $\alpha e^{-\alpha x} * U(x - c) + (1 - e^{-\alpha x}) * \delta(x - c)$.

15. Given that X and Y are independent Poisson variables with parameters λ and μ respectively.

(a) Since, X and Y are both random variables, $Z = X + Y$ would be a random variable. Its probability mass function would be given as, from total probability

$$\begin{aligned} P(X + Y = n) &= \sum_{k=0}^n P(X + Y = n | Y = k)P(Y = k) \\ &= \sum_{k=0}^n P(X = n - k)P(Y = k) \end{aligned}$$

Since, both X and Y are independent Poisson random variables, we can write the above equation as

$$= \sum_{k=0}^n \frac{e^{-\lambda} \lambda^{n-k}}{(n - k)!} * \frac{e^{-\mu} \mu^k}{k!} = e^{-(\lambda + \mu)} \sum_{k=0}^n \frac{\lambda^{n-k}}{(n - k)!} * \frac{\mu^k}{k!}$$

Multiplying and dividing by $n!$, we get

$$\begin{aligned} &= \frac{e^{-(\lambda + \mu)}}{n!} \sum_{k=0}^n \frac{*n!}{(n - k)! * k!} * \lambda^{n-k} \mu^k \\ &= \frac{e^{-(\lambda + \mu)}}{n!} \sum_{k=0}^n \binom{n}{k} \lambda^{n-k} \mu^k \end{aligned}$$

Using the binomial expansion, we get conclude that

$$\begin{aligned} &= \frac{e^{-(\lambda + \mu)}}{n!} (\lambda + \mu)^n \\ \Rightarrow P(X + Y = n) &= \frac{e^{-(\lambda + \mu)}}{n!} (\lambda + \mu)^n \end{aligned}$$

Hence, the random variable $Z = X + Y$ is a Poisson distribution with parameter $(\lambda + \mu)$.

(b) The conditional distribution of X, given $X + Y = n$, is given by

$$P(X = k | X + Y = n) = \frac{P(X = k, X + Y = n)}{P(X + Y = n)}$$

Since, X and Y are independent, the above equation can be simplified as

$$\begin{aligned} &= \frac{P(X = k)P(X + Y = n)}{P(X + Y = n)} \\ &= \frac{e^{-\lambda}\lambda^k}{k!} * \frac{e^{-\mu}\mu^{n-k}}{(n-k)!} * \frac{n!}{e^{-(\lambda+\mu)}(\lambda+\mu)^n} \\ &= \frac{e^{-\lambda}\lambda^k e^{-\mu}\mu^{n-k} n!}{k!(n-k)!e^{-(\lambda+\mu)}(\lambda+\mu)^n} \\ &= \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{\lambda+\mu}\right)^k \left(\frac{\mu}{\lambda+\mu}\right)^{n-k} \end{aligned}$$

Hence, the required conditional probability is binomial with parameters n and $\frac{\lambda}{\lambda+\mu}$.

16. Given that we observe a random variable X and then based on it, we make an attempt to predict the value of another random variable Y and that predictor of Y is $g(X)$. Now, we need to find the best possible predictor of Y, i.e., we need to find $g(X)$ that tends close to Y. One way to do this is to choose g which minimizes $E[(Y - g(X))^2]$.

Now, consider $E[(Y - g(X))^2 | X]$ and add and subtract $E[Y - X]$, so we get

$$\begin{aligned} &= E[(Y - E[Y|X] + E[Y|X] - g(X))^2 | X] \\ &= E[(Y - E[Y|X])^2 | X] + E[(E[Y|X] - g(X))^2 | X] + 2E[(Y - E[Y|X])(E[Y|X] - g(X)) | X] \end{aligned}$$

Here, it noteworthy to note that given X, $E[Y - X] - g(X)$, being a function of X, is a constant. Thus, we can write $= E[(Y - E[Y|X])(E[Y|X] - g(X)) | X]$
 $= (E[Y|X] - g(X))E[(Y - E[Y|X]) | X]$
 $= (E[Y|X] - g(X))(E[(Y|X) - E[Y|X]]) = 0$

Substituting this in the above main equation, we get

$$E[(Y - g(X))^2 | X]E[(Y - E[Y|X])^2 | X]$$

Now, since given X, all the above are just constants, on applying expectation to it once again we get the following,

$$E[(Y - g(X))^2] \geq E[(Y - E[Y|X])^2].$$

17. Given that X is a binomial random variable, with parameters (n,p). We need to show that as k goes from 0 to n, the probability mass function of X first increases monotonically and then decreases monotonically, reaching its largest value when k is the largest integer less than or equal to $(n + 1)p$.

To show the required trend, let us consider the probability of two consecutive k values.

$$\begin{aligned} \frac{p(k)}{p(k-1)} &= \frac{\binom{n}{k} p^k (1-p)^{n-k}}{\binom{n}{k-1} p^{k-1} (1-p)^{n-k+1}} \\ &= \frac{\frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}}{\frac{n!}{(n-k+1)!(k-1)!} \frac{p^k}{p} (1-p)^{n-k} (1-p)} \\ &= \frac{n!}{(n-k)!k!(k-1)!} * p^k (1-p)^{n-k} * \frac{(n-k+1)(n-k)!(k-1)!p}{n!p^k(1-p)^{n-k}(1-p)} \\ &= \frac{(n-k+1)p}{k(1-p)} \end{aligned}$$

Note that from above, we can conclude that $p(k) \geq p(k-1)$ if and only if $(n-k+1)p \geq k(1-p)$

$$\begin{aligned} &= np - kp + p \geq k - kp \\ &= np + p \geq k \\ &\Rightarrow (n+1)p \geq k. \end{aligned}$$

Thus, $p(k)$ increases monotonically and reaches its maximum when k is the largest integer less than or equal to $(n+1)p$ and thereafter decreases monotonically.

Chapter 6

Random Variables

6.1 Introduction

Till now, we have discussed the basics of set theory and probability theory. Now, we will be heading towards the most important and fundamental entity of probability theory and statistical inference, random variables. In this chapter, we would extend the so far developed theory for outcomes of numerical nature. We would also discuss the probability density (and also the mass) functions and the cumulative distribution functions which are used to describe these random variables.

6.2 Definition

Consider a random experiment with sample space S . Then, a random variable X is defined as a measurable function $X : S \rightarrow E$, from the sample space S to a measurable space E . For the most practical cases, we can define a random variable as a function $X : S \rightarrow \mathbb{R}$. In simple terms, we can define a random variable as a function that assigns some numerical value, called as the *value* of X to each sample point in the sample space S .

So, in general we can define a random variable as follows:

A random variable X is a function defined on S , which takes values on the real axis. Pictorially, it would be represented as shown in 6.1.

The sample space S of the random experiment becomes the *domain* of the random variable X and the values taken by it constitutes the *range* of the random variable X . It is important to note that two or more different sample points might take on the same value of the random variable X , but two different numbers in the range can't be assigned to the same sample point.

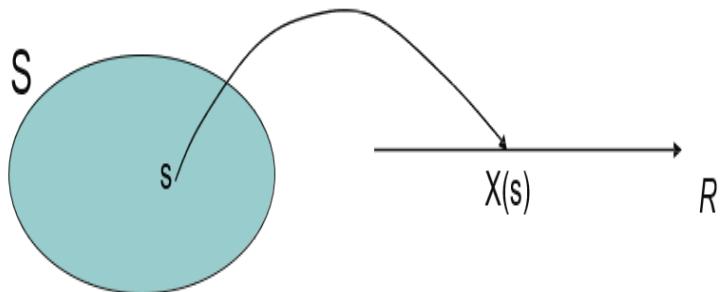


Fig. 6.1: Random Variable

In practical applications, which we will discuss later, sometimes we might interested on numerical aspects of the random experiment rather than the outcome of the random experiment itself. This laid the foundation to the concepts of *random variables*. A point to be noted here is that the concerned variable, random variable, is different from the conventional variable, which we have studied in algebra and calculus. The conventional variable have definite values and could be described by a single or group of numbers. However, random variables don't have any particular number and should be described by using probabilities.

Most of the confusion arises from the fact that both of these variables represent similar things though they aren't identical. To clear the confusion, let us consider a game in which we have a set of finite numbers and the player have to pick a number, which would then denote that player's score. So in this case, the conventional variable denotes what we actually got, which has zero variance. On the other hand, the random variable comes into play before the game starts, i.e., it would be used to describe the outcome to be expected.

For the representation part, we would denote, as per the convention, the random variable by upper case, X , while the lower case, x , to denote a quantity whose value remains constant.

In the next subsection, we would discuss some real-life examples of random variables.

6.3 Examples

Here, we would go through some of the common, yet relatable examples in the context of random variables. The real-life examples are as follows:

1. You are typing a letter in MS word and you have the option of selecting the font size of your text. The font size can be represented as a random variable.
2. Suppose you roll two dice and you are interested in the product of the two numbers obtained on each die rather than the outcome itself. So, here the product of the numbers on the two faces of the die is our random variable.
3. If you flip a coin, then random variable may takes on the values representing the number of heads (or tails) so obtained.

6.4 Types

There are mainly two types of random variables, namely,

1. Discrete Random Variable
2. Continuous Random Variable

The above classification is based on the whether the concerned random variable is countable (both finite and infinite) or uncountable. We would discuss about them in detail in the subsequent sections.

6.5 Discrete Random Variables

A random variable X is said to be *discrete random variable* if X takes on only a finite number of different values or at most, an infinite sequence of different values. More formally, we can define a discrete random variable as follows:

The random variable X , $X : S \rightarrow E$, is a discrete random variable if the range or image of the random variable is countable.

Some of the examples for a discrete random variables are as follows:

1. Consider your class of this course IC-252. Suppose the course has a strength of 320 and on a particular day, the logistic TAs mark the attendance of students present in that class. The random variable taking the number of students present would be a discrete random variable type.
2. Consider your own family. Then, the random variable taking on the number of people in the family would be a discrete random variable.
3. Consider the parking lot of our college. Then the random variable taking on the number of vehicles parked there would be a discrete random variable.

4. Consider the number of trees (or even the number of species) on the Griffon peak. The random variable taking on the number of trees (or even the number of species) there, would be a discrete random variable.

In the next sub-section, we would discuss about the probability mass function, which is used to characterize the probabilities that the discrete random variable would take on.

6.5.1 Probability Mass Function

If X is a discrete random variable, then its range, R_X , is a countable set. So, we can write its range as R_X , given by

$$R_X = \{x_1, x_2, \dots\}$$

where x_1, x_2, \dots are the possible values of the random variable X .

The random variable X , here would assume each of those values with certain probability. Mathematically, it would be useful if we represent all the probabilities of the values or elements in its *range* by a formula. The entity used to represent all the probabilities is known as the *probability mass function* or *pmf*.

Let us now define the *probability mass function* or *pmf* as follows:

Let X be a discrete random variable with range $R_X = \{x_1, x_2, \dots\}$. Then, the *probability mass function* or *pmf* of X is the set of probability values assigned to each of the values x_i taken by the discrete random variable X .

$$p_X(x_i) = P(X = x_i), \quad \text{for } i = 1, 2, 3, \dots$$

Thus, the *pmf* is a probability measure that gives us the probabilities of the possible values for a random variable. The subscript X , here, indicates that this *pmf* is for the random variable X .

Even though the *pmf* is usually defined for only the values in its range, R_X , it is customary to extend the *pmf* of X to all the real numbers. If $x \notin R_X$, we can simply write

$$p_X(x) = P(X = x) = 0$$

. Thus, in general we can write

$$p_X(x) = \begin{cases} P(X = x), & \text{if } x \in R_X \\ 0, & \text{otherwise} \end{cases}$$

Now we can now, summarize it as follows:

If $X: S \rightarrow R$ is a discrete random variable, then it has a *pmf*, $p_X(x)$ or $f(x)$, having the following properties:

1. $p_X(x) \geq 0, \forall x \in R_X$
2. $\sum_k p_X(x_k) = 1, \quad k = 1, 2, \dots$
3. $P(X = x) = p_X(x)$, where $P(X = x)$ is the probability of the outcomes $s \in S: X(s) = x$.

Let us consider two examples to get more insights into discrete random variable.

Example 1: Consider tossing a fair coin twice, and the random variable X denote the number of heads observed while tossing the coin twice. Find the range of X , R_X and the probability mass function of X .

Solution: The sample space of tossing a coin twice is S , given by

$$S = \{HH, HT, TH, TT\}$$

If we toss a coin twice, then the possible values for the number of heads is either 0, 1 or 2. Thus, the range of X , R_X , is given by

$$R_X = \{0, 1, 2\}.$$

Since, R_X is a finite set, the random variable X is a discrete random variable. It's *pmf* is given by

$$p_X(x) = P(X = x), \quad \text{for } x = 0, 1, 2$$

Now,

$$p_X(0) = P(X = 0) = P(TT) = \frac{1}{4}$$

$$p_X(1) = P(X = 1) = P(\{HT, TH\}) = \frac{2}{4}$$

$$p_X(2) = P(X = 2) = P(HH) = \frac{1}{4}$$

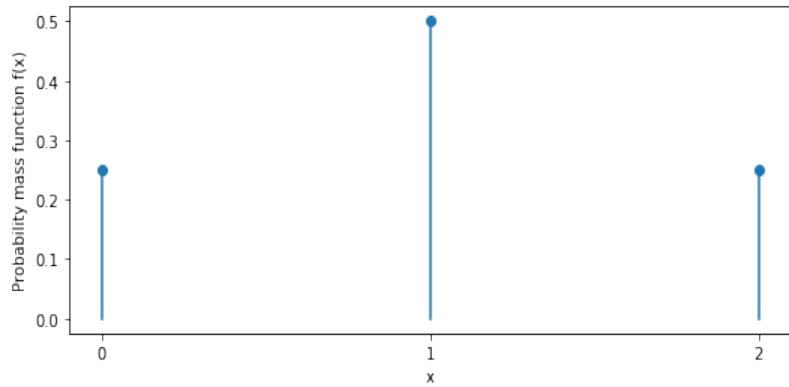


Fig. 6.2: Probability Mass Function

So, the *pmf* of X is given by

$$p_X(x) = \begin{cases} \frac{1}{4} & x = 0 \\ \frac{2}{4} & x = 1 \\ \frac{1}{4} & x = 2 \end{cases}$$

For visualization, the *pmf* is as shown in the figure 6.2:

Example 2: Consider an unfair coin for which $P(\text{Head})=p$, where $0 < p < 1$. If we toss the coin repeatedly until we observe a head for the first time and Y be the total number of coin tosses. Find the distribution of Y .

Solution: Since, Y is total number of coin tosses required to get a head for the first time, it can possibly take any positive number. So, the range of Y , R_Y would be given as

$$R_Y = \{1, 2, 3, \dots\}.$$

The *pmf* of Y is given by

$$p_Y(y) = P(Y = y)$$

So now,

$$p_Y(1) = p(Y = 1) = P(H) = p$$

$$p_Y(2) = p(Y = 2) = P(TH) = (1 - p)p$$

$$p_Y(3) = p(Y = 3) = P(TTH) = (1 - p)^2 p$$

⋮

$$p_Y(k) = p(Y = k) = P(T \dots TH) = (1 - p)^{k-1} p$$

Hence, the *pmf* of the random variable Y is given by

$$p_Y(y) = \begin{cases} (1-p)^{y-1}p & y = \{1, 2, 3, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

6.5.2 Cumulative Distribution Function

In the previous sub-section, we have discussed about the *probability mass function*, which characterizes the distribution of a discrete random variable. Now, in this sub-section, we would discuss about the *cumulative distribution function* or *cdf*, which is also used to describe the distribution of a random variable. It can be described for both the discrete and continuous random variables. Now, we will formally define the *cumulative distribution function* $F_X(x)$ as follows:

If $X: S \rightarrow R$, is a discrete random variable with *probability mass function*, $f(x)$ or $p_X(x)$, the *cumulative distribution function*, $F_X(x)$ is defined by

$$F_X(x) = P(X \leq x) \quad \text{for } -\infty < X \leq x$$

Like *pmf*, *cdf* too describes the probabilistic characteristics of the random variable X . Knowing either of these functions, we can get knowledge of the other distribution.

For a discrete random variable, the *cumulative distribution function* or *cdf* is defined as follows:

$$F(x) = \sum_{y:y \leq x} P(X = y)$$

In simple words, the *cumulative distribution function* or *cdf*, $F(x)$, is the sum of probabilities $P(y)$ such that y isn't greater than x .

The *cumulative distribution function* or *cdf* of a random variable X is an increasing step function with steps at the values taken by the random variable. The heights of the steps are the probabilities of taking these values. The *probability mass function* is related to the *cumulative distribution function* as

$$P(X = x) = F(x) - F(x^-)$$

where $F(x^-)$ is the limiting value from below of the cumulative distribution function.

Now, let us discuss the properties (and their proofs) of the *cumulative distribution function* or

cdf. Note that $F(x^+)$ and $F(x^-)$ would denote the limits,

$$F(x^+) = \lim_{\epsilon \rightarrow 0} F(x + \epsilon), \quad F(x^-) = \lim_{\epsilon \rightarrow 0} F(x - \epsilon), \quad \forall \epsilon > 0$$

1. $0 \leq F(x) \leq 1$, for $-\infty \leq x \leq \infty$

Proof. Now to prove that the *cdf* is in the above mentioned interval, we will find its extreme values.

$$F(+\infty) = P(X \leq \infty) = P(S) = 1, \quad \text{and}$$

Since, all the values of x are less than or equal to ∞ , the above set becomes the sample space and hence the above one holds true.

$$F(-\infty) = P(X \leq -\infty) = P(\phi) = 0.$$

Since, all the values of x are greater than or equal to $-\infty$, the above set becomes the null set and hence the above one holds true.

$$\Rightarrow 0 \leq F(x) \leq 1.$$

Hence, proved. □

2. The *cdf* of a random variable X is a non-decreasing function of x :

$$\text{If } (x_1 < x_2), \text{ then } F(x_1) < F(x_2).$$

Proof. We know that the *cdf* is a probability measure. The *cdf* of random variable X , $F(x)$ is a probability measure on the set $(-\infty, x]$. Similarly, for a random variable Y , $F(y)$ is a probability measure on the set $(-\infty, y]$.

Now, if $x_1 < x_2$, then we know that $(-\infty, x_1] \subset (-\infty, x_2]$. From the basic properties of probability we know that if $A \subset B$, then $P(A) < P(B)$. So, applying this property of the *cdf*, for $x_1 < x_2$, we get

$$P((-\infty, x_1]) < P((-\infty, x_2]),$$

$$\Rightarrow F(x_1) < F(x_2).$$

Hence, proved. □

3. If $F(x_0) = 0$, then $F(x) = 0$, for every $x \leq x_0$.

Proof. From property (1), we know that $0 \leq F(x) \leq 1$, for any random variable X and from property (2), we know that the *cdf* is a non-decreasing function.

Now, given that $F(x_0) = 0$. So, for $x \leq x_0$, from property (2), we can write that

$$F(x) \leq F(x_0) = F(x) \leq 0.$$

However, from property (1), we know that $0 \leq F(x)$. Thus, the inequalities, $F(x) \geq 0$ and $F(x) \leq 0$ leads to $F(x) = 0$.

Hence, proved that if $F(x_0) = 0$, then $F(x) = 0$, for every $x \leq x_0$. \square

4. $P(X > x) = 1 - F(x)$.

Proof. Consider the two mutually exclusive events $\{X \leq x\}$ and $\{X > x\}$. The union of these two sets would form the sample space, S

$$\{X \leq x\} \cup \{X > x\} = S.$$

$$\begin{aligned} &\Rightarrow P(\{X \leq x\}) + P(\{X > x\}) = P(S) \\ &= P(\{X > x\}) = 1 - P(\{X \leq x\}) \\ &\Rightarrow P(\{X > x\}) = 1 - F(x) \end{aligned}$$

Hence, proved that $P(X > x) = 1 - F(x)$. \square

5. The *cdf* of any random variable X , $F(x)$ is right continuous.

Proof. Mathematically, this means that $\lim_{\epsilon \rightarrow 0} F(x + \epsilon) = F(x)$.

Let us consider a decreasing sequence x_n such that $x_n \rightarrow x$ as $n \rightarrow \infty$. Then, since x_n is a decreasing sequence, $A_x \subseteq A_{x_n}$ for all n and A_x is the largest set for which it is true. Then we can write

$$\cap_{n=1}^{\infty} A_{x_n} = A_x$$

$$\lim_{n \rightarrow \infty} F(x_n) = \lim_{n \rightarrow \infty} P(A_{x_n}) = P(\cap_{n=1}^{\infty} A_{x_n}) = P(A_x) = F(x)$$

Since the above holds for any sequence $\{x_n\}$ such that $\{x_n\} \rightarrow x$, we can conclude that $\lim_{\epsilon \rightarrow 0} F(x + \epsilon) = F(x)$ for all x and so $F(x)$ is right continuous. \square

6. $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$, $x_2 > x_1$.

Proof. Consider the two disjoint events $\{x \leq x_1\}$ and $\{x_1 < X \leq x_2\}$. We can write

$$\begin{aligned} \{X \leq x_2\} &= \{x \leq x_1\} \cup \{x_1 < X \leq x_2\} \\ \Rightarrow P(X \leq x_2) &= P(x \leq x_1) + P(x_1 < X \leq x_2) \\ &= P(x_1 < X \leq x_2) = P(X \leq x_2) - P(x \leq x_1) \\ \Rightarrow P(x_1 < X \leq x_2) &= F(x_2) - F(x_1) \end{aligned}$$

Hence, proved that $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$, $x_2 > x_1$. \square

7. $P(X = x) = F(x) - F(x^-)$.

Proof. Assume that $x_1 = x - \epsilon$ and $x_2 = x$ in the property (6). Then, we get

$$P(x - \epsilon < X \leq x) = F(x) - F(x - \epsilon)$$

Then for $\epsilon \rightarrow 0$, we can write

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} P(x - \epsilon < X \leq x) &= \lim_{\epsilon \rightarrow 0} F(x) - F(x - \epsilon) \\ \Rightarrow P(X = x) &= F(x) - F(x^-). \end{aligned}$$

Hence, proved. \square

8. $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1^-)$, $x_2 > x_1$.

Proof. Consider the two disjoint sets $\{x_1 < X \leq x_2\}$ and $\{x = x_1\}$. Then, we can write

$$\{x_1 \leq X \leq x_2\} = \{x_1 < X \leq x_2\} \cup \{x = x_1\}.$$

Now,

$$P(\{x_1 \leq X \leq x_2\}) = P(\{x_1 < X \leq x_2\}) + P(\{x = x_1\})$$

From properties (6) and (7), we get

$$\begin{aligned} P(\{x_1 \leq X \leq x_2\}) &= F(x_2) - F(x_1) + F(x_1) - F(x_1^-) \\ \Rightarrow P(\{x_1 \leq X \leq x_2\}) &= F(x_2) - F(x_1^-) \end{aligned}$$

Hence, proved. \square

6.5.3 Functions of Random Variables

In this subsection, we would discuss about the functions of random variables.

Let us consider a random variable X , then its function, $Y = g(X)$, is a random variable itself. Hence, we can comment about its *pmf*, *cdf* and other related properties.

Note that the range of the random variable Y is given as

$$R_Y = \{g(x) | x \in R_X\}.$$

Now, to know the *pmf* of the random variable Y , we first need to know the *pmf* of the random variable X . Let the *pmf* of the random variable X is denoted as $P_X(x)$. Then for $Y = g(X)$, the *pmf* of the random variable Y is given as

$$P_Y(y) = P(Y = y) = P(g(X) = y) = \sum_{x:g(x)=y} P_X(x)$$

Let us go through some examples to get better understanding of it.

Example 1: Let X be a discrete random variable with $P_X(k) = \frac{1}{5}$, for $k = -1, 0, 1, 2, 3$. Let $Y = 2|X|$. Find the range and *pmf* of Y .

Solution: Given, $Y = 2|X|$, where the *pmf* of X is given as

$$P_X(k) = \frac{1}{5}, \text{ for } k = -1, 0, 1, 2, 3$$

Then, the range of Y would be

$$R_Y = 2|x| \quad \forall x \in R_X$$

$$\Rightarrow R_Y = \{0, 2, 4, 6\}.$$

Now, we would find the *pmf* of Y .

$$P_Y(y) = P(Y = y) \quad \forall y \in R_Y$$

$$P_Y(0) = P(Y = 0) = P(2|X| = 0) = P(X = 0) = \frac{1}{5}$$

$$P_Y(2) = P(Y = 2) = P(2|X| = 2) = P(X = -1) + P(X = 1) = \frac{2}{5}$$

$$P_Y(4) = P(Y = 4) = P(2|X| = 4) = P(X = 2) = \frac{1}{5}$$

$$P_Y(6) = P(Y = 6) = P(2|X| = 6) = P(X = 3) = \frac{1}{5}$$

We can summarize the *pmf* of Y as follows:

$$P_Y(y) = \begin{cases} \frac{1}{5} & y = 0 \\ \frac{2}{5} & y = 2 \\ \frac{1}{5} & y = 4 \\ \frac{1}{5} & y = 6 \end{cases}$$

The random variable Y satisfies all the required properties.

Example 2: Let X be a random variable with given *pmf*, $P_X(x)$ and *cdf*, $F_X(x)$. Let Y be another random variable such that $Y = a + bx$, where a and b are some constants and $b > 0$. Find the *cdf* of Y in terms of $F_X(x)$.

Solution Given, a random variable X with *pmf*, $P_X(x)$ and *cdf*, $F_X(x)$. And also Y is another random variable such that $Y = a + bx$, where a and b are some constants and $b > 0$.

The *pmf* of Y is given as

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(a + bx \leq y) = P(bx \leq y - a) = P(x \leq \frac{y - a}{b}) \\ &\Rightarrow F_Y(y) = F_X\left(\frac{y - a}{b}\right) \end{aligned}$$

where $b > 0$.

6.5.4 Special Distributions

In this sub-section, we would discuss about some of the distributions which we encounter frequently in our day-to-day life. So these are given special names. Each of these distributions describe a random experiment, which models many of our day-to-day activities.

6.5.4.1 Bernoulli Distribution

It is the simplest discrete distribution. A Bernoulli random variable takes only two possible values, usually 0 and 1. This describes the random experiments that have only two possible outcomes, usually referred to as “success” and “failure.”

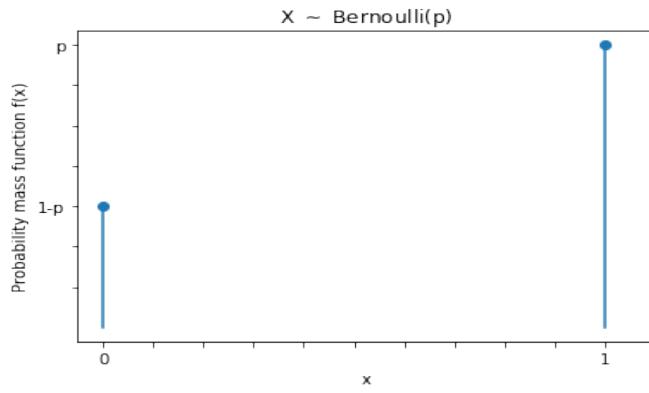


Fig. 6.3: Probability mass function of a Bernoulli random variable

Formally, it is defined as follows:

A random variable X is said to be a *Bernoulli random variable* with parameter p , $X \sim \text{Bernoulli}(p)$, if its pmf is given by

$$P_X(x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $0 \leq p \leq 1$, is known as the success probability of the experiment.

The *probability mass function* of $X \sim \text{Bernoulli}(p)$ is as shown in figure 6.3

Some of the cases, where this could be used are

1. In an exam, either we pass ($X = 1$) or fail ($X = 0$) with certain probabilities.
2. Tossing a coin once, we get either a head or tail. We can assign 1 to either of them.
3. The pen we are using to write the exam would either work ($X = 1$) or not ($X = 0$).

The *expectation* of this random variable is given by

$$E[X] = 0 * P(X = 0) + 1 * P(X = 1) = 0 * (1 - p) + 1 * p = p.$$

Also, $E[X^2] = 0^2 * P(X = 0) + 1^2 * P(X = 1) = p$. So, the variance becomes

$$\text{Var}(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1 - p)$$

6.5.4.2 Binomial Distribution

In the previous subsection, we have discussed the Bernoulli distribution, which consists of only a single trial known as *Bernoulli trial*. However, in practical applications, it is useful if we have distribution consisting of a sequence of Bernoulli trials. In such cases, the random variable of interest would be the number of *successes*, within a finite number of trials, with success being well defined. Such a distribution is known as the Binomial distribution. It is the most important discrete random variable.

Consider that there are n independent Bernoulli trials X_1, X_2, \dots, X_n , each of having a probability p of registering a 1, then the random variable

$$X = X_1 + X_2 + \dots + X_n$$

is known as the *Binomial distribution*, denoted as $X \sim Bin(n, p)$. The random variable X can take on values ranging from 0 to n and counts the number of 1's (success) in the n trials.

The *probability mass function* of a Binomial random variable X is given as

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x=0, 1, \dots, n.$$

The corresponding *cdf* is given as

$$F_X(x) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} \quad n \leq x < n+1$$

There are some properties which have to be satisfied by the random variable in order to say that it follows a Binomial distribution and they are as follows:

1. The experiment must consist of a fixed number of trials, say n .
2. Each trial must be independent of the other trial.
3. Each trial must have only two outcomes, either it should happen (success) or doesn't (failure).
4. Each of them must have a constant probability p of success.

Some of the practical applications are as follows:

1. In an exam, for say, IC-252 mid-semester, the number of students, if each of them has the same probability of clearing the exam, can be modelled as a Binomial random variable.
2. Tossing a coin n times and the number of heads we get can be modelled using this.

The expectation of a Binomial random variable, $X \sim Bin(n, p)$ is np and the variance is npq .

Example: A multiple-choice quiz consists of ten questions, each with five possible answers of which only one is correct. A student passes the quiz if seven or more correct answers are obtained. What is the probability that a student who guesses blindly at all of the questions will pass the quiz?

Solution: Given, there are 10 mcq questions, each having 5 possible answers. The student guesses blindly at all the questions. So, the probability of selecting the correct answer would be $p = \frac{1}{5} = 0.20$.

A students would pass the quiz only if he/she answers at least 7 correctly. Since, these answers are independent of each other and also the probability of selecting a correct choice remains constant, we can model the distribution using the Binomial distribution. So, if X is the random variable of the number of correct answers, then the required probability is

$$P(X \geq 7) = \binom{10}{7}(0.2)^7(0.8)^3 + \binom{10}{8}(0.2)^8(0.8)^2 + \binom{10}{9}(0.2)^9(0.8)^1 + \binom{10}{10}(0.2)^10(0.8)^0 \\ = 0.0009.$$

6.5.4.3 Geometric Distribution

The *binomial distribution*, discussed above, is the distribution of number of successes in a fixed number of Bernoulli trials, say n . However, sometimes, we might be interested in the count of the number of trials required to register our first success. Such a distribution is the *geometric distribution*. It's *probability mass function*, $P(X = x)$, is the probability that the first success would occur at the x th trial and all the previous $x - 1$ trials are failures. So, it has a *pmf* of

$$P(X = x) = (1 - p)^{x-1}p, \text{ for } x = 1, 2, \dots$$

Now, let us formally define it as follows:

The number of trials up to and including the first success in a sequence of independent Bernoulli trials with a constant success probability p has a geometric distribution with parameter p . The *probability mass function* is $P(X = x) = (1 - p)^{x-1}p$ for $x = 1, 2, 3, 4, \dots$ and the *cumulative distribution function* is

$$P(X \leq x) = 1 - (1 - p)^x$$

The geometric distribution with parameter p has an expected value and a variance of

$$E(X) = \frac{1}{p}$$

and

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

Some of the day-to-day scenarios, where this could be used are as follows:

1. The number of trials required to open the door our room for the first in some finite trials is modelled by the geometric distribution.
2. The number of trials required to start our car for the first in some finite trials is modelled by the geometric distribution.

Example: Suppose a plane's engines start successfully at a given attempt with a probability of $p = 0.75$. Find the probability that the plane's engine would start at the third attempt.

Solution: Since, the random variable X , the number of attempts required to start the plane's engine for the first time, deals with the attempt required for first success, it follows a geometric distribution with $p = 0.75$. Then, the required probability is

$$P(X = 3) = (1 - p)^{3-1}p = 0.25^2 * 0.75 = 0.047$$

6.5.4.4 Hypergeometric distribution

In the previous subsection, we have studied the Binomial distribution. One of the criteria to be met for Binomial distribution is that the success probability p must remain constant. However, in practice, we often deal with situations where the success probability has to be changed (one of the possibility is that replacement isn't allowed). In this case, the hypergeometric distribution comes into picture.

Now, we would define it formally with greater details. Consider a collection of N items of which r are of a special kind. If one of the items is chosen at random, then the probability that it is of the special kind is $\frac{r}{N}$. If n of these items are chosen *with replacement*, then the distribution of the random variable X , taking on the success count, follows Binomial distribution, $X \sim Bin(n, \frac{r}{N})$. However, if the replacement isn't allowed then X won't follow the Binomial distribution as the success probability changes. The appropriate distribution here is the hypergeometric distribution. The *probability mass function* of X is given as

$$P(X = x) = \frac{\binom{r}{x} \binom{N-r}{n-x}}{\binom{N}{n}}$$

for $\max\{0, n + r - N\} \leq x \leq \min\{n, r\}$.

The expected value and variance of the random variable X are as follows:

$$E[X] = \frac{nr}{N} \quad \text{and,}$$

$$Var(X) = \left(\frac{N-n}{N-1}\right) * n * \frac{r}{N} * \left(1 - \frac{r}{N}\right)$$

One of the practical application is that if we have N number of computer chips, out of which r are defective and we choose n chips at random without replacement. Then, to find the number of defective chips can be modelled by using hypergeometric distribution.

Example: A small lake contains 50 fish. One day a fisherman catches 10 of these fish and tags them so that they can be recognized if they are caught again. The tagged fish are released back into the lake. The next day the fisherman goes out and catches 8 fish, which are kept in the fishing boat until they are all released at the end of the day. Find the probability that 3 tagged fish are caught on the second day.

Solution: The second day's fishing can be thought of as a sample of size eight taken without replacement from the fish stock, since the fish that are caught are kept in the fishing boat until all 8 fish have been caught. Thereby eliminating the possibility of the same fish being caught twice on the second day. Consequently, given that all 50 fish are equally likely to be caught, the number of tagged fish caught on the second day has a hypergeometric distribution with $N = 50$, $r = 10$, and $n = 8$. Then, the required probability is given as

$$P(X = 3) = \frac{\binom{10}{3} \binom{40}{5}}{\binom{50}{8}} = 0.147.$$

6.5.4.5 Poisson Distribution

Till now, no distribution had considered the time or space constraint. However, we encounter events, on a daily basis, which occur within a specified time period or space. For say, consider, we might be interested in finding the number of telephone calls received by someone within a specified time period or number of balls thrown by a player within a specified period. So an appropriate distribution for these kinds of cases is the Poisson distribution.

A Poisson experiment is based on the Poisson process and has the following properties:

1. The number of outcomes occurring in one time interval or specified region of space is independent of the number that occur in any other disjoint time interval or region. This means that it has no memory.

2. The probability that a single outcome will occur during a very short time interval or in a small region is proportional to the length of the time interval or the size of the region and does not depend on the number of outcomes occurring outside this time interval or region.
3. The probability that more than one outcome will occur in such a short time interval or fall in such a small region is negligible.

Now, we can define the *Poisson distribution* as follows:

Let the random variable X be the number of events in a time interval of length t from a Poisson process, which has an average μ events happening per unit time. Then, the distribution of X is said to be *Poisson distribution* with parameter $\lambda = \mu t$. The *probability mass function* of the Poisson distribution, $X \sim Pois(\lambda)$, is given as

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots$$

where λ is the *mean rate* of the distribution, which is also known as the parameter of the Poisson distribution.

The corresponding *cumulative distribution function* is given as

$$F_X(x) = e^{-\lambda} \sum_{k=0}^n \frac{\lambda^k}{k!} \quad n \leq x < n + 1$$

The expected value and the variance of the random variable X , which follows the Poisson distribution with parameter λ are as follows:

$$E[X] = \lambda$$

$$Var(X) = \lambda$$

Some of the cases, where we can use the Poisson distribution to model events that occur in our day-to-day life is as follows:

1. The random variable taking the value of number of errors on page can be modelled using Poisson distribution.
2. The number of particles emitting can be modelled using Poisson distribution.

Example: On average there are about 25 imperfections in 100 meters of optical cable. Find the probability that there are no imperfections in 1 meter of cable.

Solution: The Poisson parameter would be $\lambda = \frac{25}{100}$. Then, the required probability is

$$P(X = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = \frac{e^{-0.25}0.25^0}{0!} = 0.7788.$$

6.6 Continuous Random Variable

In the preceding sections, we have gone through the discrete random variables. Now, in this section we would go through another type of random variables known as the *continuous random variables*.

A random variable X is said to be a *continuous random variable* if X takes on any value in an interval or a union of non-overlapping intervals. More formally, we can define a continuous random variable as follows:

The random variable X , $X : S \rightarrow E$, is a continuous random variable if the range or image of the random variable is uncountably infinite, usually over an interval.

Some of the examples for a continuous random variables are as follows:

1. Consider your class of this course IC-252. Suppose the course has a strength of 320 and on a particular day, the logistic TAs mark the height of students present in that class. The random variable taking the height of the students present would be a continuous random variable.
2. Consider your own family. Then, the random variable taking on the weight of people in the family would be a continuous random variable.
3. Consider the parking lot of our college. Then the random variable taking on the volume of air in the vehicles parked there would be a continuous random variable.
4. Consider the number of trees (or even the position of species) on the Griffon peak. The random variable taking on the distance between any two trees (or the species) there, would be a continuous random variable.

The tools developed so far for the discrete random variable can be used to develop the tools for a continuous random variable, as the theory of continuous random variables is completely analogous to the theory of discrete random variables. In the next sub-section, we would discuss about the probability density function, which is used to characterize the probabilities that the continuous random variable would take on.

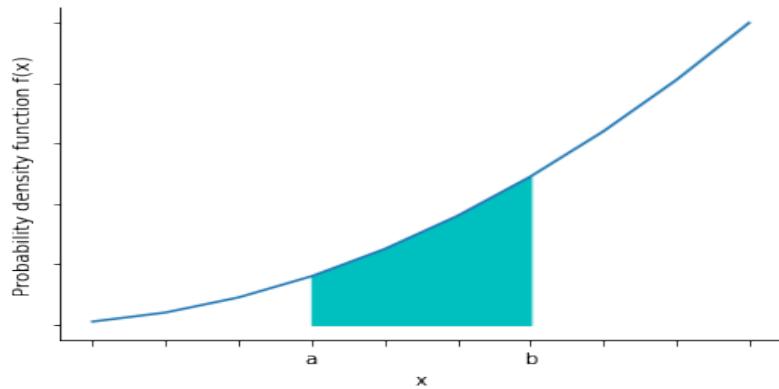


Fig. 6.4: Probability density function

6.6.1 Probability Density Function

In the case of a discrete random variable, X , which takes on a finite or countably infinite number of possible values, we defined the *probability mass function* as $P(X = x)$, for all possible values of X . However, for a continuous random variable, we can't define the probability that it takes a particular value, as it is zero, i.e., $P(X = 0) = 0$. We will soon discuss why this happens.

Instead we have to work on finding the probability that a continuous random variable X falls in some interval a, b , i.e., $P(a < X < b)$. The probability that the random variable, X , lies between two values a and b is obtained by integrating the *probability density function* or *pdf* between these two values. This probability is the area under the *probability density function* between the points a and b as shown in the figure 6.4.

Let us now formally define *probability density function* or *pdf* as follows:

A *probability density function* $f(x)$ describes the probabilistic properties of a continuous random variable and is defined by the following properties:

1. $f(x) \geq 0 \quad \forall x$
2. $\int_{-\infty}^{\infty} f(x) = 1$
3. $P(a < X < b) = \int_a^b f(x)$

Now, we will understand the *probability density function* or *pdf* through an example.

Example 1: Assume that Drongo canteen advertises that the hamburger is of 0.25 pounds weight. However, one randomly selected hamburger might weigh 0.23 pounds while another might weigh 0.27 pounds. What is the probability that a randomly selected hamburger weighs between 0.20 and 0.30 pounds?

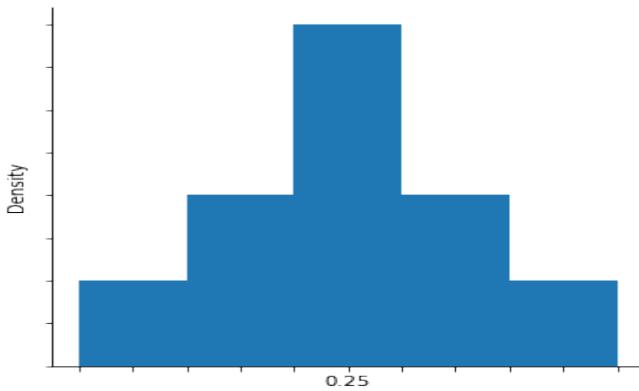


Fig. 6.5: Histogram of weights

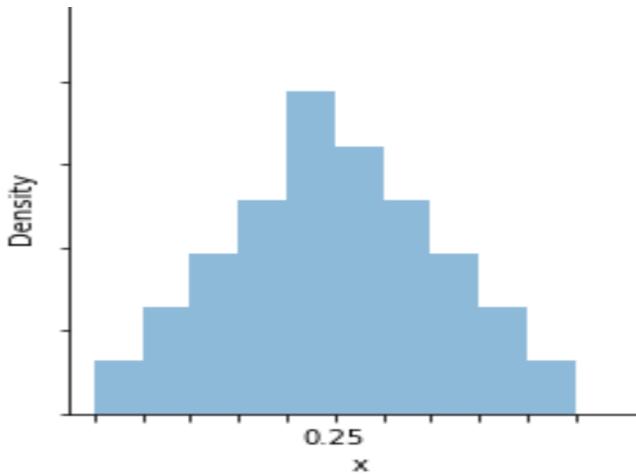


Fig. 6.6: Histogram of weights

Solution: Let us consider that the random variable X denote the weight of a randomly selected hamburger in pounds. Then, we have to find $P(0.2 < X < 0.3)$.

Now, imagine that we have randomly selected, for say, 100 hamburgers advertised to weigh a quarter-pound. If we weighed the 100 hamburgers, and created a density histogram of the resulting weights, its histogram might look as shown in figure 6.5:

The histogram illustrates that most of the sampled hamburgers do indeed weigh close to 0.25 pounds, but some are a bit more and some a bit less. Now, if we decreased the length of the class interval on that density histogram, it would look like the below histogram as shown in figure 6.6:

Now, if we push this further and decrease the intervals even more that the intervals would eventually get so small that we could represent the probability distribution of X , as a curve rather than that of density histogram. This one would be as shown in below figure: 6.7

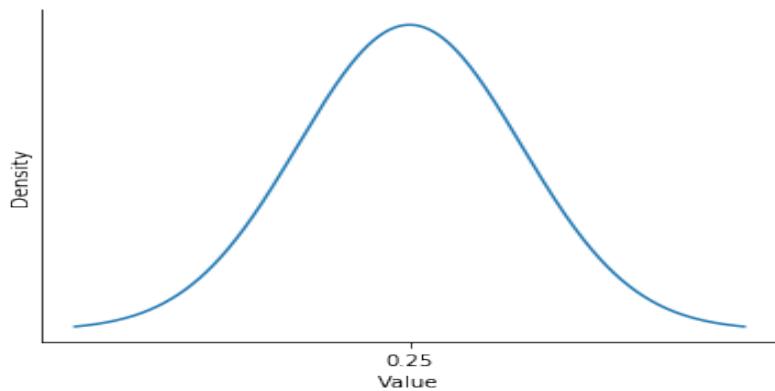


Fig. 6.7: Probability density function

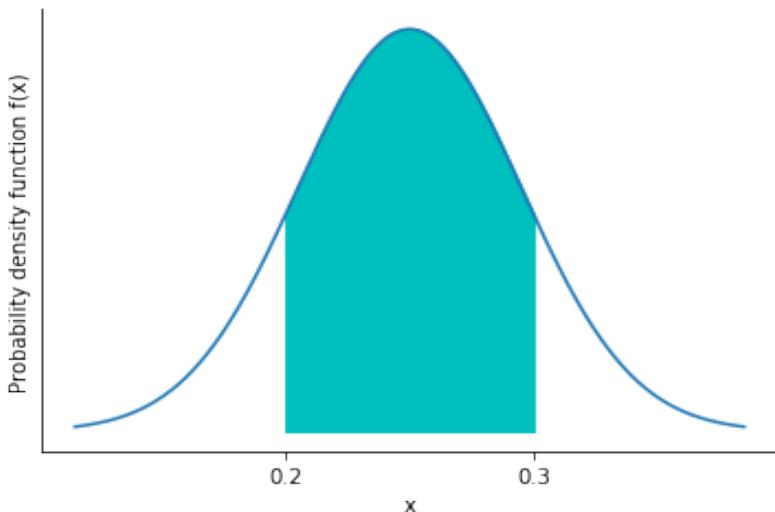


Fig. 6.8: Finding probability using pdf curve

This curve is the *probability density function* or *pdf*. As we know that a density histogram is defined so that the area of each rectangle equals the relative frequency of the corresponding class, and the area of the entire histogram equals 1. This suggests that finding the probability that a continuous random variable X falls in some interval of values involves finding the area under the curve $f(x)$ sandwiched by the endpoints of the interval. So here, the probability that a randomly selected hamburger weighs between 0.20 and 0.30 pounds is the highlighted area as shown in the figure 6.8:

Example 2: Let X be a continuous random variable, whose *pdf* is defined by

$$f(x) = 3x^2, \quad 0 < x < 1.$$

Find the probability that X falls between 0.5 and 1.

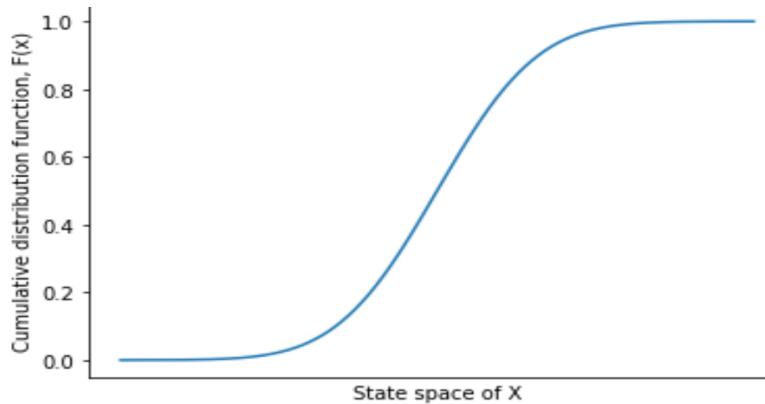


Fig. 6.9: Cumulative distribution function

Solution: Given,

$$f(x) = 3x^2, \quad 0 < x < 1.$$

Then, the required probability, $P(0.5 < X < 1)$ is given as

$$\begin{aligned} P(0.5 < X < 1) &= \int_{x=0.5}^1 f(x)dx \\ &= \int_{x=0.5}^1 3x^2 dx = x^3 \Big|_{0.5}^1 = 1 - \frac{1}{8} = \frac{7}{8} \end{aligned}$$

6.6.2 Cumulative Distribution Function

The *cumulative distribution function* of a continuous random variable X is defined in exactly the same way as for a discrete random variable, which is as follows:

$$F(x) = P(X \leq x)$$

For a continuous random variable, X , the cumulative distribution function $F(x)$ is a continuous non-decreasing function that takes the value 0 prior to and at the beginning of the state space (the range of the random variable) and increases to a value of 1 at the end of and after the state space as shown in the figure 6.9

Like the probability density function, the cumulative distribution function summarizes the probabilistic properties of a continuous random variable, and knowledge of either function allows the other function to be constructed. Suppose say, we have the *pdf* as $f(x)$, then we can get the *cumulative distribution function* as

$$F(x) = \int_{-\infty}^x f(x)dx$$

We will go through an example to gain more insights of the *cumulative distribution function*.

Example: Find the *cdf* for example 2 of the preceding section.

Solution: Given, the *pmf* of the random variable X is

$$f(x) = 3x^2, \quad 0 < x < 1.$$

We know that the *cdf* of a given *pdf*, $f(x)$, is given as

$$F(x) = \int_{-\infty}^x f(x)dx$$

Since, the *pdf* exists only for the interval $0 < x < 1$, we can obtain the required *cdf* as

$$F(x) = \int_0^x 3y^2 dy = y^3]_0^x = x^3, \quad 0 < x < 1$$

So, the *cdf* is

$$F(x) = x^3, \quad 0 < x < 1.$$

6.6.3 Special Distributions

Here, we would be discussing some of the commonly encountered continuous distribution functions, in a similar way, we discussed for the discrete random variables.

6.6.3.1 Uniform Distribution

One of the simplest continuous distribution function, one can think is of would be the a distribution having a flat response over an interval. Such a distribution is the *uniform distribution*, denoted as $U(a, b)$, where the interval under consideration is (a, b) . As we know that the area under the *pdf* must be unity, the height of the curve must be $\frac{1}{b-a}$.

One of the best example would be the case of hitting point on dartboard by throwing dart. Here, each and every point on the dartboard has an equal probability to get hit. Now, let us formally define it as follows:

A random variable X with a flat probability density function, $f(x)$, over an interval (a, b) such

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

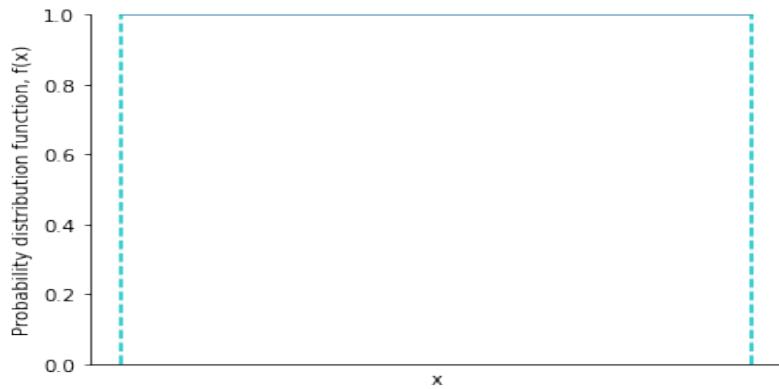


Fig. 6.10: Probability density function of a uniform random variable

is said to have a *uniform distribution*, denoted as $X \sim U(a, b)$.

It's *cdf* is given as

$$F(x) = \int_a^x \frac{1}{b-a} dy = \frac{1}{b-a} \int_a^x dy = \frac{x-a}{b-a}, \quad a \leq x \leq b$$

The expectation and variance of the random variable $X \sim U(a, b)$ are given as

$$E[X] = \frac{a+b}{2}$$

$$Var(X) = \frac{(b-a)^2}{12}$$

The *probability density function* of the uniform distribution for $a = 0$ and $b = 1$ is as shown in the figure 6.10.

Example: Consider a random X , taking on the diameter of screws (in mm), has a *pdf*, having a flat response between 0 and 10. Find its expectation, variance and the probability that X would be at least 4mm.

Solution: Given, $X \sim U(0, 10)$. Then,

$$E[X] = \frac{a+b}{2} = \frac{0+10}{2} = 5.$$

$$Var(X) = \frac{(10-0)^2}{12} = 8.33.$$

$$P(X \geq 4) = 1 - F(4) = 1 - \frac{4-0}{10-0} = 1 - \frac{4}{10} = 0.6$$

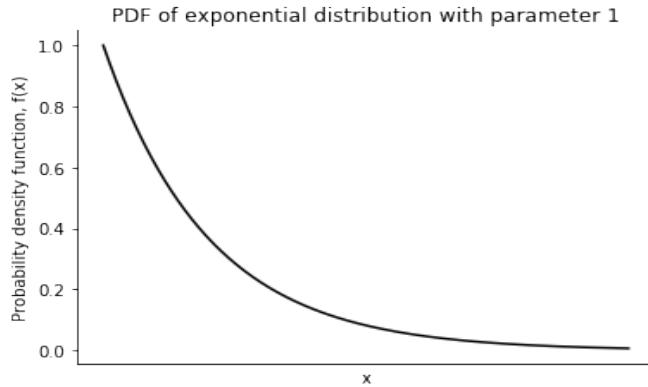


Fig. 6.11: Probability density function of an exponential random variable with $\lambda = 1$

6.6.3.2 Exponential Distribution

One of the most frequently occurred problem in our day-to-day life is to model the waiting time or failure time. For example, we need to find the probability that a mechanical components fails with a certain probability or a bus arrives in an interval. In these situations, the *exponential distribution* comes into picture. Since it deals mostly with time or non-negative entities, it has a range of only non-negative numbers. It's formal definition is as follows:

An *exponential distribution* with rate parameter $\lambda > 0$ has a *probability distribution function* given as

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and a *cumulative distribution function* as

$$F(x) = \int_0^x \lambda e^{-\lambda y} dy = \lambda \int_0^x e^{-\lambda y} dy = \lambda * \left(-\frac{1}{\lambda}\right) * e^{-\lambda y}]_0^x = 1 - e^{-\lambda x}, \quad x \geq 0.$$

The expectation and variance of the random variable $X \sim \exp(\lambda)$ are given as

$$E[X] = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}.$$

The *probability density function* of the exponential distribution, with parameter $\lambda = 1$, is as shown in the figure 6.11.

An important property of the *exponential distribution* is its memoryless property. This property states that if X has an exponential distribution with parameter λ , then conditional on $X \geq x_0$ for some fixed value x_0 , the quantity $X - x_0$ also has an exponential distribution with parameter λ . In

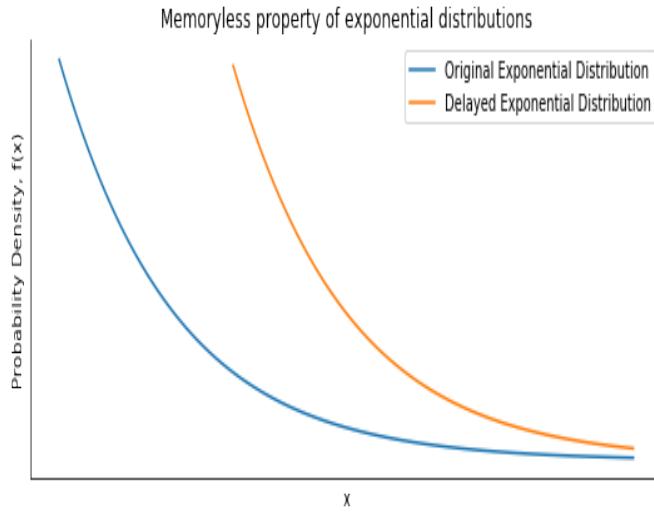


Fig. 6.12: Memoryless property of an exponential random variable

simple words, if X measures the time until a certain event occurs and the event has not occurred by time x_0 , the additional waiting time for the event to occur beyond x_0 has the same exponential distribution as X . The process seems to “forget” that a time x_0 has already elapsed and acts as though it is just starting afresh at time zero.

Mathematically, it can be shown as

$$P(X \geq x) = 1 - F(x) = e^{-\lambda x}$$

Then if the random variable Y represents the additional time beyond x_0 that elapses before the event occurs,

$$P(Y \geq y) = P(X \geq x_0 + y | X \geq x_0) = \frac{P(X \geq x_0 + y)}{P(X \geq x_0)} = \frac{e^{-\lambda(x_0+y)}}{e^{-\lambda x_0}} = e^{-\lambda y}$$

so that Y also has an exponential distribution with parameter λ .

This property is shown in the below figure 6.12. In graphical terms, the memoryless property follows from the fact that the section of the probability density function of an exponential distribution beyond a certain point x_0 is just a scaled version of the whole *probability density function* as shown in the figure 6.12 [Anthony textbook on probability].

To understand this property better, let us discuss an example. Suppose that we are waiting at a bus stop and that the time in minutes until the arrival of the bus has an exponential distribution

with $\lambda = 0.2$. The expected time that we will wait is consequently $\frac{1}{\lambda} = 5$ minutes. However, if after 1 minute the bus has not yet arrived, what is the expectation of the additional time that we must wait?

Unfortunately, it won't be reduced to 4 minutes but is still, as before, 5 minutes. This is because the additional waiting time until the bus arrives beyond the first minute during which we know the bus did not arrive still has an exponential distribution with $\lambda = 0.2$. In fact, as long as the bus has not arrived, no matter how long we have waited, we always have an expected additional waiting time of 5 minutes. This is true right up until the time we first spot the bus coming.

Example: Engineers observe that about 90% of graphite samples fracture within 5 hours when subjected to a certain stress. Find

1. If the time to fracture is modeled with an exponential distribution, what would be a suitable value for the parameter λ ?
2. Use the model to estimate the probability that a fracture occurs within 3 hours.

Solution: Given, the probability of observing graphite sample fracture within 5 hours, subjected to stress, is 0.90.

$$\Rightarrow F(5) = 0.90$$

1. Since, $F(5) = 0.90$, we can write

$$F(5) = 1 - e^{-\lambda 5} = 0.9$$

$$\Rightarrow e^{-\lambda 5} = 0.1$$

$$\Rightarrow \lambda = 0.4605.$$

2. Required probability is

$$F(3) = 1 - e^{-3\lambda} = 1 - e^{-3*0.4605} = 0.75$$

6.6.3.3 Normal Distribution

The *normal* or the *Gaussian distribution* is the most important and frequently used continuous probability distributions for many statistical inferences. It models many naturally occurring phenomena.

The random variable $X \sim N(\mu, \sigma^2)$ following a *normal* or the *Gaussian distribution* has a *probability density function*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty \leq x \leq \infty$$

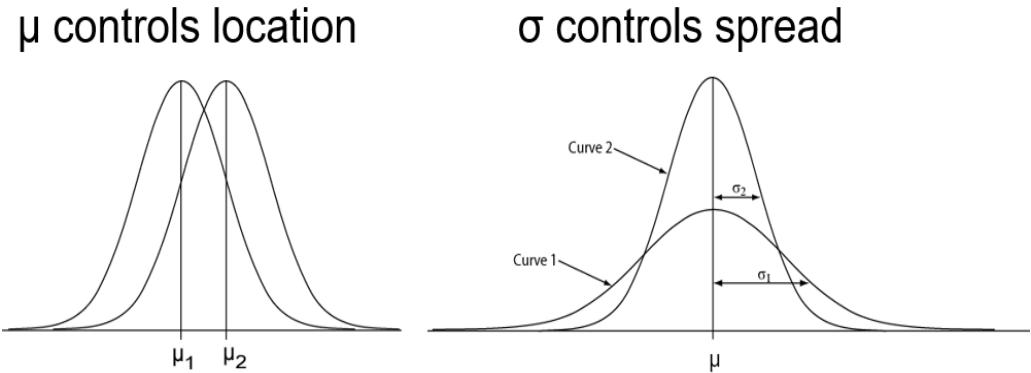


Fig. 6.13: Effect of the parameters of a normal random variable

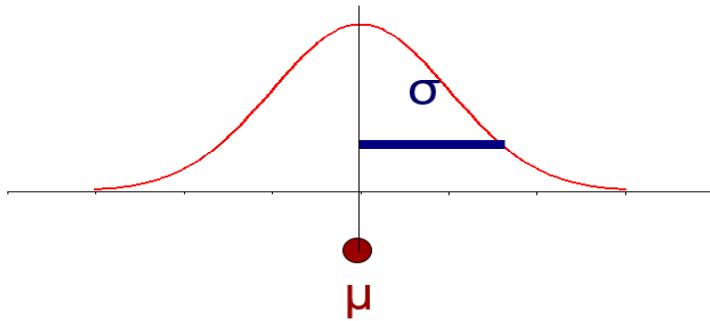


Fig. 6.14: Probability density function of a normal random variable

depending on two parameters, the mean and the variance

$$E[X] = \mu$$

$$\text{Var}(X) = \sigma^2.$$

The two parameters define the shape and position of the *pdf* curve. The mean, μ , controls the location of the curve while the standard deviation (so the variance), σ , controls the spread of the *pdf* curve. Higher the value of σ , the more is the spread in the concerned data. Here, spread is in reference to the mean. This is summarized in the below figure 6.13.

The *pdf* of a *normal distribution* is a bell shaped curve, symmetrical about μ , as shown in figure 6.14. Large values of the variance σ^2 result in long, flat bell-shaped curves, whereas small values of the variance σ^2 result in thinner, sharper bell-shaped curves. There is no simple closed-form solution for the cumulative distribution function of a normal distribution.

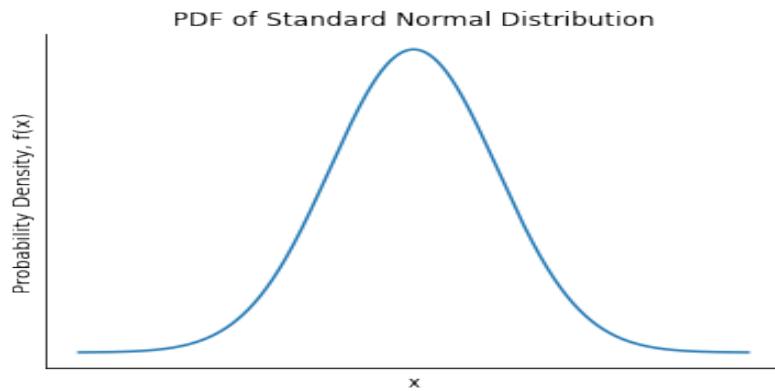


Fig. 6.15: Probability density function of a standard normal random variable

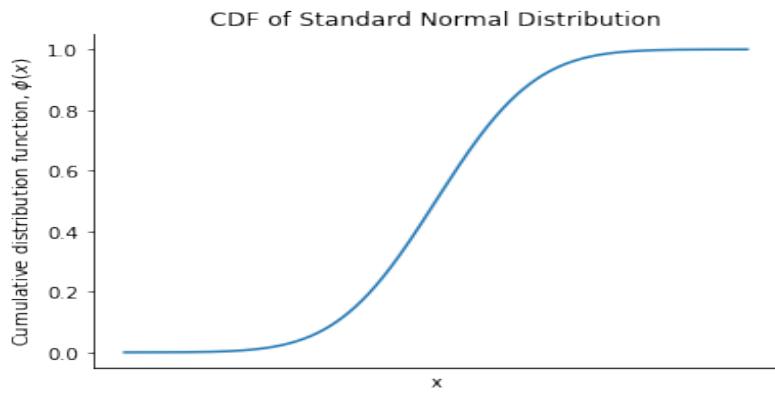


Fig. 6.16: Cumulative distribution function of a standard normal random variable

The *normal distribution* having the values of $(\mu, \sigma^2) = (0, 1)$ is known as the *standard normal distribution*. It is of importance because it eases the process to find the probability of the normal random variable. Before diving into this, let us discuss about the *standard normal distribution*.

The *pdf* of the *standard normal distribution*, $\phi(x)$, is given as

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty \leq x \leq \infty$$

It is as shown in the below figure 6.15.

The *cdf* of the *standard normal distribution*, $\Phi(x)$, is given as

$$\Phi(x) = \int_{-\infty}^x \phi(y) dy$$

It is as shown in the below figure 6.16.

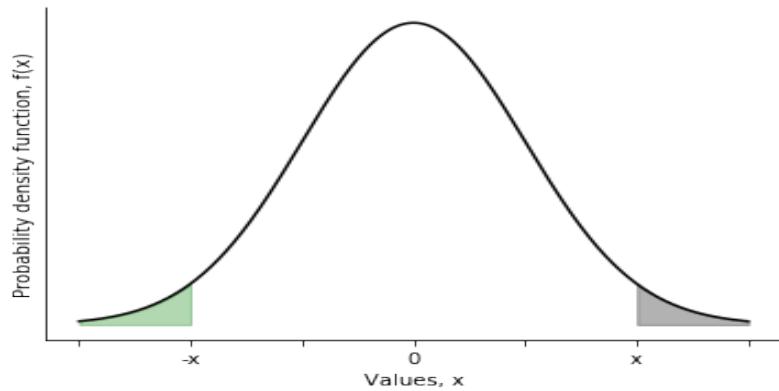


Fig. 6.17: Symmetry of the standard normal random variable

The symmetry of the *standard normal distribution* about 0 implies that if the random variable Z has a *standard normal distribution*, then

$$1 - \Phi(x) = P(Z \geq x) = P(Z \leq -x) = \Phi(-x)$$

as shown in the below figure 6.17. The area highlighted in the green color is the $\phi(-x)$ and the area highlighted in black is $1 - \phi(x)$, which are numerically equal.

$$\Rightarrow \Phi(x) + \Phi(-x) = 1.$$

The values for $\Phi(x)$ is given in a tabular form, in figures 6.31 and 6.32, and is used to find the probabilities. For example, if we want to find the probability of Z being 1.96, then along the *tenths*, we would find the location of 1.9 and along the *hundredths*, we would find the location of 0.06, then at their intersection point, we would get the value of probability that $Z = 1.96$, which is 0.9750. This is summarized in the figure 6.18.

Now, we will see how this would be useful in finding the probabilities of the *normal random variable*.

If $X \sim N(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

The random variable Z is known as the *standardized* version of the random variable X . This result implies that the probability values of a general normal distribution can be related to the cumulative distribution function of the standard normal distribution $\Phi(x)$ through the relationship

$$P(a \leq X \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

z tenths	hundredths									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Fig. 6.18: Z table example

Let us consider $X \sim N(3, 4)$. Then,

$$P(X \leq 6) = P(-\infty \leq X \leq 6) = \Phi\left(\frac{6-3}{2}\right) - \Phi\left(\frac{\infty-3}{2}\right) = \Phi(1.5) - \Phi(\infty)$$

from the tables we get the value of $\Phi(1.5)$ as 0.9332.

$$\Rightarrow P(X \leq 6) = 0.9332 - 0 = 0.9332.$$

Now, we will discuss some of the most important properties of a *normal distribution*. From above discussions, we knew how to obtain the probability of a *normal* random variable using its standardized version. Now, using the same we would discuss the significant property of the *normal distribution* known as the **68-95-99.7 rule**. This means that 68%, 95%, and 99.7% of the values lie within one, two, and three standard deviations of the mean, respectively.

This three-sigma rule of thumb (or 3σ rule) means that nearly all values are taken to lie within three standard deviations of the mean of any *normally distributed* random variable. Thus we can consider 99.7% probability as near certainty.

Mathematically, for a *normally distributed* random variable, X , it is given as

1. $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68.27\%$

Proof. Consider that the random variable X is *normally distributed*, $X \sim N(\mu, \sigma)$. Then, the probability that the values taken would be within one standard deviation is given by

$$P(\mu - \sigma \leq X \leq \mu + \sigma)$$

To find this, we would standardize X (by making its mean zero and variance one), as follows:

Subtracting mean on both sides of the inequality

$$= P(-\sigma \leq X - \mu \leq \sigma)$$

Dividing standard deviation on both sides of the inequality, we get

$$= P\left(-1 \leq \frac{X - \mu}{\sigma} \leq 1\right)$$

Since, $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$, we get proceed as follows

$$= P\left(-1 \leq \frac{X - \mu}{\sigma} \leq 1\right) = P(-1 \leq Z \leq 1)$$

$$\begin{aligned}
&\Rightarrow P(-1 \leq Z \leq 1) = \phi(1) - \phi(-1) \\
&= \phi(1) - \phi(-1) = \phi(1) - (1 - \phi(1)) = 2\phi(1) - 1 \\
&= 2(0.8413) - 1 = 1.68269 - 1 = 0.6827
\end{aligned}$$

Thus, around 68.27% of the values of X fall within one standard deviation. \square

2. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95\%$

Proof. Consider that the random variable X is *normally distributed*, $X \sim N(\mu, \sigma)$. Then, the probability that the values taken would be within two standard deviations is given by

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$$

To find this, we would standardize X (by making its mean zero and variance one), as follows:

Subtracting mean on both sides of the inequality

$$= P(-2\sigma \leq X - \mu \leq 2\sigma)$$

Dividing standard deviation on both sides of the inequality, we get

$$= P\left(-2 \leq \frac{X - \mu}{\sigma} \leq 2\right)$$

Since, $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$, we get proceed as follows

$$\begin{aligned}
&= P\left(-2 \leq \frac{X - \mu}{\sigma} \leq 2\right) = P(-2 \leq Z \leq 2) \\
&\Rightarrow P(-2 \leq Z \leq 2) = \phi(2) - \phi(-2) \\
&= \phi(2) - \phi(-2) = \phi(2) - (1 - \phi(2)) = 2\phi(2) - 1 \\
&= 2(0.9772) - 1 = 1.9544 - 1 = 0.9544
\end{aligned}$$

Thus, around 95.44% of the values of X fall within two standard deviations. \square

3. $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.7\%$

Proof. Consider that the random variable X is *normally distributed*, $X \sim N(\mu, \sigma)$. Then, the probability that the values taken would be within three standard deviations is given by

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma)$$

To find this, we would standardize X (by making its mean zero and variance one), as follows:

Subtracting mean on both sides of the inequality

$$= P(-3\sigma \leq X - \mu \leq 3\sigma)$$

Dividing standard deviation on both sides of the inequality, we get

$$= P\left(-3 \leq \frac{X - \mu}{\sigma} \leq 3\right)$$

Since, $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$, we get proceed as follows

$$= P\left(-3 \leq \frac{X - \mu}{\sigma} \leq 3\right) = P(-3 \leq Z \leq 3)$$

$$\Rightarrow P(-3 \leq Z \leq 3) = \phi(3) - \phi(-3)$$

$$= \phi(3) - \phi(-3) = \phi(3) - (1 - \phi(3)) = 2\phi(3) - 1$$

$$= 2(0.9987) - 1 = 1.9974 - 1 = 0.9974$$

Thus, around 99.74% of the values of X fall within three standard deviations. \square

We know that for a normal distribution, the area under the curve within 1 standard deviation of the mean is 0.68. So the area outside of this region is $1 - 0.68 = 0.32$. And since normal curves are symmetric, this outside area of 0.32 is evenly distributed between the two outer tails. So the area of each tail

$$= \frac{1}{2}(1 - 0.68) = \frac{1}{2}(0.32) = 0.16$$

One of the most frequently used and relatable application of the *normal distribution* is the modelling of the marks obtained by the students in an exam of a course. Before going further, let us study some definitions which would help us in understanding this example better.

Skewness is a measure of the asymmetry of the distribution. A distribution is asymmetrical when its left and right side are not mirror images. A distribution can have right (or positive), left (or negative), or zero skewness. A right-skewed distribution is longer on the right side of its peak, and a left-skewed distribution is longer on the left side of its peak. These are as shown in the figures 6.19, 6.20 and 6.21:

From the distribution parameters, i.e., mean and median, we can say whether the given distribution is right skewed or left skewed or no-skew. If mean = median = mode, then the distribution has no skewness and it is symmetric around the mean value. The *normal distribution* is the best example

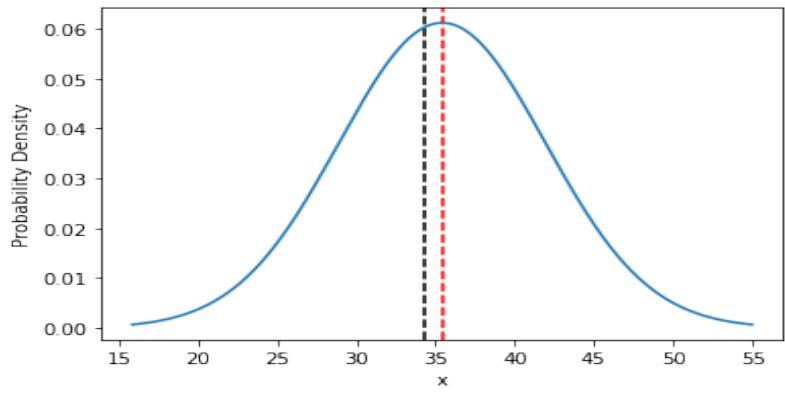


Fig. 6.19: Right skewed distribution

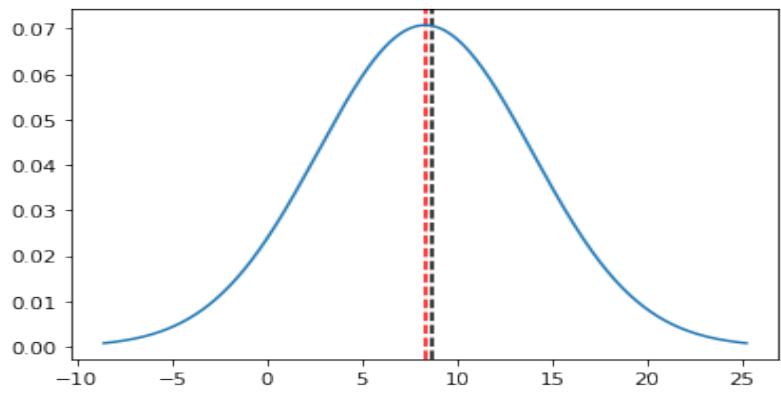


Fig. 6.20: Left skewed distribution

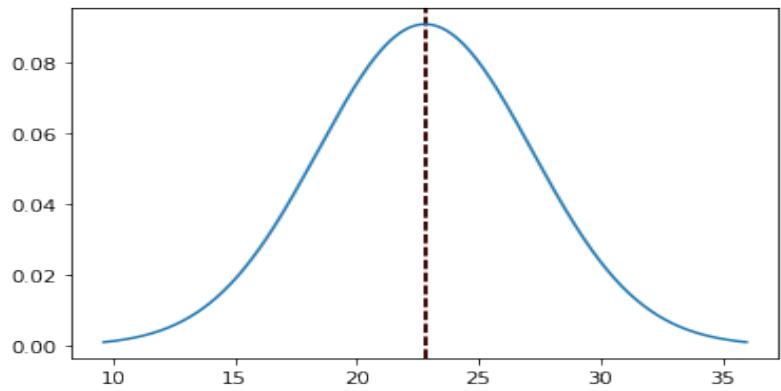


Fig. 6.21: Symmetric distribution

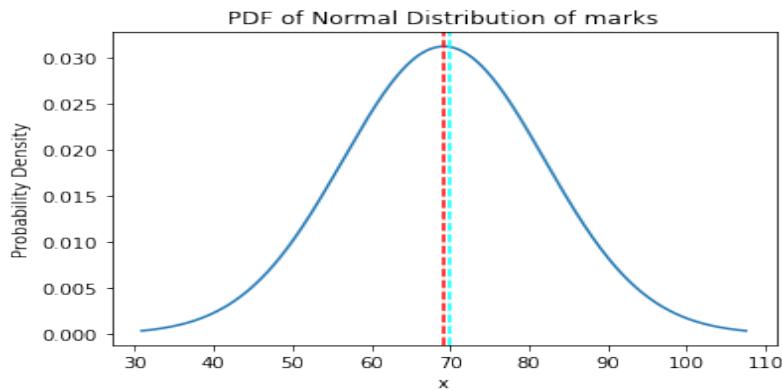


Fig. 6.22: Symmetric distribution of marks

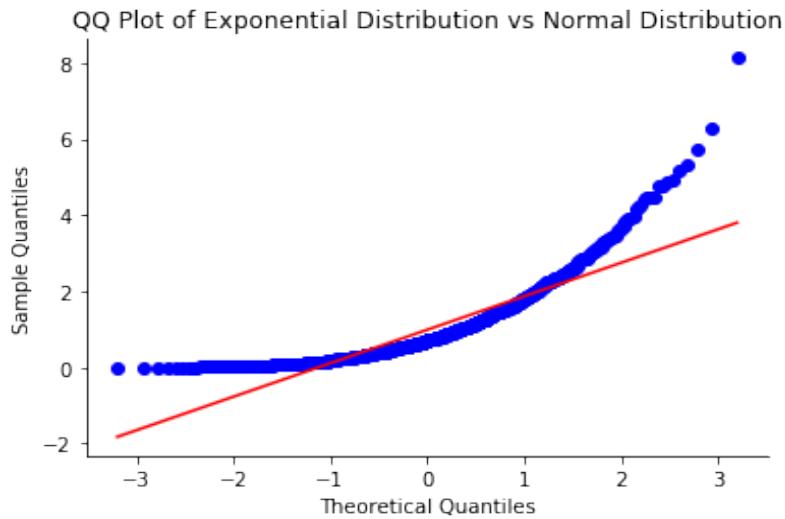


Fig. 6.23: $Q - Q$ plot of a data which isn't normally distributed

for this. If $\text{mean} > \text{median} > \text{mode}$, then the distribution is right skewed and if $\text{mean} < \text{median} < \text{mode}$, then the distribution is left skewed.

Now, we would see what would happen if the marks are modelled by right skewed distribution, left skewed and no skew distribution. The consequence of the distribution being a right skewed is that most of the students got less grades and there are very relatively few students who scored good scores in the exam. The consequence of the distribution being a left skewed is that most of the students got better grades and there are very relatively few students who scored less or poor scores in the exam. In either case, only some of the students do influence the overall average of the class' score. So, we resort to the *normal distribution* while assigning the thresholds to the grades so that the class average isn't influenced by some section of scores. In most practical cases, we prefer that most of the students' grades are centered around *B* grade and only few of them should get *O* and *E*. For say, for toy dataset

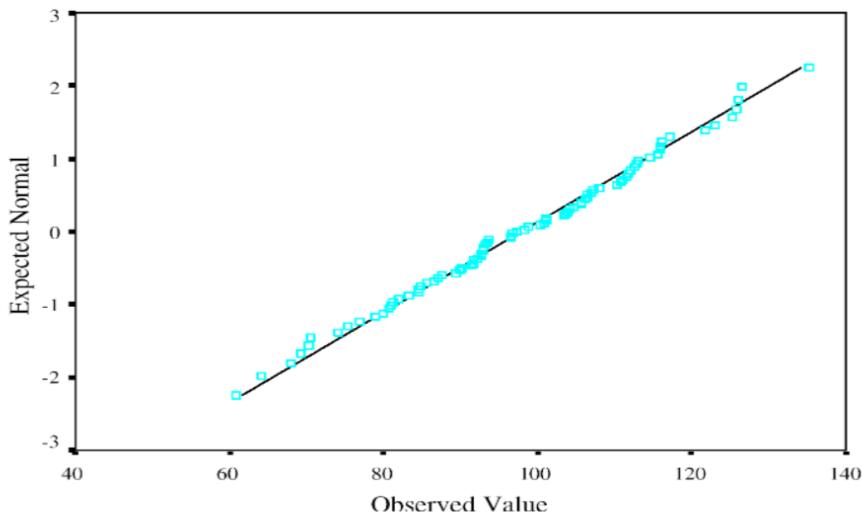


Fig. 6.24: $Q - Q$ plot of a data which is normally distributed

of marks, we get the mean as 69.189 and median as 70 and its pdf would be as shown in the figure 6.22

So far we have gained enough background of the *normal distribution*. Now, we will see how to graphically determine, whether a given data set is distributed normally or not. For this, our visualizing tool would be the *quantile-quantile plot* or *Q-Q plot*. They enable us to compare two probability distributions by plotting their quantiles against each other. If the two distributions under consideration are exactly equal then the points on the *Q-Q* plot will perfectly lie on a straight line $y = x$.

Initially, we plot the theoretical quantiles (of the standard normal variate) on the x -axis and the values of other distribution for which we want to check whether it is normally distributed or not, on the y -axis. After plotting this, we need to check the ends of the straight line. If the points at both the end of the curves aren't falling on the straight line and in fact, are significantly away from the ends of the line, then the data under consideration isn't normally distributed. On the other hand, if the points falls exactly on the ends of the line, we can conclude that the data is indeed normally distributed. The figures 6.23 and 6.24 shows the $Q - Q$ plots for data which isn't and is normally distributed, respectively.

These plots can also be used to find the Skewness of a distribution. If the bottom end of the Q-Q plot deviates from the straight line but the upper end is not, then the distribution has a longer tail to its left and hence is *left-skewed* (or negatively skewed) but when we the upper end of the Q-Q plot to deviate from the straight line and the lower and follows a straight line then the curve has a longer tail to its right and it is *right-skewed* (or positively skewed). The $Q - Q$ plots for the skewed data is shown in figures 6.25 and 6.26.

Now we will discuss about approximating the binomial distribution with normal distribution. Let

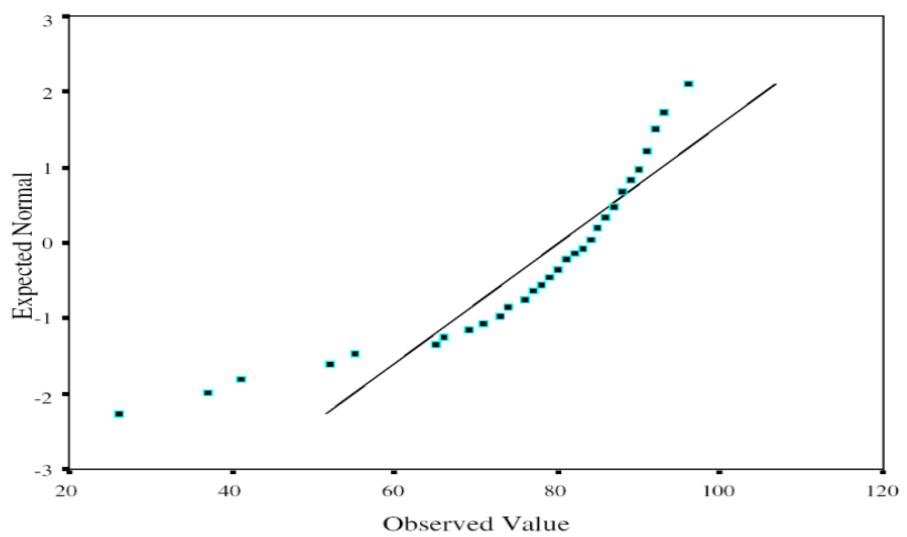


Fig. 6.25: $Q - Q$ plot of a left skewed distributed data

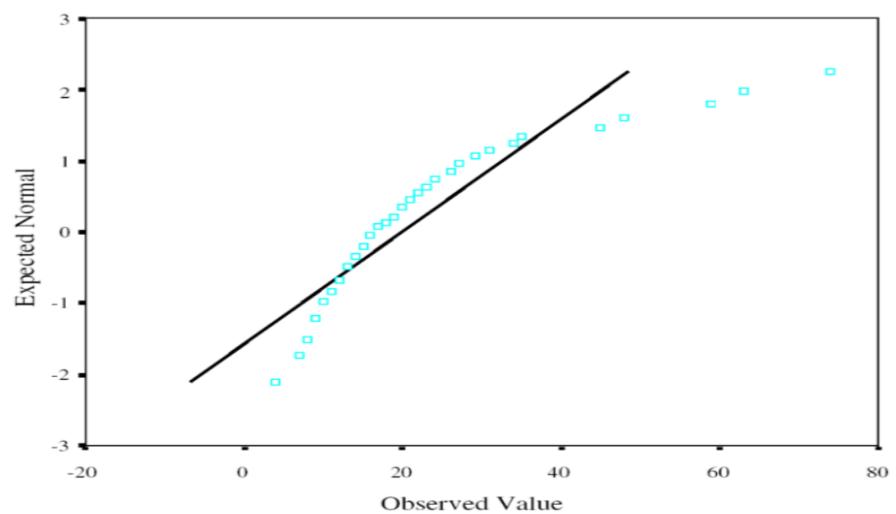


Fig. 6.26: $Q - Q$ plot of a right skewed distributed data

us consider the random variable $X \sim Bin(n, p)$. Then, for large n and a p , which isn't too close to 0 or 1, X can be approximated as $N(0, 1)$. Mathematically, it is

$$X \sim Bin(n, p) \approx Y \sim N(np, \sqrt{np(1-p)})$$

as $n \rightarrow \infty$.

Proof. We need to prove that the distribution of $X \sim Bin(n, p)$ is asymptotically equal to the distribution of the random variable $Y \sim N(np, \sqrt{np(1-p)})$. The random variable $Y \sim N(np, \sqrt{np(1-p)})$ has the following pdf

$$\frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(x-np)^2}{2np(1-p)}}$$

We consider $P(X = k)$ and assume that k doesn't deviate too much from np . Further we assume that $k - np$ should be of order \sqrt{n} .

From Stirling's formula, we get

$$m! \sim \sqrt{2\pi m} e^{-m} m^m$$

Since, $X \sim Bin(n, p)$, we get

$$\begin{aligned} P(X = k) &= \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \\ &\sim \frac{\sqrt{2\pi n} e^{-n} n^n}{\sqrt{2\pi k} e^{-k} k^k \sqrt{2\pi m} e^{-(n-k)} (n-k)^{n-k}} p^k (1-p)^{n-k} \\ &= \left(\frac{p}{k}\right)^k \left(\frac{1-p}{n-k}\right)^{n-k} n^n \sqrt{\frac{n}{2\pi k(n-k)}} \\ &= \left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k} \sqrt{\frac{n}{2\pi k(n-k)}} \end{aligned}$$

From standard identities, we get

$$\ln\left(\frac{np}{k}\right) = -\ln\left(1 + \frac{k-np}{np}\right),$$

$$\ln\left(\frac{n(1-p)}{n-k}\right) = -\ln\left(1 - \frac{n-k}{n(1-p)}\right),$$

Then, by using $\ln(1+y) \sim y - \frac{y^2}{2} + \frac{y^3}{3}$, $y \rightarrow 0$, we get

$$\ln\left(\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k}\right) = k \ln\left(\frac{np}{k}\right) + (n-k) \ln\left(\frac{n(1-p)}{n-k}\right)$$

$$\begin{aligned} &\sim k\left(-\frac{k-np}{np} + \frac{1}{2}\left(\frac{k-np}{np}\right)^2 - \frac{1}{3}\left(\frac{k-np}{np}\right)^3\right) + (n-k)\left(-\frac{k-np}{n(1-p)} + \frac{1}{2}\left(\frac{k-np}{n(1-p)}\right)^2 - \frac{1}{3}\left(\frac{k-np}{n(1-p)}\right)^3\right) \\ &\sim -\frac{(k-np)^2}{2np(1-p)} \end{aligned}$$

Thus,

$$\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k} \sim e^{-\frac{(k-np)^2}{2np(1-p)}}$$

At the beginning of this proof, we have assumed that $k - np$ is of order \sqrt{n} ,

$$\begin{aligned} k - np &\approx \sqrt{n} \\ \Rightarrow k &\approx np + \sqrt{n} \\ \Rightarrow n - k &\approx n - (np + \sqrt{n}) \approx n(1-p) - \sqrt{n} \\ \Rightarrow k(n - k) &\approx n^2 p(1-p) \\ \Rightarrow \sqrt{\frac{n}{2\pi k(n - k)}} &\approx \frac{1}{\sqrt{2\pi np(1-p)}}. \\ \Rightarrow P(X = k) &\approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}. \end{aligned}$$

However, we need to note that the binomial distribution is a discrete distribution and the normal distribution is a continuous distribution. Hence, a continuity correcting factor of 0.5 is added to make the approximation better. So, if $X \sim Bin(n, p)$, then

$$\begin{aligned} P(X \leq x) &\approx \phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right), \quad \text{and,} \\ P(X \geq x) &\approx 1 - \phi\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

□

Now, we would go through an example to understand this approximation.

Example: Consider the distributions $Bin(16, 0.5)$ and $N(8, 4)$. The shapes of $Bin(16, 0.5)$ and $N(8, 4)$ are quite similar even though the former distribution is a discrete random variable and the latter is a continuous random variable. This similarity is as shown in the figure 6.27.

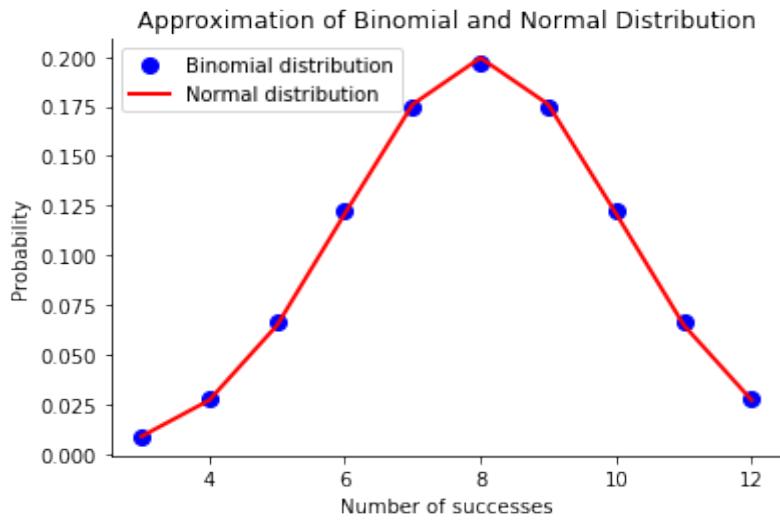


Fig. 6.27: Comparison of the *pmf* of $Bin(16, 0.5)$ with the *pdf* of $N(8, 4)$.

Then, the probability that the binomial takes no larger than 5 is

$$P(X \leq 5) = \sum_{x=0}^5 \binom{16}{x} (0.5)^x (0.5)^{16-x} = 0.1051$$

Now, the approximation using the normal distribution is

$$P(Y \leq 5.5) = \phi\left(\frac{5.5 - \mu}{\sigma}\right) = \phi\left(\frac{5.5 - 8}{2}\right) = \phi(-1.25) = 1 - \phi(1.25) = 0.1056$$

6.7 Theory Assignment 6

1. For $X \sim N(\mu, \sigma^2)$, verify that
 - (a) the mean is μ
 - (b) the variance is σ^2
2. Eve sends a message to Mallory via the internet, an insecure channel. Acknowledgement received by Eve upon successful message transmission(1) is +2 and upon disrupted transmission is -2 summed with a number Y sent by Mallory i.e the content received by Eve is $X+Y$ where $Y \sim N(0, 1)$. If $X+Y$ is greater than 1.5 we consider the internet to be secure and insecure otherwise. What is the probability that
 - (a) Internet is considered secure given that disrupted transmission occurred?
 - (b) Internet is considered insecure given that successful transmission occurred?

3. The time taken to serve a customer at a supermarket has a mean of 65.00 seconds and a standard deviation of 7.00 seconds. Using Chebyshev inequality, calculate the time interval that has 75% probability of containing a particular service time.
4. Dogs are sterilized. Suppose that a particular shot has a probability of 0.0004 of causing death when administered to a dog. Suppose that the shot is to be administered to 800,000 dogs. Find the approximate value of the probability that at most 200 dogs die.
5. Let X and Z be random variables, such that $X \sim N(0, 1)$ and $Z \sim N(1, 4)$, find
 - (a) $P(|X| \leq 0.31)$
 - (b) the value of x for which $P(X \leq x) = 0.27$
 - (c) the value of x for which $P(|X| \geq x) = 0.44$
 - (d) Upper bound of $P(|Z - 1| \geq 3)$
 - (e) $P(|Z - 1| \geq 3)$

6.7.1 Theory Assignment 6 Solutions:

1. given $f_X(x)$

(a)

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

we know,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

thus,

$$E[X] = \int_{-\infty}^{\infty} x \cdot \left(\frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) dx$$

substitute , $t = (x - \mu)/\sigma\sqrt{2}$

$$E[X] = \sigma\sqrt{2} \int_{-\infty}^{\infty} (\sigma\sqrt{2}t + \mu) \cdot \left(\frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-t^2} \right) dt$$

$$E[X] = \frac{1}{\sqrt{\pi}} \left(\int_{-\infty}^{\infty} (\sigma\sqrt{2}t \cdot e^{-t^2}) dt + \int_{-\infty}^{\infty} (\mu \cdot e^{-t^2}) dt \right)$$

$$E[X] = \frac{\mu\sqrt{\pi}}{\sqrt{\pi}} = \mu$$

(b)

$$Var[X] = E[X^2] - E[X]^2$$

$$Var[X] = \int_{-\infty}^{\infty} x^2 \cdot \left(\frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) dx - \mu^2$$

substitute , $t = (x - \mu)/\sigma\sqrt{2}$

solving yields

$$Var[X] = \sigma^2$$

2. (a)

$$\begin{aligned} P(Secure|disrupted\ transmission) &= P(X + Y > 1.5) \\ &= P(-2 + Y > 1.5) \\ &= P(Y > 3.5) \\ &= 1 - P(Y < 3.5) \\ &= 1 - \phi(3.5) \end{aligned} \tag{6.1}$$

(b)

$$\begin{aligned} P(Insecure|successful\ transmission) &= P(X + Y < 1.5) \\ &= P(2 + Y < 1.5) \\ &= P(Y < -0.5) \\ &= \phi(-0.5) \end{aligned} \tag{6.2}$$

3. Let y be the time taken to serve a customer at a supermarket,

$$\mu = 65.00sec$$

$$\sigma = 7.00sec$$

From Chebyshev Inequality,

$$P(\mu - c\sigma \leq Y \leq \mu + c\sigma) \geq 1 - \frac{1}{c^2}$$

$$1 - \frac{1}{c^2} = 0.75$$

$$c = 2$$

$$P(65 - 14 \leq Y \leq 65 + 14) \geq 0.75$$

Thus, Y lies in interval $[51, 79]$ with probability 0.75.

4. If the underlying distribution of dogs affected is X

$$E[X] = \mu_X = 0.0004$$

$$Var[X] = \sigma_X^2 = 0.0004 * (1 - 0.0004) = 0.0039984$$

We can approximate the distribution as normal distribution $Y \sim N(n\mu_X, n\sigma_X^2)$

where $n = 800,000$

$$Z \sim N(0, 1)$$

$$Y \sim N(320, 319.872)$$

$$Y = 320 + \sqrt{319.872}Z$$

$$P(X \leq 200) \approx P(Y \leq 200)$$

$$P(Y \leq 200) = P(Z \leq \frac{200 - 320}{\sqrt{319.872}})$$

$$P(X \leq 200) = P(Z \leq -6.7095)$$

5. (a)

$$P(|X| \leq 0.31) = P(-0.31 \leq X \leq 0.31)$$

$$P(|X| \leq 0.31) = \phi(0.31) - \phi(-0.31)$$

(b)

$$P(X \leq x) = 0.27$$

$$\phi(X \leq x) = 0.27$$

check the table

(c)

$$P(|X| \geq x) = 0.44$$

$$1 - P(|X| \leq x) = 0.44$$

$$P(|X| \leq x) = 0.56$$

$$1 - (\phi(x) - \phi(-x)) = 0.56$$

$$2 - 2\phi(x) = 0.56$$

$$\phi(x) = \frac{2 - 0.56}{2}$$

(d) We know

$$P(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2}$$

where X has mean μ and variance σ^2

$$P(|Z - 1| \geq 3) \leq \frac{4}{9}$$

(e) $Z = 1 + 4X$

$$P(|Z - 1| \geq 3) = P(|1 + 4X - 1| \geq 3) = P(|4X| \geq 3) = P(|X| \geq 0.75)$$

$$P(-0.75 \leq X \leq 0.75) = 2 \cdot \phi(X \leq 0.75)$$

6.8 Lab Assignment 4

Question 1 In this problem, the random variable X is the sum of the numbers appearing on two die.

1. Write a code that calculates the probabilities $P(X = i)$ where $i \in \{1, 2, 3, \dots, 13\}$. Do all these probabilities add to 1, i.e., $\sum_i P(X = i) = 1$?
2. Your program should throw an error when this isn't true.
3. Make a plot of $P(X = i)$ as a function of i .
4. The *cumulative distribution function* $F(x)$ gives you the probability that the random variable X takes values $\leq x$.
 - (a) Calculate $F(x)$ using a computer program. Your program should be able to calculate $F(x)$ for any given $x \in \mathbb{R}$.
 - (b) Calculate in particular $F(\pi)$ and $F(\sqrt{30})$.
 - (c) Make a plot of $F(x)$ as a function of $x \in [-1, 20]$.

```
1 import math
2 import matplotlib.pyplot as plt
3
4 #a
5 p = []
6 c = 0
7 for i in range(13):
8     if i < 6:
9         p.append(c/36)
10    c+=1
11 else:
12     p.append(c/36)
13    c-=1
```

```

14
15 prob = sum(p)
16 print('sum of probabilities: ',prob)
17 if round(prob, 2) == 1:
18     print('Valid')
19 else:
20     print('Invalid')
21
22 plt.bar([i for i in range(1, 14)], p)
23 plt.show()
24
25 #b
26 x = [i for i in range(-1, 21)]
27 y = []
28 for i in x:
29     if i < 1 or i>13:
30         y.append(0)
31     else:
32         s = sum(p[::(i - 1)])
33         y.append(s)
34
35 plt.bar(x, y)
36 plt.show()
37
38 print('F(pi) = ', y[x.index(math.ceil(math.pi))])
39 print('F(sqrt(30)) = ', y[x.index(math.ceil(math.sqrt(30)))])
```

Question 2 Consider the *random walk problem* in one dimension in which a person takes steps (*step length 1 m*) randomly either to the left (with probability p) or right (with probability $q = 1 - p$). Let us say the person starts from the origin, i.e., $x = 0$, shown as Lamp post in Figure 6.28. In this case, the random variable X is *displacement* of the person from the origin after a given number of steps, say n .

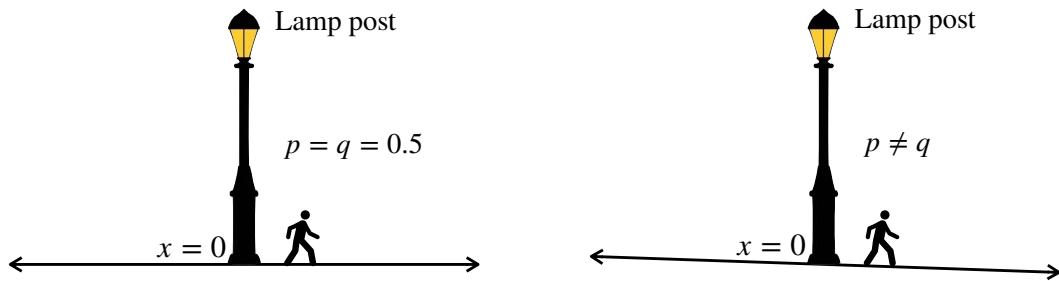


Fig. 6.28: The random walk problem. *Left:* Equal probability of moving in both directions. *Right:* Because of the inclination, the person has more probability of moving towards right as compared to the left direction.

1. Write a computer program and calculate $P(X = -n)$, $P(X = -n + 1)$, ..., $P(X = n - 1)$ and $P(X = n)$ after n steps. Put a check condition in your code that ensures all these probabilities add to 1.
2. Check your program for $n = 5$ and $n = 10$ fixing $p = 0.5$. Note that this corresponds to the left panel of Figure 6.28 where the walker has equal probability of going to the left or right.
3. Using **Matplotlib**, plot the *probability mass function* $P(X)$ as a function of X . In the present case, X is **BINOMIAL RANDOM VARIABLE**.
4. Calculate the cumulative distribution function $F(x) = P(X \leq x)$. Make a plot of $F(x)$.
5. Repeat the above analysis when $p = 0.7$. This situation might arise when there is a slope as shown in the *right* panel of Figure 6.28.

```

1 import math
2 import matplotlib.pyplot as plt
3
4 def prob(n, x, p):
5     r = int((n + x)/2)
6     return math.comb(n, r) * math.pow(p, r) * math.pow(1-p, n-r)

```

```

7
8 def list_p(n, p):
9     l = []
10    a = []
11    for x in range(-n, n+1, 2):
12        a.append(x)
13        l.append(prob(n, x, p))
14    return a, l
15
16 def dis(n, p):
17    l = []
18    a = []
19    s = 0
20    for x in range(-n, n+1, 2):
21        a.append(x)
22        s+=prob(n, x, p)
23        l.append(s)
24    return a, l
25
26 def check(n, p):
27    x, y = list_p(n, p)
28    s = sum(y)
29    print(f'For n = {n} and p = {p}')
30    if round(s, 2) == 1:
31        print('Valid PMF, sum = ', s)
32    else:
33        print('Invalid PMF, sum = ', s)
34
35 #p= q
36 check(5, 0.5)
37 check(10, 0.5)
38
39 x, y = list_p(5, 0.5)
40 plt.bar(x, y)
41 plt.show()
42
43 x, y = list_p(10, 0.5)
44 plt.bar(x, y)
45 plt.show()
46
47 x, y = dis(10, 0.5)
48 plt.bar(x, y)
49 plt.show()
50
51 x, y = dis(10, 0.5)
52 plt.bar(x, y)

```

```

53 plt.show()
54
55 #p != q
56
57 check(5, 0.7)
58 check(10, 0.7)
59
60 x, y = list_p(5, 0.7)
61 plt.bar(x, y)
62 plt.show()
63
64 x, y = list_p(10, 0.7)
65 plt.bar(x, y)
66 plt.show()
67
68 x, y = dis(10, 0.7)
69 plt.bar(x, y)
70 plt.show()
71
72 x, y = dis(10, 0.7)
73 plt.bar(x, y)
74 plt.show()

```

Question 3 A discrete random variable X follows a probability mass function (PMF), with parameter $\lambda > 0$, given by:

$$f[k; \lambda] = Pr[X = k] = \frac{\lambda^k e^{-\lambda}}{k!} \quad (6.3)$$

where k is the number of occurrences $k \in \{0, 1, 2, \dots\}$, e is Eulers number, and $!$ is the factorial function.

1. Obtain one plot of $Pr[X = k]$ vs. k with the following values of $\lambda \in \{1, 4, 10\}$.
2. Obtain one plot of $Pr[X \leq k]$ vs. k with the following values of $\lambda \in \{1, 4, 10\}$.
3. Calculate the mean and the variance of the random variable X (experimentally, i.e., running the experiments multiple times and not mathematically).
4. Reason and infer the name of the PMF in (6.3).

```

1 import math
2 import matplotlib.pyplot as plt

```

```

3 import numpy as np
4 import statistics
5
6 def f(l, k):
7     p = pow(l, k) * pow(math.e, -l)/ math.factorial(k)
8     return p
9
10 N = 20
11 k = [i for i in range(N)]
12 l = [1, 4, 10]
13
14 pl1 = []
15 pl2 = []
16 for i in l:
17     p1 = []
18     p2 = []
19     s = 0
20     for j in k:
21         p = f(i, j)
22         s+=p
23         p1.append(p)
24         p2.append(s)
25     pl1.append(p1)
26     pl2.append(p2)
27
28 #plot 1
29 ind = np.arange(N)
30 width = 0.25
31 bar1 = plt.bar(ind, pl1[0], width, color = 'r')
32 bar2 = plt.bar(ind+width, pl1[1], width, color = 'g')
33 bar3 = plt.bar(ind+width*2, pl1[2], width, color = 'b')
34
35 plt.xlabel("k")
36 plt.ylabel('p')
37 plt.title("Pr [X=k]")
38
39 plt.xticks(ind+width,[i for i in range(N)])
40 plt.legend( (bar1, bar2, bar3), ('lambda = 1', 'lambda = 4', 'lambda = 10') )
41 plt.show()
42
43 #plot2
44 bar1 = plt.bar(ind, pl2[0], width, color = 'r')
45 bar2 = plt.bar(ind+width, pl2[1], width, color = 'g')
46 bar3 = plt.bar(ind+width*2, pl2[2], width, color = 'b')
47
48 plt.xlabel("k")

```

```

49 plt.ylabel('p')
50 plt.title("Pr[X<=k]")
51
52 plt.xticks(ind+width,[i for i in range(N)])
53 plt.legend( (bar1, bar2, bar3), ('lambda = 1', 'lambda = 4', 'lambda = 10') )
54 plt.show()
55
56 print('For lambda = 1')
57 m = 0
58 sd = 0
59 for i in range(N):
60     m+=i * pl1[0][i]
61     sd+=i*i*pl1[0][i]
62 v = sd - m*m
63 print('mean = ', statistics.mean(pl1[0]))
64 print('statistical var = ', statistics.variance(pl1[0]))
65 print('expectation = ', m)
66 print('var = ', v)
67
68 print('For lambda = 4')
69 m = 0
70 sd = 0
71 for i in range(N):
72     m+=i * pl1[1][i]
73     sd+=i*i*pl1[1][i]
74 v = sd - m*m
75 print('mean = ', statistics.mean(pl1[1]))
76 print('statistical var = ', statistics.variance(pl1[1]))
77 print('expectation = ', m)
78 print('var = ', v)
79
80 print('For lambda = 10')
81 m = 0
82 sd = 0
83 for i in range(N):
84     m+=i * pl1[2][i]
85     sd+=i*i*pl1[2][i]
86 v = sd - m*m
87 print('mean = ', statistics.mean(pl1[2]))
88 print('statistical var = ', statistics.variance(pl1[2]))
89 print('expectation = ', m)
90 print('var = ', v)
91
92 print('Poisson Distribution')

```

Question 4 A discrete random variable X , probability of getting exactly k successes in n independent trials is given by the following probability mass function, with parameter $n \in \mathbb{N}$, $p \in [0, 1]$:

$$f[k; n, p] = Pr[X = k] = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \quad (6.4)$$

for $k \in \{0, 1, 2, \dots, n\}$, and $!$ is the factorial function.

1. Obtain one plot of $Pr[X = k]$ vs. k with the following value pairs of $(p, n) \in \{(0.5, 20), (0.7, 20), (0.5, 40)\}$.
2. Obtain one plot of $Pr[X \leq k]$ vs. k with the following value pairs of $(p, n) \in \{(0.5, 20), (0.7, 20), (0.5, 40)\}$.
3. Calculate the mean and the variance of the random variable X (experimentally, i.e., running the experiments multiple times and not mathematically).
4. Reason and infer the name of the PMF in (6.4).

```

1 import math
2 import matplotlib.pyplot as plt
3 import numpy as np
4 import statistics
5
6 def f(n, k, p):
7     return math.comb(n, k) * math.pow(p, k) * math.pow(1-p, n-k)
8
9 P = [0.5, 0.7, 0.5]
10 N = [20, 20, 40]
11
12 p1 = []
13 p2 = []
14 for i in range(len(N)):
15     p1 = []
16     p2 = []
17     s = 0
18     for k in range(N[i] + 1):
19         p = f(N[i], k, P[i])
20         p1.append(p)
21         s+=p
22         p2.append(s)
23     p1.append(p1)

```

```

24     pl2.append(p2)
25
26 #plot 1
27 ind = np.arange(max(N) + 1)
28 width = 0.25
29 pd = pl1[0]
30 for i in range(20):
31     pd.append(0)
32 bar1 = plt.bar(ind, pd, width, color = 'r')
33 pd = pl1[1]
34 for i in range(20):
35     pd.append(0)
36 bar2 = plt.bar(ind+width,pd, width, color = 'g')
37 pd = pl1[2]
38 bar3 = plt.bar(ind+width*2, pd, width, color = 'b')
39
40 plt.xlabel("k")
41 plt.ylabel('p')
42 plt.title("Pr[X=k]")
43
44 plt.xticks(ind+width,[i for i in range(max(N) + 1)])
45 plt.legend( (bar1, bar2, bar3), ('(p,n)=(0.5,20)', '(p,n)=(0.7,20)', '(p,n)=(0.5,40)')
46 plt.show()
47
48 #plot2
49 pd = pl2[0]
50 for i in range(20):
51     pd.append(0)
52 bar1 = plt.bar(ind, pd, width, color = 'r')
53 pd = pl2[1]
54 for i in range(20):
55     pd.append(0)
56 bar2 = plt.bar(ind+width,pd, width, color = 'g')
57 pd = pl2[2]
58 bar3 = plt.bar(ind+width*2, pd, width, color = 'b')
59
60 plt.xlabel("k")
61 plt.ylabel('p')
62 plt.title("Pr[X<=k]")
63
64 plt.xticks(ind+width,[i for i in range(max(N) + 1)])
65 plt.legend( (bar1, bar2, bar3), ('(p,n)=(0.5,20)', '(p,n)=(0.7,20)', '(p,n)=(0.5,40)')
66 plt.show()
67

```

```

68 print('For (p,n)=(0.5,20)')
69 m = 0
70 sd = 0
71 for i in range(N[0] + 1):
72     m+=i * pl1[0][i]
73     sd+=i*i*pl1[0][i]
74 v = sd - m*m
75 print('mean = ', statistics.mean(pl1[0]))
76 print('statistical var = ', statistics.variance(pl1[0]))
77 print('expectation = ', m)
78 print('var = ', v)
79
80 print('For (p,n)=(0.7,20)')
81 m = 0
82 sd = 0
83 for i in range(N[1] + 1):
84     m+=i * pl1[1][i]
85     sd+=i*i*pl1[1][i]
86 v = sd - m*m
87 print('mean = ', statistics.mean(pl1[1]))
88 print('statistical var = ', statistics.variance(pl1[1]))
89 print('expectation = ', m)
90 print('var = ', v)
91
92 print('For (p,n)=(0.5,40)')
93 m = 0
94 sd = 0
95 for i in range(N[2] + 1):
96     m+=i * pl1[2][i]
97     sd+=i*i*pl1[2][i]
98 v = sd - m*m
99 print('mean = ', statistics.mean(pl1[2]))
100 print('statistical var = ', statistics.variance(pl1[2]))
101 print('expectation = ', m)
102 print('var = ', v)
103
104 print('Binomial Distribution')

```

6.9 Mock-Mid term exam

1. On a TV game show, hosted by Monty Hall, a contestant chooses one of three closed doors, two of which have a goat behind them and one of which has a car. Monty, who knows where the car is, then opens one of the two remaining doors. The door he opens always has a goat behind it (he never reveals the car!). If he has a choice, then he picks a door at random with equal

probabilities. Monty then offers the contestant the option of switching to the other unopened door. If the contestant's goal is to get the car, should he switch doors?

2. Draw a line graph of the probability mass function. Also construct and plot the cumulative distribution function for
 - (a) $\text{Geom}(1/4)$
 - (b) $\text{Bin}(2, 1/2)$
3. Roll two fair 6-sided dice. Let $T = 2X + Y$ be a new random variable, where X and Y are the individual rolls.
 - (a) What is the PMF of T ?
 - (b) What is the probability that T is in the interval $[3, 5]$?
4. The input messages to a channel are chosen from the set a, b such that probability $P(X = a) = .35$ and $P(X = b) = .65$. Output of the channel is a stream of messages from the set A, B with probability $P(Y = A)$ and $P(Y = B)$. In the channel, the input message a is altered to b with probability r and the input message b is altered to a with probability s . If $r = .45$, $s = .35$, what is the PMF of Y ?
5. A random variable Y can take the value 1, 2, or any other positive integer.
 - (a) could we have $P(Y = i) = 2c/i$ for some value of c ?
 - (b) could we have $P(Y = i) = 3c/(i^2)$ for some value of c ?
6. A n -component automatic door system functions properly if at least $2/5$ the components function properly. The independent probability of them functioning properly is q . What could be the value of q such that 5 component automatic door functions better than 10 component automatic doors?
7. A fair coin is tossed three times, and the random variable X is the number of heads in the first two tosses and the random variable Y is the number of heads in the last two tosses.
 - (a) What is the joint probability mass function of X and Y ?
 - (b) What are the marginal probability mass functions of X and Y ?
 - (c) Are the random variables X and Y independent?
8. A web server model follows Poisson distribution for probability such that k is the no. of jobs and λ as always the system parameter. Suppose 0.1 requests arrive per second. What is the probability that an interval of 10 seconds elapsed without any requests arriving?

6.9.1 Mock-Mid Term Exam-Solutions

1. Let X be the event that the person wins.

There are 2 strategies: (i) switch and (ii) no switch.

Let C_i be the event that car is behind the door i

- (a) In strategy 1:

$$\begin{aligned}
 P(X) &= \sum_{i=1}^3 P(X|C_i) \\
 &= (0 * 1/3) + (1 * 1/3) + (1 * 1/3) \\
 &= \frac{2}{3}
 \end{aligned} \tag{6.5}$$

- (b) In strategy 2:

$$\begin{aligned}
 P(X) &= \sum_{i=1}^3 P(X|C_i) \\
 &= (1 * 1/3) + (0 * 1/3) + (0 * 1/3) \\
 &= \frac{1}{3}
 \end{aligned} \tag{6.6}$$

Hence, strategy (1) is better i.e. switch.

2. $T = 2X+Y$

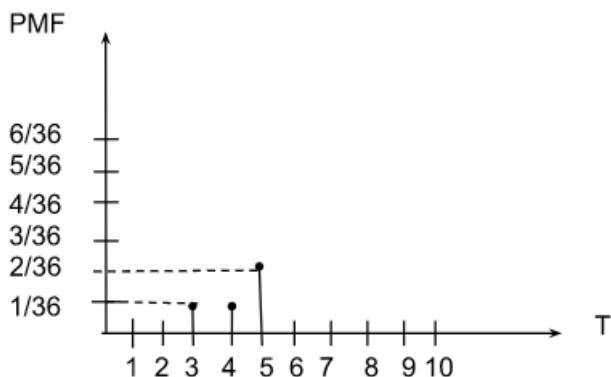
- (a) Support of $T = 3, 4, \dots, 18$

$$P_T(3) = P(T = 3) = P(X = 1, Y = 1) = 1/36$$

$$P_T(4) = P(T = 4) = P(X = 1, Y = 2) = 1/36$$

$$P_T(5) = P(T = 5) = P(X = 2, Y = 1) = 2/36$$

Similarly, find for others.



(b)

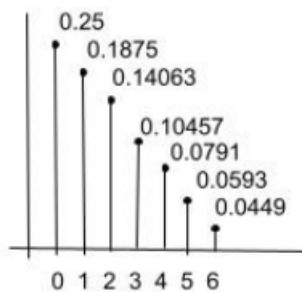
$$P(T \in [3, 5]) = P(T = 3 \text{ or } T = 4 \text{ or } T = 5) = 1/36 + 1/36 + 2/36 = 1/9$$

3. (a) As we know that if $X \sim \text{Geom}(p)$ then PMF of X is given by

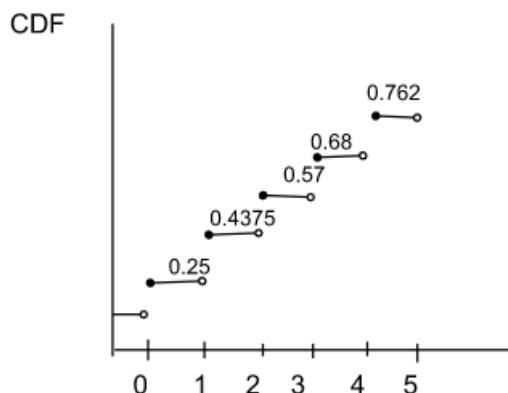
$$P(X=k) = (1-p)^k p \text{ for } k = 0, 1, 2, \dots$$

Hence, applying the formula for k ranging 1 to 6 we get, PMF plot as follows

$\text{Geom}(1/4)$



Using PMF plot the CDF plot obtained is as follows



- (b) As we know that $X \sim \text{Bin}(n, p)$ then PMF(X) is given by

$$P_X(k) = P(X = k) = \binom{n}{k} \times p^k (1-p)^{n-k}$$

Use the theorem to draw the PMF for

$\text{Bin}(2, 1/2)$

4. $P(Y=A|X=a) = 1-r$

$$P(Y=B|X=a) = r$$

$$P(Y=B|X=b) = 1-s$$

$$P(Y=A|X=b) = s$$

$$\begin{aligned}
P_Y(A) &= P_{X,Y}(a, A) + P_{X,Y}(b, A)) \\
&= P(Y = A|X = a).P(X = a) + P(Y = A|X = b).P(X = b) \quad (6.7) \\
&= (1 - r) * (0.35) + s * 0.65
\end{aligned}$$

$$\begin{aligned}
P_Y(B) &= P_{X,Y}(a, B) + P_{X,Y}(b, B)) \\
&= P(Y = B|X = a).P(X = a) + P(Y = B|X = b).P(X = b) \quad (6.8) \\
&= (1 - s) * (0.65) + r * 0.35
\end{aligned}$$

Substitute values of $r=0.45$ and $s=0.65$

5. A functioning component is a binary event, hence we can apply binomial distribution

Let the probability of a 5-component automatic door system be $P_5(q)$ and that of a 3-component automatic door system be $P_3(q)$

$$P_5(q) = P(\text{at least two component function}) \quad (6.9)$$

$$= P(2 - \text{component}) + P(3 - \text{component}) + P(4 - \text{component}) + P(5 - \text{component}) \quad (6.10)$$

$$= (\binom{5}{2} * q^2 * (1 - q)^{(5-2)}) + (\binom{5}{3} * q^3 * (1 - q)^{(5-3)}) + \quad (6.11)$$

$$(\binom{5}{4} * q^4 * (1 - q)^{(5-4)}) + (\binom{5}{5} * q^5 * (1 - q)^{(5-5)}) \quad (6.12)$$

$$P_3(q) = P(\text{at least 1.2 component function}) \quad (6.13)$$

$$= P(2 - \text{component}) + P(3 - \text{component}) \quad (6.14)$$

$$= (\binom{3}{2} * q^2 * (1 - q)^{(3-2)}) + (\binom{3}{3} * q^3 * (1 - q)^{(3-3)}) \quad (6.15)$$

We need to find range of q such that $P_5(q) > P_3(q)$, solve for q .

6.

OUTCOME	No. of heads in 1st two tosses (X)			No. of heads in last two tosses (X)		
	2	1	0	2	1	0
HHH	yes			yes		
HHT	yes				yes	
HTT		yes				yes
TTT			yes			yes
TTH			yes		yes	
THH		yes			yes	
THT		yes		yes		
HTH		yes			yes	

(a) The joint probability mass function is

$X Y$	0	1	2
0	1/8	1/8	0
1	1/8	2/8	0
2	0	1/8	0

Table 6.1: Table for Question-7a

(b) Using the marginal distribution formula,

PMF for X

$$P(X = 0) = 1/8 + 1/8 + 0 = 1/4$$

$$P(X = 1) = 1/8 + 1/8 + 1/8 = 1/2$$

$$P(X = 2) = 0 + 1/8 + 1/8 = 1/4$$

PMF for Y

$$P(Y = 0) = 1/8 + 1/8 + 0 = 1/4$$

$$P(Y = 1) = 1/8 + 1/4 + 1/8 = 1/2$$

$$P(Y = 2) = 0 + 1/8 + 1/8 = 1/4$$

(c) $P(X = 0) \cdot P(Y = 0) = 1/16$

$$P(X = 0, Y = 0) = 1/8$$

Since above 2 equations are not equal hence, X and Y are not independent random variables.

7. The arrival is modeled using a Poisson distribution.

The parameter is the rate of arrival.

$$P(k \text{ jobs in } t \text{ secs}) = (\lambda t)^k * e^{-\lambda t} / k!$$

$$P(0 \text{ jobs in } 10 \text{ secs}) = e^{-0.1*10} = 0.3678$$

6.10 Bivariate Distributions

Till now, we have discussed the probabilistic properties of a single random variable. Using this knowledge, we can generalize the concept of a random variable to the joint distributions of two (even to some finite number n) random variables. The joint distributions describes the relationship between the two (multiple) concerned random variables.

Some of its practical applications are:

1. We might measure the amount of precipitate P and volume V of gas released from a controlled chemical experiment, giving rise to a two-dimensional sample space consisting of the outcomes (p, v) .
2. We might be interested in the hardness H and tensile strength T of cold-drawn copper, resulting in the outcomes (h, t) .

6.10.1 Joint Probability Distributions

If the random variables, X and Y , are discrete, then the *joint probability mass function* consists of probability values

$$P(X = x_i, Y = y_j) = p_{ij} \geq 0$$

satisfying $\sum_i \sum_j p_{ij} = 1$.

If the random variables are continuous, X and Y , then the *joint probability density function* is a function $f(x, y) \geq 0$ satisfying

$$\int \int_{\text{state space}} f(x, y) \, dx \, dy = 1$$

The probability that $a \leq X \leq b$ and $c \leq Y \leq d$ is obtained from the *joint probability density function* as

$$\int_{x=a}^b \int_{y=c}^d f(x, y) \, dx \, dy$$

The *joint cumulative distribution function* is defined as

$$F(x, y) = P(X \leq x, Y \leq y),$$

which is

$$F(x, y) = \sum_i \sum_j p_{ij},$$

for the discrete random variables, and

$$F(x, y) = \int_{w=-\infty}^x \int_{z=-\infty}^y f_{W,Z}(w, z) dw dz,$$

for the continuous random variables.

Let us now, summarize the above theory as follows:

The joint probability distribution of two random variables X and Y is specified by a set of probability values $P(X = x_i, Y = y_j) = p_{ij}$ for discrete random variables, or a joint probability density function $f(x, y)$ for continuous random variables. In either case, the joint *cumulative distribution function* is defined to be

$$F(x, y) = P(X \leq x, Y \leq y)$$

Now, we would first summarize all the properties that a joint *cumulative distribution function* satisfies and then will go through some examples to have a better understanding of the *joint probability distributions*.

1. Marginal *cdf* of X , $F_X(x) = F_{XY}(x, \infty)$, for any x ;
2. Marginal *cdf* of Y , $F_Y(y) = F_{XY}(\infty, y)$, for any y ;
3. $F_{XY}(\infty, \infty) = 1$;
4. $F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0$;
5. $P(x_1 < X < x_2, y_1 < Y < y_2) = F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1)$;
6. If X and Y are independent, then $F_{XY}(x, y) = F_X(x)F_Y(y)$.

Example 1: A company that services air conditioner units in residences and office blocks is interested in how to schedule its technicians in the most efficient manner. Specifically, the company is interested in how long a technician takes on a visit to a particular location, and the company recognizes that this mainly depends on the number of air conditioner units at the location that need to be serviced.

If the random variable X , taking the values 1, 2, 3, and 4, is the service time in hours taken at a particular location, and the random variable Y , taking the values 1, 2, and 3, is the number of air conditioner units at the location, then these two random variables can be thought of as jointly distributed.

Assume that their joint probability mass function p_{ij} is given in the table 6.2. The table 6.2 indicates, for example, that there is a probability of 0.12 that $X = 1$ and $Y = 1$, so that there is

a probability of 0.12 that a particular location chosen at random has one air conditioner unit that takes a technician one hour to service. Note that this is a valid probability mass function since

$$\sum_i \sum_j p_{ij} = 0.12 + 0.08 + \dots + 0.07 = 1$$

Y/X	1	2	3	4
1	0.12	0.08	0.07	0.05
2	0.08	0.15	0.21	0.13
3	0.01	0.01	0.02	0.07

Table 6.2: Joint probability mass function

The *joint cumulative distribution function* is

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_i \sum_j p_{ij}$$

as shown in the table 6.3. Now, the probability that a location has no more than two air conditioner units that take no more than two hours to service is

$$F(2, 2) = p_{11} + p_{12} + p_{21} + p_{22} = 0.12 + 0.08 + 0.08 + 0.15 = 0.43$$

Y/X	1	2	3	4
1	0.12	0.20	0.27	0.32
2	0.20	0.43	0.71	0.89
3	0.21	0.45	0.75	1.00

Table 6.3: Joint cumulative function

Example 2: In order to determine the economic viability of mining in a certain area, a mining company obtains samples of ore from the location and measures their zinc content and their iron content. Suppose that the random variable X is the zinc content of the ore, taking values between 0.5 and 1.5, and that the random variable Y is the iron content of the ore, taking values between 20.0 and 35.0. Further assume that their joint probability density function is given as

$$f(x, y) = \frac{39}{400} - \frac{17(x-1)^2}{50} - \frac{(y-25)^2}{10,000}, \quad 0.5 \leq x \leq 1.5, \quad 20.0 \leq y \leq 35.0.$$

The probability that a randomly chosen sample of ore has a zinc content between 0.8 and 1.0 and an iron content between 25 and 30 is

$$\begin{aligned} & \int_{x=0.8}^1 \int_{y=25}^{30} f(x, y) dx dy = \int_{x=0.8}^1 \int_{y=25}^{30} \frac{39}{400} - \frac{17(x-1)^2}{50} - \frac{(y-25)^2}{10,000} dx dy \\ &= 0.092 \end{aligned}$$

6.10.2 Marginal Probability Distributions

Suppose we have knowledge of *joint probability distribution* but our interest is of only a single random variable. Then, it is appropriate to consider the probability distribution of that random variable alone. This is known as the *marginal distribution* of the random variable and can be obtained quite simply by summing or integrating the joint probability distribution over the values of the other random variable.

For example, for two discrete random variables X and Y , the probability values of the marginal distribution of X are

$$P(X = x_i) = \sum_j p_{ij}$$

and for two continuous random variables, the probability density function of the marginal distribution of X is

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

where in practice the summation and integration limits can be curtailed at the appropriate boundaries of the state space. Note that the marginal distribution of a random variable X and the marginal distribution of a random variable Y do not uniquely determine their joint distribution.

Let us summarize the above theory as follows:

The marginal distribution of a random variable X is obtained from the joint probability distribution of two random variables X and Y by summing (if they are discrete random variables) or integrating (if they are continuous random variables) over the values of the random variable Y . The marginal distribution is the individual probability distribution of the random variable X considered alone.

Now, we will go through some examples to understand these concepts.

Example 1: Consider the case of AC maintenance problem (example 1) of the previous section. The *marginal probability mass function* of X , the time taken to service the air conditioner units at a particular location, is given in the below figure and is obtained by summing the appropriate values of the joint probability mass function. For example,

$$P(X = 1) = \sum_{j=1}^3 p_{1j} = 0.12 + 0.08 + 0.01 = 0.21$$

The *marginal probability mass function* of Y , the number of air conditioner units at a particular

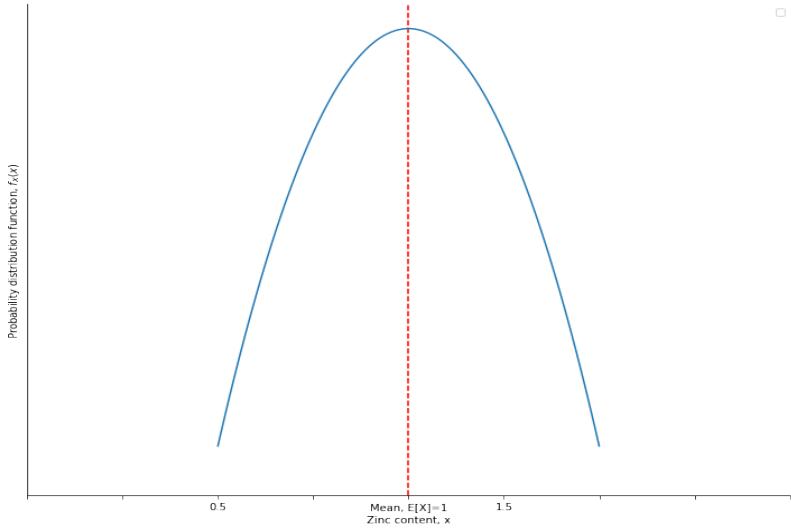


Fig. 6.29: Marginal probability density function of X

location, is given in the table 6.4. Here,

$$P(Y = 1) = \sum_{i=1}^4 p_{i1} = 0.12 + 0.08 + 0.07 + 0.05 = 0.32$$

Y/X	1	2	3	4	Marginal Y
1	0.12	0.08	0.07	0.05	0.32
2	0.08	0.15	0.21	0.13	0.57
3	0.01	0.01	0.02	0.07	0.11
Marginal X	0.21	0.24	0.30	0.25	1.0

Table 6.4: Marginal probability mass function

Example 2: Consider the mineral deposit problem (example 2) of the previous section. The *marginal probability density function* of X , the zinc content of the ore, is

$$\begin{aligned} f_X(x) &= \int_{y=20}^{35} f(x, y) dy = \int_{y=20}^{35} \left(\frac{39}{400} - \frac{17(x-1)^2}{50} - \frac{(y-25)^2}{10,000} \right) dy \\ &= \left[\frac{39y}{400} - \frac{17y(x-1)^2}{50} - \frac{(y-25)^3}{30,000} \right]_{y=20}^{35} = \frac{57}{40} - \frac{51(x-1)^2}{10} \end{aligned}$$

for $0.5 \leq x \leq 1.5$. This is shown in the figure 6.29.

Similarly, the *marginal probability density function* of Y , the iron content in the ore, is given as

$$f_Y(y) = \int_{x=0.5}^{1.5} f(x, y) dx = \int_{x=0.5}^{1.5} \left(\frac{39}{400} - \frac{17(x-1)^2}{50} - \frac{(y-25)^2}{10,000} \right) dy$$

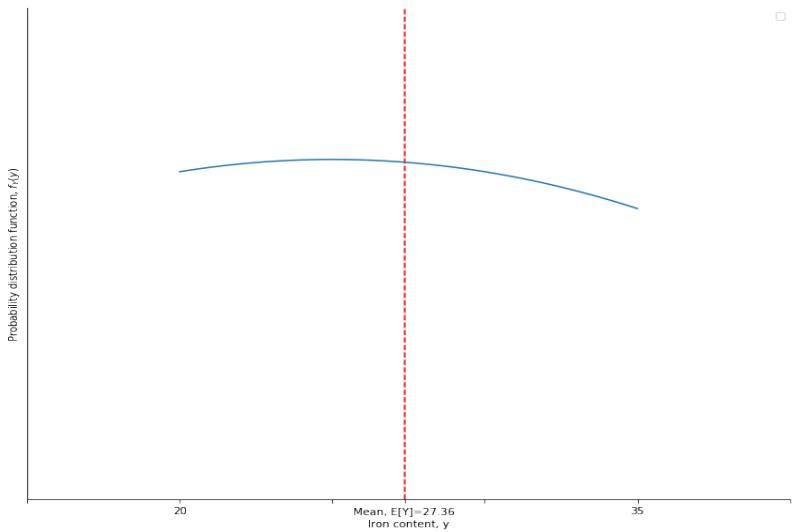


Fig. 6.30: Marginal probability density function of Y

$$= \left[\frac{39x}{400} - \frac{17(x-1)^3}{150} - \frac{x(y-25)^2}{10,000} \right]_{y=0.5}^{1.5} = \frac{83}{1200} - \frac{(y-25)^2}{10,000}$$

for $20 \leq y \leq 35$. This is shown in the figure 6.30.

6.10.3 Conditional Probability Distributions

If two random variables X and Y are jointly distributed, then we might be interested in the distribution of one random variable conditional on the other random variable having taken a particular value.

If two continuous random variables X and Y are jointly distributed, then the conditional distribution of random variable X conditional on the event $Y = y$ has a probability density function

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

where $f(y)$ is the *marginal distribution* of the random variable Y .

Now, we would summarize the above concepts as:

The *conditional distribution* of a random variable X conditional on a random variable Y taking a particular value describes the probabilistic properties of the random variable X under the knowledge

provided by the value of Y . It consists of the probability values

$$p_{i|Y=y_j} = P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{\sum_i p_{ij}} p + j$$

for discrete random variables or the probability density function

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

for continuous random variables, where $f_Y(y)$ is the marginal distribution of the random variable Y .

It is important to recognize the difference between a marginal distribution and a conditional distribution. The marginal distribution for X is the appropriate distribution for the random variable X when nothing is known about the random variable Y . In contrast, the conditional distribution for X conditional on a particular value y of Y is the appropriate distribution for the random variable X when the random variable Y is known to take the value y .

Now, we would go through some examples to understand it.

Example 1: Continuing with the AC maintenance problem, assume that a technician is visiting a location that is known to have three air conditioner units. Then this event has a probability of

$$P(Y = 3) = \sum_i p_{i3} = p_{13} + p_{23} + p_{33} + p_{43} = 0.01 + 0.01 + 0.02 + 0.07 = 0.11$$

Then, the conditional distribution of the service time X consists of the probability values

$$p_{1|Y=3} = P(X = 1 | Y = 3) = \frac{P(X = 1, Y = 3)}{P(Y = 3)} = \frac{0.01}{0.11} = 0.091$$

$$p_{2|Y=3} = P(X = 2 | Y = 3) = \frac{P(X = 2, Y = 3)}{P(Y = 3)} = \frac{0.01}{0.11} = 0.091$$

$$p_{3|Y=3} = P(X = 3 | Y = 3) = \frac{P(X = 3, Y = 3)}{P(Y = 3)} = \frac{0.02}{0.11} = 0.182$$

$$p_{4|Y=3} = P(X = 4 | Y = 3) = \frac{P(X = 4, Y = 3)}{P(Y = 3)} = \frac{0.07}{0.11} = 0.636$$

Hence, the *conditional distribution* of X is given as

$$f_{X|Y=y}(x) = \begin{cases} 0.091 & x = 1 \\ 0.091 & x = 2 \\ 0.182 & x = 3 \\ 0.636 & x = 4 \end{cases}$$

6.10.4 Independence

Two random variables X and Y are said to be independent if the knowledge of the value taken by the random variable Y does not influence the distribution of the random variable X , and vice versa. More formally, it is defined as

Two random variables X and Y are defined to be independent if their joint probability mass function or joint probability density function is the product of their two marginal distributions. If the random variables are discrete, then they are independent if

$$p_{ij} = \sum_j p_{ij} = \sum_i p_{ij}$$

for all values of x_i and y_j . If the random variables are continuous, then they are independent if

$$f(x, y) = f_X(x) f_Y(y)$$

for all values of x and y . If two random variables are independent, then the probability distribution of one of the random variables does not depend upon the value taken by the other random variable.

Example: Assume that the random variables X and Y have a *joint probability density function*, $f(x, y)$, given as

$$f_{X,Y}(x, y) = \begin{cases} 6xy^2 & 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Comment whether the random variables X and Y are independent or not.

Solution: Given, the *joint probability density function*, $f(x, y)$, of the random variables X and Y as

$$f_{X,Y}(x, y) = \begin{cases} 6xy^2 & 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Then, the *marginal distribution* of X is

$$f_X(x) = \int_{y=0}^1 f_{X,Y}(x,y) \, dx \, dy = \int_{y=0}^1 6xy^2 \, dx \, dy = 2x$$

for $0 \leq x \leq 1$.

Similarly, the *marginal distribution* of Y is

$$f_Y(y) = \int_{x=0}^1 f_{X,Y}(x,y) \, dx = \int_{x=0}^1 6xy^2 \, dx = 3y^2$$

for $0 \leq y \leq 1$.

Now, since

$$f_X(x) \cdot f_Y(y) = 2x * 3y^2 = 6xy^2 = f_{X,Y}(x,y)$$

the random variables X and Y are independent random variables.

6.11 Summary

Axioms of Probability:

1. $0 \leq P(A) \leq 1$, for any event A .
2. The probability of the whole sample space is 1, i.e., $P(S) = 1$.
3. If A and B are disjoint events i.e., $P(A \cap B) = 0$, then $P(A \cup B) = P(A) + P(B)$.

Conditional Probability:

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

Independence:

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

Bayes formula:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})}$$

Bayes' general form: Suppose that B_1, B_2, \dots, B_k form a partition of S : $B_i \cap B_j = \emptyset$ and $\cup_i B_i = S$. Then, the general form of Bayes' formula is

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_k)P(B_k)}$$

Probability distribution and mass functions

The probability mass function of a discrete random variable X is defined as

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

$$f(x) \geq 0 \text{ and,}$$

Similarly the probability distribution function of a continuous random variable X is defined as

$$P(a \leq X \leq b) = \int_a^b f(x) \, dx$$

$$f(x) \geq 0$$

Cumulative Distribution Function

The *cumulative distribution function* for a discrete random variable is given as

$$F_X(x) = P(X \leq x) = \sum_{t \leq x} P(X = t)$$

and, the *cumulative distribution function* for a continuous random variable is given as

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

It is conventional to use a capital F for a *cdf* and lower case f for *pmf* or *pdf*.

For a discrete random variable

$$P(a \leq X \leq b) = F(b) - F(a) \quad \text{and,}$$

$$P(X = x) = F(x) - F(x^-)$$

For a continuous random variable

$$P(a \leq X \leq b) = F(b) - F(a),$$

$$F(x) = \int_{-\infty}^x f(t) dt \quad \text{and,}$$

$$f(x) = \frac{dF(x)}{dx}$$

Cumulative conditional distribution function

$$F_X\left(\frac{x}{M}\right) = \frac{P(X \leq x, M)}{P(M)}$$

Joint Distributions: For a discrete random variable X , the joint distribution is $P(X = x, Y = y) = f(x, y)$, the probability that both $X = x$ and $Y = y$ and the joint cumulative discrete distribution is

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{s \leq x} \sum_{t \leq y} f(s, t)$$

The marginal probability mass function for X is

$$g(x) = F(x, \infty) = \sum_{t \in R_y} f(s, t)$$

The marginal probability mass function for Y is

$$h(y) = F(\infty, y) = \sum_{s \in R_x} f(s, t)$$

For a continuous random variable X , the joint density function is

$$P(a < X < b, c < Y < d) = \int_{x=a}^b \int_{y=c}^d f(x, y) \ dx \ dy$$

The marginal densities of X and Y are given as follows:

$$g(x) = \int_{-\infty}^{\infty} f(x, y) \ dy, \text{ for all } 0.1cm x$$

$$h(y) = \int_{-\infty}^{\infty} f(x, y) \ dx, \text{ for all } 0.1cm y$$

Conditional discrete distribution:

$$f(x|y) = \frac{f(x, y)}{h(y)}$$

Independence:

$$f(x, y) = g(x) * h(y)$$

Expectations: The mean or expected value of a discrete random variable X is given as

$$\mu_X = E[X] = \sum_x x f(x)$$

and for a continuous random variable

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f(x) \ dx$$

It's properties are:

1. $E[aX + b] = aE[X] + b$
2. $E[X + Y] = E[X] + E[Y]$
3. If X and Y are independent, then $E[XY] = E[X]E[Y]$

The variance of a discrete random variable is

$$\sigma^2 = Var(X) = E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - \mu^2 = \sum_x (x - \mu)^2 f(x)$$

For continuous random variable

$$\sigma^2 = Var(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \ dx$$

The standard deviation is the positive root of the variance

$$\sigma = \sqrt{Var(X)}$$

$$Var(aX + b) = a^2 Var(X)$$

The covariance between X and Y in discrete case is

$$\sigma_{XY} = Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = \sum_x \sum_y (x - \mu_x)(y - \mu_y)f(x, y)$$

and in continuous case is

$$\sigma_{XY} = Cov(X, Y) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y) dx dy$$

The covariance between two random variables X and Y with means μ_x and μ_y respectively is

$$\sigma_{XY} = Cov(X, Y) = E[XY] - \mu_x \mu_y$$

$$Cov(X, X) = Var(X)$$

If X and Y are independent random variables, then $Cov(X, Y) = 0$.

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov(X, Y)$$

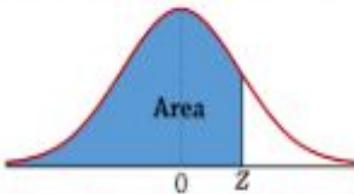
If X and Y are independent random variables, then $Var(X + Y) = Var(X) + Var(Y)$.

The correlation coefficient of X and Y is

$$\rho_{XY} = \frac{Cov(X, Y)}{\rho_x \rho_y}$$

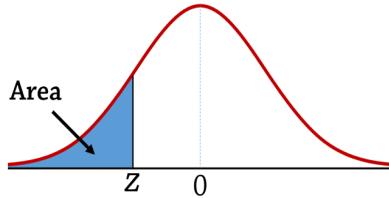
It holds that $-1 \leq \rho_{xy} \leq 1$. If X and Y are independent, $\rho_{xy} = 0$.

Table of Standard Normal Probabilities for Positive z-coordinates



<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Fig. 6.31: Z distribution table-1



<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Fig. 6.32: *Z* distribution table-2

6.12 Practice Problem Set-1

1. A floor in charge supervises the operation of three machines, namely X , Y , and Z . At any given time, each of them can be classified as either working, denoted by 1 or being idle, denoted by 0. The notation $(0, 1, 0)$ is used to represent the situation where machine Y is working but machines X and Z are both idle. Then give the sample space for the status of the three machines at a particular point in time.
2. Two fair dice are thrown, one red and one blue. What is the probability that the red die has a score that is strictly greater than the score of the blue die?
3. A die is loaded in such a way that an even number is twice as likely to occur as an odd number. If E is the event that a number less than 4 occurs on a single toss of the die, find the probability of the event E , $P(E)$.
4. If the probabilities are, respectively, 0.09, 0.15, 0.21, and 0.23 that a person purchasing a new automobile will choose the colour green, white, red, or blue, what is the probability that a given buyer will purchase a new automobile that comes in one of those colours?
5. If the probabilities that an automobile mechanic will service 3, 4, 5, 6, 7, or 8 or more cars on any given workday are, respectively, 0.12, 0.19, 0.28, 0.24, 0.10, and 0.07, what is the probability that he will service at least 5 cars on his next day at work?
6. When rolling a fair die, an event A is defined as the event of getting an even number and event B is defined as the event of getting a high score and they are given as $A = \{2, 4, 6\}$ and, $B = \{4, 5, 6\}$. Then find:
 - (a) Probability of their intersection, $P(A \cap B)$.
 - (b) Probability of their union, $P(A \cup B)$.

6.13 Practice Problem Set-1 Solutions

1. Given, there are three machines, namely X , Y , and Z . Each of them can be either in working or idle state. The notation $(0,1,0)$ is used to represent the situation where machine Y is working but machines X and Z are both idle.

The sample space of an experiment contains all the possible outcomes of it. Since, there are only machines and each of them can be in either two states represented by 1 and 0, the sample space would be

$$S = \{(0, 0, 0)(0, 0, 1)(0, 1, 0)(0, 1, 1)(1, 0, 0)(1, 0, 1)(1, 1, 0)(1, 1, 1)\}.$$

2. Without loss of generality, we can consider that the first die is a red die and second one is a blue die.

Since there are two fair dice, the total possible outcomes of it are 36, denoted by $n(S) = 6 * 6 = 36$.

Let E be the event of getting a greater number on the red die and $\{(a, b)\}$ represent an event, where 'a' represents the number that appeared on the red die and 'b' represents the number that appeared on the blue die. Then, the possible outcomes of E are given as

$$E = \{(2, 1), (3, 1), (3, 2), (4, 1), (4, 2), (4, 3), (5, 1), (5, 2), (5, 3), (5, 4), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5)\}$$

Total number of possible outcomes of E , $n(E) = 15$.

$$\text{Then, probability of the event } E, P(E) = \frac{n(E)}{n(S)} = \frac{15}{36} = \frac{5}{12}.$$

Therefore, the probability that the red die has a score that is strictly greater than the score of the blue die is $\frac{5}{12}$.

3. Given, that an even number is twice as likely to occur as an odd number and E is the event that a number less than 4 occurring on a single toss of the die.

Let us consider that if the probability of an odd number is x , then the probability of an even number would be $2x$. Thus, we get

$$P(1) = P(3) = P(5) = x \quad \text{and},$$

$$P(2) = P(4) = P(6) = 2x$$

But we know that the total probability would be 1. So we get,

$$P(1) + P(3) + P(5) + P(2) + P(4) + P(6) = 1$$

$$\Rightarrow x + x + x + 2x + 2x + 2x = 1$$

$$\Rightarrow 9x = 1$$

$$\Rightarrow x = \frac{1}{9}.$$

$$\Rightarrow P(1) = P(3) = P(5) = \frac{1}{9} \quad \text{and } P(2) = P(4) = P(6) = \frac{2}{9}.$$

The possible outcomes of the event $E = \{1, 2, 3\}$. Then,

$$P(E) = P(1) + P(2) + P(3) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9}.$$

Thus, the probability of the event E is $\frac{4}{9}$.

4. Given, the probabilities that a person purchasing a new automobile will choose the colour green, white, red, or blue are 0.09, 0.15, 0.21, and 0.23, respectively.

Let E be an event that a given buyer will purchase a new automobile that comes in one of these colours. Then,

$$\begin{aligned} P(E) &= P(\text{green}) + P(\text{white}) + P(\text{red}) + P(\text{blue}) \\ &= 0.09 + 0.15 + 0.21 + 0.23 = 0.68. \end{aligned}$$

Therefore, the probability that a given buyer will purchase a new automobile that comes in one of those colours is 0.68.

5. Given, the probabilities that an automobile mechanic will service 3, 4, 5, 6, 7, or 8 or more cars on any given workday are, respectively, 0.12, 0.19, 0.28, 0.24, 0.10, and 0.07.

Let E be an event that he will service at least 5 cars on his next day at work. Then,

$$\begin{aligned} P(E) &= P(5) + P(6) + P(7) + P(8 \text{ or more}) \\ &= 0.28 + 0.24 + 0.10 + 0.07 = 0.69. \end{aligned}$$

Therefore, the probability that he will service at least 5 cars on his next day at work is 0.69.

6. Given, an event A , defined as the event of getting an even number, when a fair die is rolled and event B defined as the event of getting a high score and they are given as $A = \{2, 4, 6\}$ and $B = \{4, 5, 6\}$.

Then, the events $A \cap B$ and $A \cup B$ are given as $\{4, 6\}$ and $\{2, 4, 5, 6\}$, respectively. Then,

$$\begin{aligned} (\text{a}) \quad P(A \cap B) &= \frac{n(A \cap B)}{n(S)} = \frac{2}{6} = \frac{1}{3} \\ (\text{b}) \quad P(A \cup B) &= P(A) + P(B) - P(A \cap B) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{2}{3}. \end{aligned}$$

6.14 Practice Problem Set-2

1. Define a discrete random variable and suppose that the probability of having x accidents is given by the following probability mass function (pmf)

$$P(X = x) = \frac{1}{2^{x+1}}$$

Then, comment on the validity of the pmf.

2. If there is a probability of 0.261 that a milk container is underweight and these milk containers are shipped to retail outlets in boxes of 20 containers. Then,
 - (a) What is the distribution of the number of underweight containers in a box, if the underweight containers are independent of each other?
 - (b) What is the probability that exactly 7 of them would be underweight?
3. Telephone ticket sales for an event are handled by a bank of telephone salespersons who start accepting calls at a specified time. In order to get through to an operator, a caller has to be lucky enough to place a call at just the time when a salesperson has become free from a previous client. Suppose that the chance of this is 0.1. Then,
 - (a) What is the distribution of the number of calls that a person needs to make until a salesperson is reached?
 - (b) What is the probability that 15 or more calls are needed?
4. Suppose that a plane's engines start successfully at a given attempt with a probability of 0.75. Any time that the mechanics are unsuccessful in starting the engines, they must wait five minutes before trying again. Then, what is the probability that the plane is launched within 10 minutes of the first attempt?
5. Suppose that the number of errors in a piece of software has a Poisson distribution with parameter $\lambda = 3$. Then, what is the probability that there are three or more errors in a piece of software?
6. A quality inspector at a glass manufacturing company inspects sheets of glass to check for any slight imperfections. Suppose that the number of these flaws in a glass sheet has a Poisson distribution with parameter $\lambda = 0.5$. Then, find the probability that there are
 - (a) no flaws in a sheet
 - (b) at least 2 flaws

6.15 Practice Problem Set-2 Solutions

1. A random variable is a function that assigns a numerical value to the outcomes of an experiment. A discrete random variable is a random variable that takes on a finite set of values.

The probability of having x accidents is given by the following probability mass function (pmf)

$$P(X = x) = \frac{1}{2^{x+1}}$$

We know that the total sum of probabilities must sum to 1. Hence, for the above pmf to be valid,

$$\sum_{x=0}^{\infty} P(X = x) = 1$$

$$\Rightarrow \sum_{x=0}^{\infty} \frac{1}{2^{x+1}} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1$$

Hence, the given pmf is valid.

2. Given, there is a probability of 0.261 that a milk container is underweight and 0.739 for its complementary event.

In each box, there are a total of 20 containers.

- (a) Given that the underweight containers are independent of each other and there are only two possible outcomes for each trial i.e., being underweight and not being underweight, the distribution of the number of underweight containers in a box would be a binomial distribution with parameters $n = 20$ and $p = 0.261$.
- (b) If we consider that X is a random variable that takes on the number of underweight containers, then the probability that exactly 7 of them would be underweight is given by

$$P(X = 7) = \binom{20}{7} (0.261)^7 (0.739)^{13} = 0.125.$$

3. Given, the probability that the caller is lucky enough to place a call at just the time when a salesperson has become free from a previous client is 0.1.

- (a) Here, placing a call represents a Bernoulli trial with a success probability of $p = 0.1$ and the quantity of interest is the number of calls made until the first success. So, the distribution of the required number of calls would be a geometric distribution.
- (b) The probability that 15 or more calls are needed is

$$P(X \geq 15) = 1 - P(X \leq 14) = 1 - (1 - 0.9^{14}) = 0.9^{14} = 0.229$$

4. Here, the quantity of interest is at the number of trials until the first success. Hence, the geometric distribution is the appropriate distribution for the number of trials required to start a plane's engine.

The probability that the plane is launched within 10 minutes of the first attempt to start the engine is the probability that no more than three attempts are required, is

$$P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3)$$

$$= 0.75 + 0.25 * 0.75 + 0.25^2 * 0.75 = 0.9845$$

5. Given that the number of errors in a piece of software has a Poisson distribution with parameter $\lambda = 3$. The probability that there are three or more errors in a piece of software is

$$P(X \geq 3) = 1 - P(X = 0) - P(X = 1) - P(X = 2)$$

$$= 1 - \frac{e^{-3} * 3^0}{0!} - \frac{e^{-3} * 3^1}{1!} - \frac{e^{-3} * 3^2}{2!} = 0.577$$

6. Given, the number of flaws in a glass sheet has a Poisson distribution with parameter $\lambda = 0.5$.

- (a) The probability that there are no flaws is

$$P(X = 0) = \frac{e^{-0.5} * 0.5^0}{0!} = e^{-0.5} = 0.607$$

- (b) The probability that there are at least 2 errors is

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1)$$

$$= 1 - \frac{e^{-0.5} * 0.5^0}{0!} - \frac{e^{-0.5} * 0.5^1}{1!} = 1 - 0.910 = 0.090.$$

6.16 Practice Problem Set-3

- Consider an exam consisting of 20 multiple choice questions. Each of them have four possible options and only one of them is correct. Suppose you know the answers to only 10 questions and have no idea about the other 10 questions and you choose to answer randomly. Let X be a random variable taking the total number of correct answers. Then find the *probability mass function (pmf)* of the random variable X and also the probability $P(X > 15)$.
- Let X be a discrete random variable with the following *probability mass function (pmf)*

$$f_X(x) = \begin{cases} 0.1, & x = 0.2 \\ 0.2, & x = 0.4 \\ 0.2, & x = 0.5 \\ 0.3, & x = 0.8 \\ 0.2, & x = 1.0 \\ 0, & \text{otherwise} \end{cases}$$

- (a) Find the range of the random variable X .
(b) Find $P(0.25 < X < 0.75)$.
(c) Find $P(X = 0.2|X < 0.6)$.
- Consider an experiment of rolling two dice, one red and one blue. Let X be a random variable taking twice the value of the blue die if the red die has an even value and to be the value of the red die minus the value of the blue die if the red die has an odd value. Construct its *probability mass function* of the random variable X .
- Suppose that the random variable X is the time taken by a garage to service a car. These times are distributed between 0 and 10 hours with a *cumulative distribution function (cdf)*

$$F(x) = A + B * \ln(3x + 2), \quad 0 \leq x \leq 10.$$

Now, find the

- (a) values of A and B such that the given *cdf* is valid.
(b) probability that a repair job takes longer than two hours.
- Suppose that two persons A and B can transmit messages between them and another person C has the ability to eavesdrop them, with a probability of 0.8, independently of the other messages.

On a certain day, A and B sent each other a total of 8 messages and let X be the number of those 8 messages that were eavesdropped by C . Then,

- (a) Find the *pmf* of X and $E[X]$.
 - (b) What is the most probable number of messages eavesdropped by C ?
 - (c) Find the probability that exactly two messages were eavesdropped, given that at least one message was eavesdropped.
6. Consider a game in which a fair coin is tossed ten times and a player would win if at least four coin tosses show tails. Let X be the random variable denoting the number of tails observed in the game, and let W denote the event that the player wins.
- (a) Find $P(W)$
 - (b) Find $P(X \leq 5|W)$.
7. Suppose vehicles arrive at a road signal at an average rate of 360 per hour and the cycle of the traffic lights is set at 40 seconds. In what percentage of cycles will the number of vehicles arriving will be
- (a) exactly 5
 - (b) less than 5
8. A committee has been formed to decide whether to base a recovery vehicle on a stretch of road to clear accidents as quickly as possible. The concerned road carries over 5000 vehicles during peak rush hour period. On average, the number of incidents during the peak hour is 5. The committee wont base a vehicle on the road if the percentage of having more than 5 incidents is less than 30 %. Comment whether the committee would approve of the vehicle or not.
9. Suppose that n students are selected at random without replacement from a class containing T students, of whom A are boys and $T - A$ are girls. Let X denote the number of boys that are obtained. Find the sample size n such that the variance of X is maximum.
10. A bus arrives every 10 minutes at a bus stop. It is assumed that the waiting time for a particular individual is a random variable with a continuous uniform distribution.
- (a) What is the probability that the individual waits more than 7 minutes?
 - (b) What is the probability that the individual waits between 2 and 7 minutes?

11. A new battery supposedly with a charge of 1.5 volts actually has a voltage with a uniform distribution between 1.43 and 1.60 volts. Then,
- What is the expectation and the standard deviation of the voltage?
 - What is its *cumulative distribution function*?
 - If a box contains 50 batteries, what is the expectation and variance of the number of batteries in the box with a voltage less than 1.5 volts?
12. Suppose that components have failure times that are independent and can be modeled with an exponential distribution with parameter value of 0.0065 per day. If a box contains ten components, what is the probability that the box has at least eight components that last longer than 150 days?
13. Suppose a certain mechanical component produced by a company has a width that is normally distributed with mean of 2600 and a standard deviation of 0.6. Find
- the proportion of the components having a width in the range 2599 to 2601.
 - If the company needs to be able to guarantee to its purchaser that no more than 1 in 1000 of the components have a width outside the range 2599 to 2601, by how much does the standard deviation need to be reduced?
14. Find the probability density function if its cumulative density function is given as
- $$F(x) = (1 - e^{-x})U(x - c), \alpha > 0$$
- and $U(x) = 1, \forall x \geq 0$ and 0 otherwise. Consider the derivative of $U(x)$ as $\delta(x)$.
15. Let X and Y be independent Poisson variables with parameters λ and μ respectively. Show that:
- $X + Y$ is Poisson distributed with parameter $\lambda + \mu$.
 - the conditional distribution of X , given $X+Y = n$, is binomial and also find its parameters.
16. Suppose we observe a random variable X and then based on it, we make an attempt to predict the value of another random variable Y . Let $g(X)$ be the predictor of Y . Then, find the best possible predictor of Y .
17. Let X be a binomial random variable with parameters (n, p) . Show that as k goes from 0 to n , the *probability mass function* of X first increases monotonically and then decreases monotonically, reaching its largest value when k is the largest integer less than or equal to $(n + 1)p$.

6.17 Practice Problem Set-3 Solutions

1. Given that there are 20 multiple-choice questions, and each of them have 4 possible options.

We know answers of only 10 questions and for the remaining 10 questions we randomly choose an option. The random variable X is the total number of correct answers.

Let us consider that Y is the random variable taking the number of correct answers to the 10 questions that we randomly answer. Then, the total score becomes $X = Y + 10$.

Since, there are 4 options for every question, the success probability for randomly answering the question is $\frac{1}{4}$. So, we need to perform 10 independent Bernoulli trials with parameter $\frac{1}{4}$ and Y is the number of successes. Thus, we can say that $Y \sim \text{Binomial}(10, \frac{1}{4})$.

Now, as $X = Y + 10$, the range of X is $(10, 11, \dots, 20)$. Since, we know the pmf of Y , we can derive the pmf of X as, where k is in the range of X ,

$$P_X(k) = P(X = k) = P(Y + 10 = k) = P(Y = k - 10)$$

$$\Rightarrow P_X(k) = \binom{10}{k-10} \left(\frac{1}{4}\right)^{k-10} \left(\frac{3}{4}\right)^{10-(k-10)} = \binom{10}{k-10} \left(\frac{1}{4}\right)^{k-10} \left(\frac{3}{4}\right)^{20-k}$$

So the pmf of the random variable X would be given as

$$P_X(k) = \begin{cases} \binom{10}{k-10} \left(\frac{1}{4}\right)^{k-10} \left(\frac{3}{4}\right)^{20-k}, & k = 10, 11, \dots, 20 \\ 0, & \text{otherwise} \end{cases}$$

The probability of getting a score of more than 15 is given as

$$\begin{aligned} P(X > 15) &= P_X(16) + P_X(17) + P_X(18) + P_X(19) + P_X(20) \\ &= \binom{10}{6} \left(\frac{1}{4}\right)^6 \left(\frac{3}{4}\right)^4 + \binom{10}{7} \left(\frac{1}{4}\right)^7 \left(\frac{3}{4}\right)^3 + \binom{10}{8} \left(\frac{1}{4}\right)^8 \left(\frac{3}{4}\right)^2 + \binom{10}{9} \left(\frac{1}{4}\right)^9 \left(\frac{3}{4}\right)^1 + \binom{10}{10} \left(\frac{1}{4}\right)^{10} \left(\frac{3}{4}\right)^0 \\ &= 210 * 0.772 * 10^{-4} + 120 * 0.257 * 10^{-4} + 45 * 0.0853 * 10^{-4} + 10 * 0.0286 * 10^{-4} + 0.0095 * 10^{-4} \\ &= 0.016212 + 0.003089 + 0.000386 + 0.0000286 + 0.00000095 = 0.0197 \end{aligned}$$

Hence, the required probability is 0.0197.

2. Given a discrete random variable X .

(a) The range of the random variable X consists of all possible values of X . So, from the given pmf, the range of X is obtained as

$$R_X = \{0.2, 0.4, 0.5, 0.8, 1\}.$$

$$(b) P(0.25 < X < 0.75) = P(X = 0.4) + P(X = 0.5) = 0.2 + 0.2 = 0.4$$

(c) The conditional probability, $P(X = 0.2|X < 0.6)$ is given as

$$\begin{aligned} P(X = 0.2|X < 0.6) &= \frac{P(\{X = 0.2 \cap X < 0.6\})}{P(X < 0.6)} \\ &= \frac{P(X = 0.2)}{P(X < 0.6)} = \frac{P(X = 0.2)}{P(X = 0.2) + P(X = 0.4) + P(X = 0.5)} \\ &= \frac{0.1}{0.1 + 0.2 + 0.2} = 0.2 \end{aligned}$$

3. Given an experiment of rolling two dice, one red and one blue and X is a random variable taking twice the value of the blue die if the red die has an even value and to be the value of the red die minus the value of the blue die if the red die has an odd value.

Without loss of generality, we can assume that the first die is the red one and the second one is the blue die and the output of this experiment is denoted as (r,b) , where 'r' takes the value of the red die and 'b' takes the value of blue die.

The sample space of rolling two dice is $\{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$.

The value of the random variable X is given by

$$X = \begin{cases} 2b, & r = \text{even} \\ r - b, & r = \text{odd} \end{cases}$$

So, for an even number on the red die, X would take on the following values $\{2, 4, 6, 8, 10, 12\}$ and for odd values of red die, X would take on the following values $\{-5, -4, -3, -2, -1, 0, 1, 3\}$. So the range of X is $\{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 6, 8, 10, 12\}$.

The pmf of the random variable X is given as follows:

$$P(X = -5) = P((r,b) = (1,6)) = \frac{1}{36}$$

$$P(X = -4) = P((r,b) = (1,5)) = \frac{1}{36}$$

$$P(X = -3) = P((r,b) = (1,4)) + P((r,b) = (3,6)) = \frac{2}{36}$$

$$P(X = -2) = P((r, b) = (1, 3)) + P((r, b) = (3, 5)) = \frac{2}{36}$$

$$P(X = -1) = P((r, b) = (1, 2)) + P((r, b) = (3, 4)) + P((r, b) = (5, 6)) = \frac{3}{36}$$

$$P(X = 0) = P((r, b) = (1, 1)) + P((r, b) = (3, 3)) + P((r, b) = (5, 5)) = \frac{3}{36}$$

$$P(X = 1) = P((r, b) = (3, 2)) + P((r, b) = (5, 4)) = \frac{2}{36}$$

$$P(X = 2) = P((r, b) = (2, 1)) + P((r, b) = (3, 1)) + P((r, b) = (4, 1)) + P((r, b) = (5, 3)) +$$

$$P((r, b) = (6, 1))$$

$$\Rightarrow P(X = 2) = \frac{5}{36}$$

$$P(X = 3) = P((r, b) = (5, 2)) = \frac{1}{36}$$

$$P(X = 4) = P((r, b) = (2, 2)) + P((r, b) = (4, 2)) + P((r, b) = (5, 1)) + P((r, b) = (6, 2)) = \frac{4}{36}$$

$$P(X = 6) = P((r, b) = (2, 3)) + P((r, b) = (4, 3)) + P((r, b) = (6, 3)) = \frac{3}{36}$$

$$P(X = 8) = P((r, b) = (2, 4)) + P((r, b) = (4, 4)) + P((r, b) = (6, 4)) = \frac{3}{36}$$

$$P(X = 10) = P((r, b) = (2, 5)) + P((r, b) = (4, 5)) + P((r, b) = (6, 5)) = \frac{3}{36}$$

$$P(X = 12) = P((r, b) = (2, 6)) + P((r, b) = (4, 6)) + P((r, b) = (6, 6)) = \frac{3}{36}$$

So the pmf of the random variable X is given as

$$f_X(x) = \begin{cases} \frac{1}{36}, & X = -5 \\ \frac{1}{36}, & X = -4 \\ \frac{2}{36}, & X = -3 \\ \frac{2}{36}, & X = -2 \\ \frac{3}{36}, & X = -1 \\ \frac{3}{36}, & X = 0 \\ \frac{2}{36}, & X = 1 \\ \frac{5}{36}, & X = 2 \\ \frac{1}{36}, & X = 3 \\ \frac{4}{36}, & X = 4 \\ \frac{3}{36}, & X = 6 \\ \frac{3}{36}, & X = 8 \\ \frac{3}{36}, & X = 10 \\ \frac{3}{36}, & X = 12 \end{cases}$$

4. Given a random variable X , which is the time taken by a garage to service a car distributed between 0 and 10 hours with a cumulative distribution function (cdf)

$$F(x) = A + B * \ln(3x + 2), 0 \leq x \leq 10$$

(a) From the properties of the cdf, we know that $F(0) = 0$ (lower limit)

$$\begin{aligned} \Rightarrow A + B \ln(3 * 0 + 2) &= 0 \\ = A + B \ln(2) &= 0 \end{aligned} \tag{1}$$

$$\begin{aligned} F(10) &= 1 \text{(upper limit)} \\ \Rightarrow A + B \ln(3 * 10 + 2) &= 1 \\ = A + B \ln(32) &= 1 \end{aligned} \tag{2}$$

Subtracting (1) from (2) gives us,

$$\begin{aligned} B \ln(32) - B \ln(2) &= 1 \\ = B \ln(16) &= 1 \\ \Rightarrow B &= \frac{1}{\ln(16)} \end{aligned}$$

This gives us $A = -0.25$. Hence, the values of A and B are -0.25 and 0.361 respectively.

So, the cdf becomes

$$F(x) = -0.25 + 0.361 \ln(3x + 2), 0 \leq x \leq 10.$$

- (b) The probability that a repair job takes longer than two hours is

$$\begin{aligned} P(X > 2) &= 1 - P(X \leq 2) = 1 - F(2) \\ &= 1 - (-0.25 + 0.361 * \ln(3 * 2 + 2)) = 0.4993 \end{aligned}$$

5. Given that two persons A and B transmit messages between them and another person C eavesdrops on them, with a probability of 0.8, independently of the other messages. On a certain day, A and B sent each other a total of 8 messages and X is the number of those 8 messages that were eavesdropped by C .

- (a) Since C can either eavesdrop the messages or can't, with a success probability of 0.8 and also C eavesdrops each message independently. Hence, the distribution of the random variable X is Binomial distribution, $\text{Bin}(8, 0.8)$. Thus, the pmf of X is

$$P_X(k) = \binom{8}{k} \left(\frac{8}{10}\right)^k \left(\frac{2}{10}\right)^{8-k}$$

Since, it is a Binomial distribution, its mean would be given as

$$E[X] = np = 8 * 0.8 = 6.4$$

- (b) The most probable number of messages eavesdropped by C is the mode of the Binomial distribution, which is given by

$$\lfloor (n+1)p \rfloor = \lfloor (8+1)0.8 \rfloor = \lfloor 7.2 \rfloor = 7$$

- (c) The probability that exactly two messages were eavesdropped, given that at least one message was eavesdropped is given as

$$\begin{aligned} P(X = 2 | X \geq 1) &= \frac{P(X = 2 \cap X \geq 1)}{P(X \geq 1)} = \frac{P(X = 2)}{P(X \geq 1)} \\ &= \frac{P(X = 2)}{1 - P(X = 0)} \\ &= \frac{\binom{8}{2} \left(\frac{8}{10}\right)^2 \left(\frac{2}{10}\right)^{8-2}}{1 - \binom{8}{0} \left(\frac{8}{10}\right)^0 \left(\frac{2}{10}\right)^{8-0}} \\ &= 4.096 * 10^{-5} \end{aligned}$$

6. Given a game in which a fair coin is tossed ten times and a player wins if at least four coin tosses show tails. The random variable X denotes the number of tails observed in the game, and W denotes the event that the player wins.

- (a) The probability that the player wins is $P(W)$ given as

$$P(W) = P(X \geq 4) = 1 - P(X < 4)$$

Since each trial is independent of other trials and also each trial has only two outcomes, the distribution of the random variable X is Binomial distribution, $Bin(10, 0.5)$. Hence, the required probability is given as

$$\begin{aligned} P(W) &= 1 - \left[\binom{10}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{10} + \binom{10}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^9 + \binom{10}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^8 + \binom{10}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^7 \right] \\ &= 0.828 \end{aligned}$$

$$(b) P(X \leq 5|W) = \frac{P((X \leq 5) \cap (X \geq 4))}{P(X \geq 4)} = \frac{P(X=5)+P(X=4)}{P(X \geq 4)}$$

$$= \frac{\binom{10}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^6 + \binom{10}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^5}{P(X \geq 4)} = 231424 = 0.5448$$

7. Given that vehicles arrive at a road signal at an average rate of 360 per hour and the cycle of the traffic lights is set at 40 seconds. So the average rate for 40 seconds would be 4, since the rate of 360 per hour converges to 0.1 per second.

Since here the problem of interest is to find the probability of vehicles arriving in an interval of time, the random variable X , denoting the number of vehicles arriving, takes on the Poisson distribution with parameter, $\lambda = 4$ per second.

- (a) The probability that there would be exactly 5 vehicles arriving is

$$P(X = 5) = \frac{\lambda^5}{5!} e^{-\lambda} = \frac{4^5}{5!} e^{-4} = 8.533 * e^{-4} = 0.15629.$$

Hence, the required percentage is 15.62.

- (b) The probability that there would be less than 5 vehicles arriving is

$$P(X < 5) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

$$= e^{-4} \left[1 + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \frac{\lambda^4}{4!} \right]$$

$$= e^{-4} \left[1 + 4 + 8 + \frac{32}{3} + \frac{32}{3} \right] \\ = 0.6288.$$

Hence, the required percentage is 62.88.

8. Given, a committee has been formed to decide whether to base a recovery vehicle on a stretch of road to clear accidents as quickly as possible. The concerned road carries over 5000 vehicles during peak rush hour period. On average, the number of incidents during the peak hour is 5. The committee won't base a vehicle on the road if the percentage of having more than 5 incidents is less than 30 %.

Since here the problem of interest is to find the probability of accidents in an interval of time, the random variable X , denoting the number of accidents, takes on the Poisson distribution with parameter, $\lambda = 5$ per hour.

The probability of having more than 5 incidents is given as

$$P(X > 5) = 1 - P(X \leq 5) \\ = 1 - [e^{-5} \left(1 + \frac{5^1}{1!} + \frac{5^2}{2!} + \frac{5^3}{3!} + \frac{5^4}{4!} + \frac{5^5}{5!} \right)] \\ = 1 - [0.00674 + 0.03369 + 0.08422 + 0.14037 + 0.17547 + 0.17547] \\ = 1 - 0.61596 = 0.384.$$

So, the percentage of having more than 5 accidents is 38.4, which is more than 30 %. Hence, the committee should place the recovery vehicle.

9. Given that n students are selected at random without replacement from a class of T students, of whom A are boys and $T - A$ are girls. The random variable X denotes the number of boys obtained. Now, we have to find a sample size n such that the variance is maximum.

Here, the core is that n students are drawn without replacement out of T students, of which A are boys and $T - A$ are girls. The number of boys is our success rate. Hence, the random variable X takes on the hyper-geometric distribution with parameters $A, T - A, n$.

The variance of X is given as

$$\text{Var}(X) = n * \frac{A}{T} * (1 - \frac{A}{T}) * \frac{T - n}{T - 1} \\ = \frac{nA(T - A)(T - n)}{T^2(T - 1)}$$

$$= \frac{A(T - A)(nT - n^2)}{T^2(T - 1)}$$

To get the value of n such that the variance is maximum is obtained by differentiating $\text{Var}(X)$ wrt n and equating it to 0,

$$\frac{A(T - A)(T - 2n)}{T^2(T - 1)} = 0$$

$$\Rightarrow n = \frac{T}{2}$$

Since, n can be integer only, the required n value will be $\frac{T}{2}$ if T is an even number and if T is odd, then the required n value will be $\frac{T-1}{2}$ and $\frac{T+1}{2}$.

10. Given that a bus arrives every 10 minutes at a bus stop and the waiting time for a particular individual is a random variable, X , with a continuous uniform distribution. So, we can say that $X \sim \text{Uni}(0, 10)$.

- (a) The probability that the individual waits more than 7 minutes is given as

$$\begin{aligned} P(X > 7) &= \frac{1}{10 - 0} \int_{x=7}^{10} dx \\ &= \frac{1}{10} * \frac{10 - 7}{1} = 0.3 \end{aligned}$$

- (b) The probability that the individual waits between 2 and 7 minutes is given as

$$\begin{aligned} P(2 \leq X \leq 7) &= \frac{1}{10 - 0} \int_{x=2}^{7} dx \\ &= \frac{1}{10} * \frac{7 - 2}{1} = 0.5 \end{aligned}$$

11. Given that a new battery supposedly with a charge of 1.5 volts actually has a voltage with a uniform distribution between 1.43 and 1.60 volts. So the random variable, X , denoting the voltage of the battery is $X \sim \text{Uni}(1.43, 1.60)$.

- (a) The expectation of the voltage is given as

$$E[X] = \frac{\text{sum of limits}}{2} = 1.43 + 1.602 = 1.515$$

and the standard deviation of the voltage

$$\begin{aligned} \sigma &= \sqrt{\frac{(\text{difference between the limits})^2}{12}} = \frac{1.60 - 1.43}{\sqrt{12}} \\ &= 0.0491 \end{aligned}$$

(b) Its cumulative distribution function is given as

$$\begin{aligned} F(x) &= \int_{1.43}^x \frac{1}{1.60 - 1.43} dx = \frac{x - 1.43}{1.60 - 1.43} \\ &= \frac{x - 1.43}{0.17}, \quad 1.43 \leq x \leq 1.60 \end{aligned}$$

(c) Given that a box contains 50 batteries. Let X be the random variable denoting the number of batteries with a voltage less than 1.5 volts. Then, X would take on the Binomial distribution with parameters $n = 50$ and a success probability equals the probability that a battery will have a voltage less than 1.5 volts.

The probability that a battery will have a voltage less than 1.5 volts is given as $F(1.5) = \frac{1.5 - 1.43}{0.17} = \frac{0.07}{0.17} = 0.412$. Hence, the random variable $X \sim \text{Bin}(50, 0.412)$. So, the expected value is $E[X] = np = 50 * 0.412 = 20.6$ and the variance of X is $\text{Var}(X) = np(1 - p) = 50 * 0.412 * 0.588 = 12.11$.

12. Given that components have failure times that are independent are modeled with an exponential distribution with parameter value of 0.0065 per day. A box contains ten components. Let T be the random variable denoting the failure time.

Let X be the random variable denoting the number of components that last longer than 150 days. Hence, the random variable X takes on the Binomial distribution with parameters $n = 10$ and a success probability, p , equals the probability that it will last longer than 150 days. Now, since failure times are modeled by exponential distribution, we can get p as follows,

$$\begin{aligned} p &= P(T > 150) = 1 - P(T \leq 150) \\ &= 1 - (1 - e^{-0.0065*150}) = e^{-0.0065*150} = 0.377 \end{aligned}$$

Then, the required probability is

$$\begin{aligned} P(X \geq 8) &= P(X = 8) + P(X = 9) + P(X = 10) \\ &= \binom{10}{8} (0.377)^8 (0.623)^2 + \binom{10}{9} (0.377)^9 (0.623)^1 + \binom{10}{10} (0.377)^1 0 (0.623)^0 \\ &= 0.00713 + 0.00096 + 0.00006 = 0.00815. \end{aligned}$$

13. Given that a certain mechanical component produced by a company has a width that is normally distributed with mean of 2600 and a standard deviation of 0.6. Let X be the random variable taking the width value, then $X \sim N(2600, 0.36)$.

(a) The probability of the components having a width in the range 2599 to 2601 is given by

$$\begin{aligned} P(2599 \leq X \leq 2601) &= \phi\left(\frac{2601 - 2600}{0.6}\right) - \phi\left(\frac{2599 - 2600}{0.6}\right) \\ &= 0.9522 - 0.0478 = 0.9044. \end{aligned}$$

(b) Given that the company needs to be able to guarantee to its purchaser that no more than 1 in 1000 of the components have a width outside the range 2599 to 2601 and we need to find a suitable standard deviation.

$$P(2599 \leq X \leq 2601) = 1 - 0.001 = 0.999$$

$$\begin{aligned} \Rightarrow \phi\left(\frac{2601 - 2600}{\sigma}\right) - \phi\left(\frac{2599 - 2600}{\sigma}\right) &= 0.999 \\ &= \phi\left(\frac{1}{\sigma}\right) - \phi\left(-\frac{1}{\sigma}\right) = 0.999 \\ &= \phi\left(\frac{1}{\sigma}\right) - (1 - \phi\left(\frac{1}{\sigma}\right)) = 0.999 \\ &= 2\phi\left(\frac{1}{\sigma}\right) - 1 = 0.999 \\ \Rightarrow \phi\left(\frac{1}{\sigma}\right) &= \frac{1 + 0.999}{2} = 0.9995 \\ \frac{1}{\sigma} &= \phi^{-1}(0.9995) = 3.2905 \\ \Rightarrow \sigma &= \frac{1}{3.2905} = 0.304. \end{aligned}$$

So the required standard deviation is 0.304.

14. Given the cumulative density function is

$$F(x) = (1 - e^{-x})U(x - c), \alpha > 0$$

and $U(x) = 1, \forall x \geq 0$ and 0 otherwise.

Its pdf can be obtained as

$$\begin{aligned} \frac{d(F(x))}{dx} &= \frac{d((1 - e^{-x})U(x - c))}{dx} \\ &= U(x - c) * \frac{d(1 - e^{-x})}{dx} + (1 - e^{-x}) * \frac{d(U(x - c))}{dx} \\ &= U(x - c) * (0 - e^{-\alpha x} * (-\alpha)) + (1 - e^{-\alpha x}) * \delta(x - c), \end{aligned}$$

$$= \alpha e^{-\alpha x} * U(x - c) + (1 - e^{-\alpha x}) * \delta(x - c),$$

Hence, the required pdf is $\alpha e^{-\alpha x} * U(x - c) + (1 - e^{-\alpha x}) * \delta(x - c)$.

15. Given that X and Y are independent Poisson variables with parameters λ and μ respectively.

- (a) Since, X and Y are both random variables, $Z = X + Y$ would be a random variable. Its probability mass function would be given as, from total probability

$$\begin{aligned} P(X + Y = n) &= \sum_{k=0}^n P(X + Y = n | Y = k) P(Y = k) \\ &= \sum_{k=0}^n P(X = n - k) P(Y = k) \end{aligned}$$

Since, both X and Y are independent Poisson random variables, we can write the above equation as

$$= \sum_{k=0}^n \frac{e^{-\lambda} \lambda^{n-k}}{(n-k)!} * \frac{e^{-\mu} \mu^k}{k!} = e^{-(\lambda+\mu)} \sum_{k=0}^n \frac{\lambda^{n-k}}{(n-k)!} * \frac{\mu^k}{k!}$$

Multiplying and dividing by $n!$, we get

$$\begin{aligned} &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^n \frac{*n!}{(n-k)! * k!} * \lambda^{n-k} \mu^k \\ &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^n \binom{n}{k} \lambda^{n-k} \mu^k \end{aligned}$$

Using the binomial expansion, we get conclude that

$$\begin{aligned} &= \frac{e^{-(\lambda+\mu)}}{n!} (\lambda + \mu)^n \\ \Rightarrow P(X + Y = n) &= \frac{e^{-(\lambda+\mu)}}{n!} (\lambda + \mu)^n \end{aligned}$$

Hence, the random variable $Z = X + Y$ is a Poisson distribution with parameter $(\lambda + \mu)$.

- (b) The conditional distribution of X , given $X + Y = n$, is given by

$$P(X = k | X + Y = n) = \frac{P(X = k, X + Y = n)}{P(X + Y = n)}$$

Since, X and Y are independent, the above equation can be simplified as

$$= \frac{P(X = k) P(X + Y = n)}{P(X + Y = n)}$$

$$\begin{aligned}
&= \frac{e^{-\lambda} \lambda^k}{k!} * \frac{e^{-\mu} \mu^{n-k}}{(n-k)!} * \frac{n!}{e^{-(\lambda+\mu)} (\lambda + \mu)^n} \\
&= \frac{e^{-\lambda} \lambda^k e^{-\mu} \mu^{n-k} n!}{k! (n-k)! e^{-(\lambda+\mu)} (\lambda + \mu)^n} \\
&= \frac{n!}{(n-k)! k!} \left(\frac{\lambda}{\lambda + \mu}\right)^k \left(\frac{\mu}{\lambda + \mu}\right)^{n-k} \\
&= \frac{n!}{(n-k)! k!} \left(\frac{\lambda}{\lambda + \mu}\right)^k \left(1 - \frac{\lambda}{\lambda + \mu}\right)^{n-k}
\end{aligned}$$

Hence, the required conditional probability is binomial with parameters n and $\frac{\lambda}{\lambda + \mu}$.

16. Given that we observe a random variable X and then based on it, we make an attempt to predict the value of another random variable Y and that predictor of Y is $g(X)$. Now, we need to find the best possible predictor of Y , i.e., we need to find $g(X)$ that tends close to Y . One way to do this is to choose g which minimizes $E[(Y - g(X))^2]$.

Now, consider $E[(Y - g(X))^2 | X]$ and add and subtract $E[Y|X]$, so we get

$$\begin{aligned}
&= E[(Y - E[Y|X] + E[Y|X] - g(X))^2 | X] \\
&= E[(Y - E[Y|X])^2 | X] + E[(E[Y|X] - g(X))^2 | X] + 2E[(Y - E[Y|X])(E[Y|X] - g(X)) | X]
\end{aligned}$$

Here, it noteworthy to note that given X , $E[Y|X] - g(X)$, being a function of X , is a constant. Thus, we can write

$$\begin{aligned}
&= E[(Y - E[Y|X])(E[Y|X] - g(X)) | X] \\
&= (E[Y|X] - g(X))E[(Y - E[Y|X]) | X] \\
&= (E[Y|X] - g(X))(E[(Y|X) - E[Y|X]]) = 0
\end{aligned}$$

Substituting this in the above main equation, we get

$$E[(Y - g(X))^2 | X] E[(Y - E[Y|X])^2 | X]$$

Now, since given X , all the above are just constants, on applying expectation to it once again we get the following,

$$E[(Y - g(X))^2] \geq E[(Y - E[Y|X])^2].$$

17. Given that X is a binomial random variable, with parameters (n, p) . We need to show that as k goes from 0 to n , the probability mass function of X first increases monotonically and then decreases monotonically, reaching its largest value when k is the largest integer less than or equal to $(n+1)p$.

To show the required trend, let us consider the probability of two consecutive k values.

$$\begin{aligned}
\frac{p(k)}{p(k-1)} &= \frac{\binom{n}{k} p^k (1-p)^{n-k}}{\binom{n}{k-1} p^{k-1} (1-p)^{n-k+1}} \\
&= \frac{\frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}}{\frac{n!}{(n-k+1)!(k-1)!} \frac{p^k}{p} (1-p)^{n-k} (1-p)} \\
&= \frac{n!}{(n-k)!k(k-1)!} * p^k (1-p)^{n-k} * \frac{(n-k+1)(n-k)!(k-1)!p}{n!p^k (1-p)^{n-k} (1-p)} \\
&= \frac{(n-k+1)p}{k(1-p)}
\end{aligned}$$

Note that from above, we can conclude that $p(k) \geq p(k-1)$ if and only if

$$\begin{aligned}
(n-k+1)p &\geq k(1-p) \\
np - kp + p &\geq k - kp \\
np + p &\geq k \\
\Rightarrow (n+1)p &\geq k.
\end{aligned}$$

Thus, $p(k)$ increases monotonically and reaches its maximum when k is the largest integer less than or equal to $(n+1)p$ and thereafter decreases monotonically.

Goals, X	Probability, $P(X)$
0	0.18
1	0.34
2	k
3	0.02
4	0.11

Table 6.5: Probability Distribution of goals

6.18 Practice Problem set-4

1. Two dice are rolled together and the sum of outcome of each die is observed.
 - (a) Find the probability distribution of random variable representing the sum.
 - (b) Find the probability for sum greater than 6 given one of die has 4 as outcome.
2. A fair coin is tossed and a dice is thrown simultaneously.
 - (a) Find the joint probability of the two events.
 - (b) Find the probability the coin tossed is tail and dice has number less than 5.
 - (c) Find value of k for the distribution given in table 1.1.
3. Two random variable X and Y have the joint probability distribution

$$F_{XY}(x, y) = \begin{cases} \frac{5}{4} \left(\frac{x+e^{-(x+1)y^2}}{x+1} - e^{-y^2} \right) & 0 \leq x \leq 4, \quad y \geq 0 \\ 1 + \frac{1}{4}e^{-5y^2} - \frac{5}{4}e^{-y^2} & 4 \leq x, \quad y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- (a) Find the marginal distribution $F_X(x)$.
 - (b) Find the marginal distribution $F_Y(y)$.
 - (c) Find the probability $P(3 < X \leq 5, 1 < Y \leq 2)$.
4. The joint probability mass function of X and Y is given in the figure 1.22.

$Y \setminus X$	1	2
1	1/8	1/8
2	1/4	1/2

Fig. 6.33: Joint probability mass function

- (a) Compute the marginal pmfs $f_X(x)$ and $f_Y(y)$.
- (b) Compute the conditional mass function of X given $Y = i$, where $i = 1, 2$.
- (c) Find $P(X|Y > 1)$.
5. The power reflected from an aircraft of complicated shape that is received by a radar can be described by a random variable X . The *pdf* of X is

$$f_X(x) = \begin{cases} \frac{1}{W} e^{-\frac{x}{W}} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where W is the average amount of power received. Find the probability that the received power is larger than the power received on an average.

6.19 Practice set -4 solutions

1. Let X and Y denote the random variables taking values on the first and the second die, respectively. Then, S be the random variable denoting the sum $X + Y$.

- (a) Then, the range of S would be

$$R_S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Their probabilities are $P(S = 2) = P(\{1, 1\}) = \frac{1}{36}$.

$$P(S = 3) = P(\{1, 2\}) + P(\{2, 1\}) = \frac{2}{36}.$$

$$P(S = 4) = P(\{1, 3\}) + P(\{3, 1\}) + P(\{2, 2\}) = \frac{3}{36}.$$

$$P(S = 5) = P(\{1, 4\}) + P(\{4, 1\}) + P(\{2, 3\}) + P(\{3, 2\}) = \frac{4}{36}.$$

$$P(S = 6) = P(\{1, 5\}) + P(\{5, 1\}) + P(\{2, 4\}) + P(\{4, 2\}) + P(\{3, 3\}) = \frac{5}{36}.$$

$$P(S = 7) = P(\{1, 6\}) + P(\{6, 1\}) + P(\{2, 5\}) + P(\{5, 2\}) + P(\{3, 4\}) + P(\{4, 3\}) = \frac{6}{36}.$$

$$P(S = 8) = P(\{2, 6\}) + P(\{6, 2\}) + P(\{3, 5\}) + P(\{5, 3\}) + P(\{4, 4\}) = \frac{5}{36}.$$

$$P(S = 9) = P(\{3, 6\}) + P(\{6, 3\}) + P(\{4, 5\}) + P(\{5, 4\}) = \frac{4}{36}.$$

$$P(S = 10) = P(\{4, 6\}) + P(\{6, 4\}) + P(\{5, 5\}) = \frac{3}{36}.$$

$$P(S = 11) = P(\{5, 6\}) + P(\{6, 5\}) = \frac{2}{36}.$$

$$P(S = 12) = P(\{6, 6\}) = \frac{1}{36}.$$

Thus, the *pmf* of S is given as

$$f_S(s) = \begin{cases} \frac{1}{36} & s = 2 \\ \frac{2}{36} & s = 3 \\ \frac{3}{36} & s = 4 \\ \frac{4}{36} & s = 5 \\ \frac{5}{36} & s = 6 \\ \frac{6}{36} & s = 7 \\ \frac{5}{36} & s = 8 \\ \frac{4}{36} & s = 9 \\ \frac{3}{36} & s = 10 \\ \frac{2}{36} & s = 11 \\ \frac{1}{36} & s = 12 \end{cases}$$

Y,X	1	2	3	4	5	6
H	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
T	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

Table 6.6: Joint Probability Mass Function

$$(b) P((S > 6|X = 4) \cup S > 6|Y = 4)$$

$$= P(S > 6|X = 4) + P(S > 6|Y = 4) - P(S > 6|X = 4, Y = 4)$$

$$\begin{aligned} &= \frac{P(S > 6, X = 4)}{P(X = 4)} + \frac{P(S > 6, Y = 4)}{P(Y = 4)} - \text{frac}P(S > 6, X = 4, Y = 4)P(X = 4, Y = 4) \\ &= \frac{2}{3} + \frac{2}{3} - 1 = \frac{4}{3} - 1 = \frac{1}{3}. \end{aligned}$$

2. Given, a fair coin is tossed and a die is thrown simultaneously. The sample space of tossing a fair coin is S_c ,

$$S_c = \{H, T\}$$

And, the sample space of rolling a die is S_d

$$S_d = \{1, 2, 3, 4, 5, 6\}$$

- (a) Let X be a random variable for number on the die i.e.,

$$X \in \{1, 2, 3, 4, 5, 6\}$$

And, Y be the random variable for tossing a fair coin

$$Y \in \{H, T\}$$

These two are independent experiments. Hence, we obtain the probabilities as

$$P(X = 1, Y = H) = \frac{1}{6} * \frac{1}{2}$$

$$P(X = 2, Y = T) = \frac{1}{6} * \frac{1}{2}$$

This holds for other events too. Hence, the joint probability mass function is as given in table 1.2.

- (b) The probability the coin tossed is tail and die has a number less than 5 is given by

$$P(Y = T, X < 5) = P(Y = T, X = 1) + P(Y = T, X = 2) + P(Y = T, X = 3) + P(Y =$$

$$\begin{aligned} T, X = 4) \\ &= \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} = \frac{4}{12} = \frac{1}{3} \end{aligned}$$

(c) We know that the sum of probabilities must be one. Hence,

$$0.18 + 0.34 + k + 0.02 + 0.11 = 1$$

$$\Rightarrow k + 0.65 = 1 \Rightarrow k = 0.35$$

3. Given, the joint probability distribution of the random variables X and Y as

$$F_{XY}(x, y) = \begin{cases} \frac{5}{4} \left(\frac{x+e^{-(x+1)y^2}}{x+1} - e^{-y^2} \right) & 0 \leq x \leq 4, \quad y \geq 0 \\ 1 + \frac{1}{4}e^{-5y^2} - \frac{5}{4}e^{-y^2} & 4 \leq x, \quad y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) The marginal cdf, $F_X(x)$, is given as

$$F_X(x) = F_{X,Y}(x, \infty) = \int_0^\infty \frac{5}{4} \left(\frac{x+e^{-(x+1)y^2}}{x+1} - e^{-y^2} \right) dy$$

for $0 \leq x \leq 4$.

$$= \frac{5x}{4(x+1)}$$

$$\Rightarrow F_X(x) = \begin{cases} \frac{5x}{4(x+1)} & 0 \leq x \leq 4 \\ 0 & x < 0 \\ 1 & x \geq 4 \end{cases}$$

(b) In the same way, the marginal cdf, $F_Y(y)$, is given as $F_Y(y) = F_{X,Y}(\infty, y)$

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ 1 + \frac{1}{4}e^{-5y^2} - \frac{5}{4}e^{-y^2} & y \geq 0 \end{cases}$$

(c) The probability $P(3 < X \leq 5, 1 < Y \leq 2)$ is given by

$$P(3 < X \leq 5, 1 < Y \leq 2) = F_{X,Y}(5, 2) + F_{X,Y}(3, 1) - F_{X,Y}(5, 1) - F_{X,Y}(3, 2)$$

$$F_{XY}(5, 2) = 1 + \frac{1}{4}e^{-5(2)^2} - \frac{5}{4}e^{-(2)^2} = 0.9771$$

$$F_{XY}(3, 1) = \frac{5}{4} \left(\frac{3+e^{-4*1}}{4} \right) - e^{-1} = 0.4883$$

$$F_{XY}(5, 1) = 1 + \frac{1}{4}e^{-5(1)^2} - \frac{5}{4}e^{-(1)^2} = 0.541$$

$$F_{XY}(3, 2) = \frac{5}{4}(\frac{3 + e^{-4*4}}{4}) - e^{-4} = 0.914$$

$$= 0.9771 + 0.4883 - 0.541 - 0.914 = 0.0104.$$

4. Given, the joint probability mass function of X and Y as shown in below table.

Y,X	1	2
1	$\frac{1}{8}$	$\frac{1}{8}$
2	$\frac{1}{4}$	$\frac{1}{2}$

(a) Then, the marginal pmf of X is obtained as

$$f_X(1) = \sum_{y=y_i} P(X = 1, y = y_i) = P(X = 1, Y = 1) + P(X = 1, Y = 2) = \frac{1}{8} + \frac{1}{4} = \frac{3}{8}$$

$$f_X(2) = \sum_{y=y_i} P(X = 2, y = y_i) = P(X = 2, Y = 1) + P(X = 2, Y = 2) = \frac{1}{8} + \frac{1}{2} = \frac{5}{8}$$

$$\Rightarrow f_X(x) = \begin{cases} \frac{3}{8} & x = 1 \\ \frac{5}{8} & x = 2 \end{cases}$$

and, the marginal pmf of Y is obtained as

$$f_Y(1) = \sum_{x=x_i} P(X = x_i, y = 1) = P(X = 1, Y = 1) + P(X = 2, Y = 1) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$$

$$f_Y(2) = \sum_{x=x_i} P(X = x_i, y = 2) = P(X = 1, Y = 2) + P(X = 2, Y = 2) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$

$$\Rightarrow f_Y(y) = \begin{cases} \frac{1}{4} & y = 1 \\ \frac{3}{4} & y = 2 \end{cases}$$

(b) The conditional mass function of X given $Y = i$, where $i = 1, 2$ can be obtained as

$$f_{X=1|Y=1} = \frac{P(X = 1, Y = 1)}{P(Y = 1)} = \frac{\frac{1}{8}}{\frac{1}{8} + \frac{1}{8}} = \frac{\frac{1}{8}}{\frac{1}{4}} = \frac{1}{2}$$

$$f_{X=2|Y=1} = \frac{P(X = 2, Y = 1)}{P(Y = 1)} = \frac{\frac{1}{8}}{\frac{1}{8} + \frac{1}{8}} = \frac{\frac{1}{8}}{\frac{1}{4}} = \frac{1}{2}$$

$$f_{X=1|Y=2} = \frac{P(X = 1, Y = 2)}{P(Y = 2)} = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{2}} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

$$f_{X=2|Y=2} = \frac{P(X=2, Y=2)}{P(Y=2)} = \frac{\frac{1}{2}}{\frac{1}{4} + \frac{1}{2}} = \frac{\frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}$$

(c) The pmf of $P(X|Y > 1)$ is given as

$$P(X|Y > 1) = \frac{P(X=x, Y=2)}{P(Y=2)}$$

$$P(X=1|Y > 1) = \frac{P(X=1, Y=2)}{P(Y=2)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

$$P(X=2|Y > 1) = \frac{P(X=2, Y=2)}{P(Y=2)} = \frac{\frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}$$

5. Given, the pdf of X

$$f_X(x) = \begin{cases} \frac{1}{W} e^{\frac{-x}{W}} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Then, the probability that the power received is greater than the average power W is given as

$$\begin{aligned} P(X > W) &= 1 - P(X \leq W) = 1 - \int_{-\infty}^W f_X(x) dx \\ &= 1 - \frac{1}{W} \int_0^W e^{\frac{-x}{W}} dx \\ &= 1 - \frac{1}{W} (-W) e^{\frac{-x}{W}} \Big|_0^W \\ &= 1 + e^{-1} - e^0 = e^{-1}. \end{aligned}$$

6.20 Lab Assignment 6

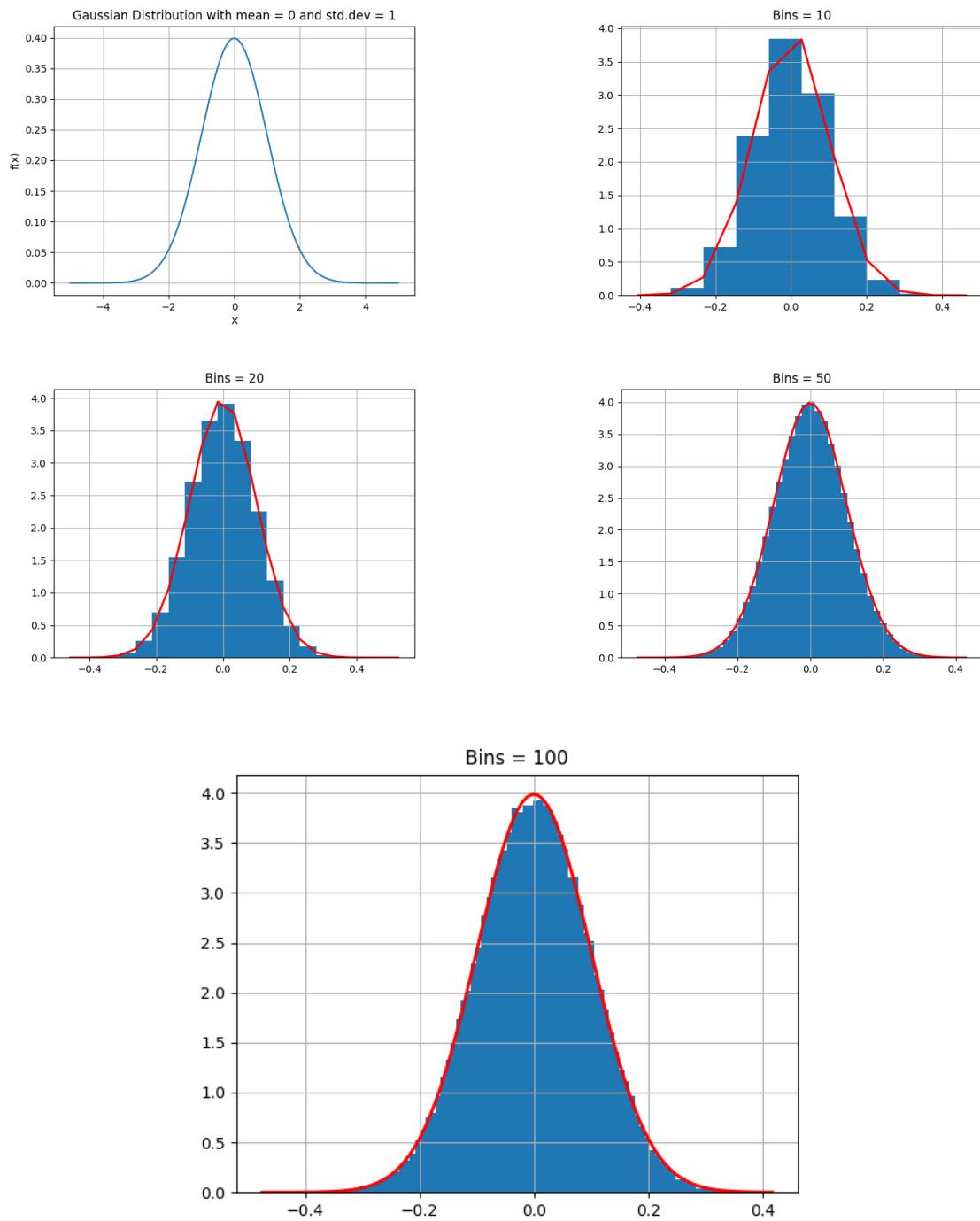
1. Let a continuous random variable X follows a Gaussian probability density function (PDF) $f_X(x)$ with mean μ and variance σ^2 , then
 - (a) Write a code to plot the PDF $f_X(x)$ of the random variable X with mean $\mu = 0$ and variance $\sigma^2 = 1$ (Use the PDF expression to plot).
 - (b) Write a code to generate data X , $N = 100,000$ sample points, following a Normal PDF with mean $\mu = 0$ and variance $\sigma^2 = 1$ (Can use the numpy function `np.random.randn`).
 - (c) Redo the part 1b for decreasing value of the bin width of the histogram and state your observations. Can the two axes be representative of something? Hence, explain the significance of the obtained histograms in relation to the $f_X(x)$ of part 1a.

```
1 from scipy.integrate import quad
2 import math
```

```

3 import numpy as np
4 import random
5 import matplotlib.pyplot as plt
6 ### Q1
7 from scipy import stats
8
9 ### Two ways to define gaussian distribution
10
11 # Way 1 (Using norm.pdf from scipy.stats)
12 x_data = np.arange(-5, 5, 0.001)
13 y_data = stats.norm.pdf(x_data, 0, 1)
14
15 # Way 2 (define ur own function)
16 def normal_dist(x , mean , sd):
17     prob_density = (math.sqrt(np.pi*2)*sd)
18     * np.exp(-0.5*((x-mean)/sd)**2)
19     return prob_density
20
21 plt.plot(x_data,y_data)
22 plt.xlabel("X")
23 plt.ylabel("f(x)")
24 plt.title("Gaussian Distribution with mean = 0 and std.dev = 1")
25 plt.grid()
26 plt.show()
27
28 ## B
29
30 mu, sigma = 0, 0.1 # mean and standard deviation
31 rand_data = np.random.normal(mu, sigma, 100000)
32 count, bins, ignored = plt.hist(rand_data, 100, density=True)
33 plt.plot(bins, 1/(sigma * np.sqrt(2 * np.pi)) *
34           np.exp( - (bins - mu)**2 / (2 * sigma**2) ),
35           linewidth=2, color='r')
36 plt.show()

```



2. Suppose that X is a continuous random variable whose probability distribution function is given by :

$$f(x) = \begin{cases} C(4x - 2x^2)e^{-x} & \text{if } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Calculate the constant C using the fact that

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

You might want to use `scipy.integrate` method for performing integration.

(b) Using **Matplotlib**, $f(x)$ plot as a function of x.

(c) Calculate the probability $P(1 < X < 2)$.

```
1 # Q2
2 # # A
3 from scipy.integrate import quad
4 import math
5 import numpy as np
6 import random
7 def integrand(x):
8     if(x>0 and x<2):
9         f = (4*(x)-2*(x**2))*np.exp(-x)
10    else:
11        f = 0
12    return f
13 I = quad(integrand, 0, 2)
14
15 # We know that C*I = 1, So C = 1/I
16
17 C = (1/I[0])
18 print("Constant C = ",C)
19
20 # # B
21
22 import matplotlib.pyplot as plt
23
24 # creating the test data
25 x = [x/100 for x in range(-400,400,1)]
26 y= [integrand(x/100) for x in range(-400,400,1) ]
27
28 fig,ax= plt.subplots()
29 ax.plot(x,y)
30 ax.set_title('C*(4x - 2x^2)e^-x for x E (0,2)')
31 ax.set_xlabel('X')
32 ax.set_ylabel('f(X)')
33 ax.grid()
34
35 plt.show()
```

```

37 # C
38
39 I = quad(integrand,1,2)
40 P = C*I[0]
41 print("P(1<X<2) = ",P)

```

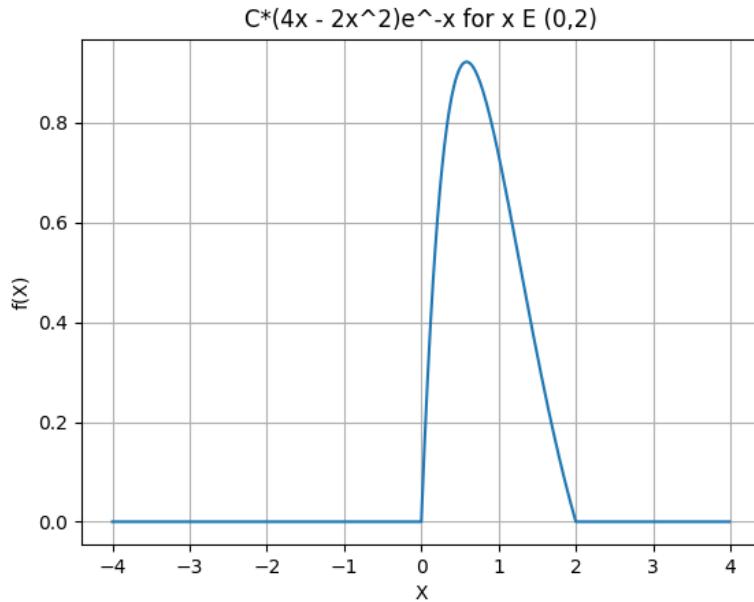


Fig. 6.34: Plot: Q2

3. The normal distribution is the most common type of distribution used in statistical analysis. In many fields which deal with large data you will notice that the data can be approximated to follow Normal distribution. Here, you are given a *marks.csv* file which contains the marks and grades of the students from a certain course at IIT Mandi. You are expected to do the following with the given data :

- (a) Find the mean and standard deviation of the given marks and check if the marks distribution follows the 3-sigma rule.
- (b) Assign the grades to the students based on the following rules :
 - Score > Average +1.5 SD -gives O
 - Average +1.5 SD > Score > Average + SD -gives A
 - Average + SD > Score > Average +0.5 SD -gives B
 - Average + 0.5 SD > Score > Average -gives C
 - Average > Score > Average - 0.5 SD -gives D

- Average - 0.5 SD > Score > Average - SD -gives E
- Average - SD > Score > Average -1.5 SD -gives F

Check whether the originally assigned grades match the grades you assigned using this scheme.

- Check whether the given data follows Normal distribution or not. Plot a histogram of the marks data. Generate random numbers following normal distribution with mean and standard deviation that you calculated in part (a). Now plot these random numbers by superimposing the curve over the histogram and see if the distribution fits our data or not.
- Standardize the random variable, marks, by applying the Z-transformation, and calculate the mean and standard deviation of new variable.
- Repeat a-d for IC252 marks.

```

1 # Importing Libraries
2
3 import pandas as pd
4 import numpy as np
5 import matplotlib.pyplot as plt
6 import statistics as st
7 from scipy.stats import norm
8 import seaborn as sns
9
10 # Calculating Mean and Standard deviation of Marks
11
12 def compute_mean_std(marks_df):
13     marks = list(marks_df['Marks'])
14     for i in range(len(marks)):
15         if(marks[i]=='A'):
16             marks[i] = 0
17             marks[i] = float(marks[i])
18     mean = st.mean(marks)
19     std_dev = st.stdev(marks)
20     print(mean, std_dev)
21     return marks,mean,std_dev
22
23 # Checking 3-sigma Rule
24
25 def check_three_sigma(marks_df):
26     marks,mean,std_dev = compute_mean_std(marks_df)
27     num_1 = []
28     num_2 = []

```

```

29     num_3 = []
30     outlier = []
31     total_num = len(marks)
32     for i in marks:
33         if (i > (mean - std_dev) and i < (mean + std_dev)):
34             num_1.append(i)
35     for i in marks:
36         if (i > (mean - (2 * std_dev)) and i < (mean + (2 * std_dev))):
37             num_2.append(i)
38     for i in marks:
39         if ((i > (mean - (3 * std_dev)) and i < (mean + (3 * std_dev)))):
40             num_3.append(i)
41     else :
42         outlier.append(i)
43     return num_1, num_2, num_3, total_num, outlier
44
45 # Theoretically Values should be - 68 %, 95 %, 99.7 % for a
46 normal distribution
47
48 def print_values(marks_df):
49     num_1, num_2, num_3, total_num, outlier =
50     check_three_sigma(marks_df)
51     print("Values within 1 std. dev =
52     ", round(len(num_1)*100/total_num,2), "%")
53     print("Values within 2 std. dev =
54     ", round(len(num_2)*100/total_num,2), "%")
55     print("Values within 3 std. dev =
56     ", round(len(num_3)*100/total_num,2), "%")
57     print("Outliers = ", round(len(outlier)*100/total_num,2), "%")
58
59 # Assigning Grades
60
61 def assign_grades(marks_df):
62     marks, mean, std_dev = compute_mean_std(marks_df)
63     Grades_Assigned = []
64     for i in marks:
65         if(i >= (mean + 1.5*std_dev)):
66             Grades_Assigned.append('O')
67         elif(i < (mean + 1.5*std_dev) and i >= (mean + std_dev)):
68             Grades_Assigned.append('A')
69         elif(i < (mean + std_dev) and i >= (mean + 0.5*std_dev)):
70             Grades_Assigned.append('B')
71         elif(i < (mean + 0.5*std_dev) and i >= (mean)):
72             Grades_Assigned.append('C')
73         elif(i < mean and i >= (mean - 0.5*std_dev)):
74             Grades_Assigned.append('D')

```

```

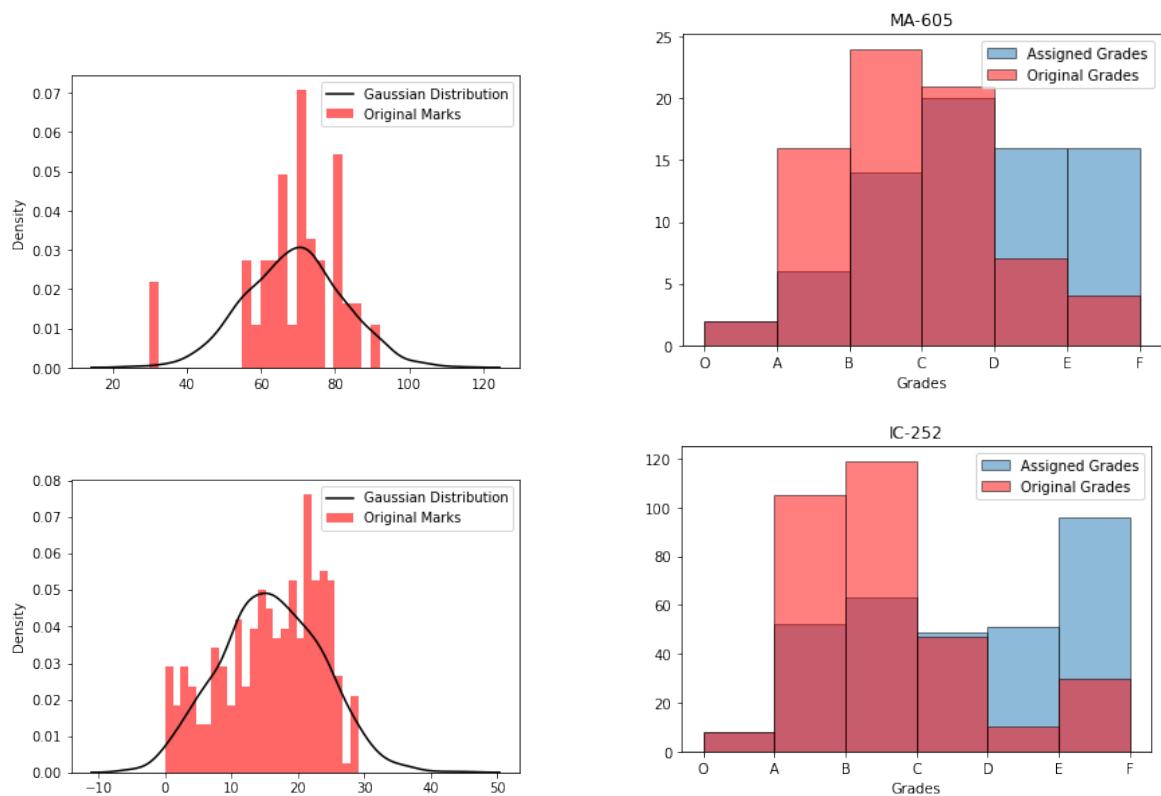
75         elif(i< (mean - 0.5*std_dev) and i>=(mean - std_dev)):
76             Grades_Assigned.append('E')
77         else:
78             Grades_Assigned.append('F')
79     return Grades_Assigned
80 def print_grades(marks_df, Original_Grades):
81     Grades_Assigned = assign_grades(marks_df)
82     print(Grades_Assigned, "\n", Original_Grades)
83
84 # Plot of Histogram + Normal distribution
85
86 def plot(marks_df):
87     marks, mean, std_dev = compute_mean_std(marks_df)
88     data = np.random.normal(mean, std_dev, size = 1000)
89
90     # Fit a normal distribution to
91     # the data:
92     # mean and standard deviation
93     mu, std = norm.fit(data)
94
95     # Plot the histogram.
96     plt.hist(marks, bins=25, density=True, alpha=0.6, color='red')
97     # Plot the PDF.
98     xmin, xmax = plt.xlim()
99     x = np.linspace(xmin, xmax, 1000)
100    p = norm.pdf(x, mu, std)
101    sns.distplot(data, hist=False, color='black')
102    plt.legend(['Gaussian Distribution', 'Original Marks'])
103    plt.show()
104
105 # Z-Transformation
106
107 def z_transform(marks_df):
108     marks, mean, std_dev = compute_mean_std(marks_df)
109     standardized_marks = []
110     for i in marks:
111         standardized_marks.append((i-mean)/std_dev)
112     return standardized_marks
113
114 df_ma605 = pd.read_csv('marks.csv')
115 df_ic252 = pd.read_csv('IC252_Midsem.csv')
116
117 # Results for MA-605 Marks List
118
119 marks1, mean1, std_dev1 = compute_mean_std(df_ma605)
120 print_values(df_ma605)

```

```

121 print_grades(df_ma605, list(df_ma605['Grades']))
122 plot(df_ma605)
123 standardized_marks = z_transform(df_ma605)
124 print(st.mean(standardized_marks), st.stdev(standardized_marks))
125
126 marks1, mean1, std_dev1 = compute_mean_std(df_ic252)
127 print_values(df_ic252)
128 print_grades(df_ic252, None)
129 plot(df_ic252)
130 standardized_marks = z_transform(df_ic252)
131 print(st.mean(standardized_marks), st.stdev(standardized_marks))

```



4. A Galton Board is a device used for statistical experiments and demonstrations. It consists of a vertical board with a series of evenly spaced pegs or nails arranged in a triangular pattern. At the top of the board, a large number of small balls or beads are dropped, and they bounce off the pegs as they descend, randomly falling into slots or compartments at the bottom of the board.

Let the number of balls be N and the number of levels be L (i.e. number of times the ball will hit pegs). Simulate a simple galton board in python under the following assumptions.

- (a) The balls do not interact with each other.

(b) The probability of a ball going to either side of a peg on impact is 0.5 each.

Now using the simulated galton board experiment, do the following tasks.

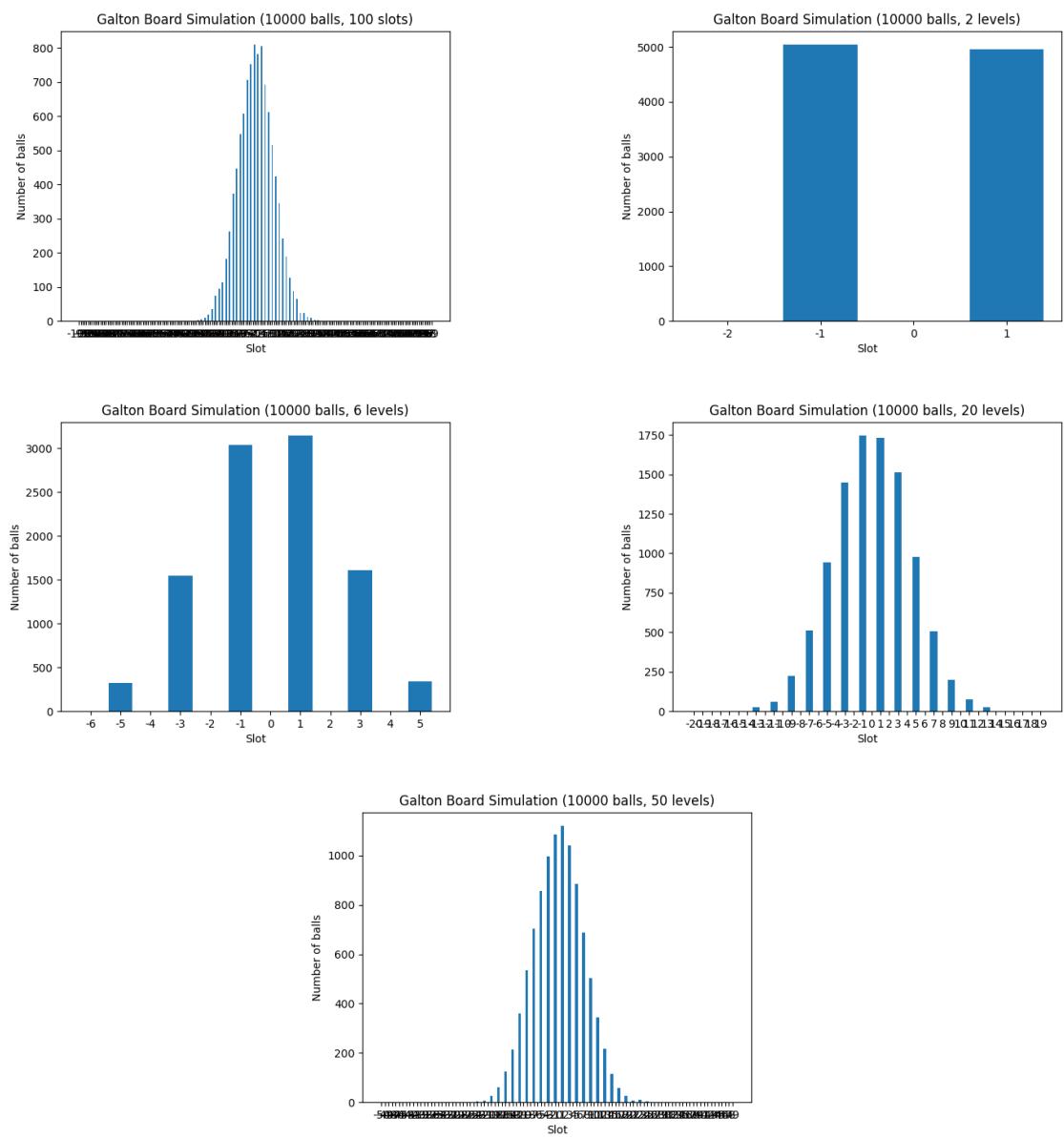
- (a) Plot the distribution of the balls in the slots for $L = [2, 6, 20, 100]$
- (b) Find the Mean and Variance of the distributions.
- (c) Difference in odd and even L.

```
1 # Q4, simulation of galton board
2
3 import random
4 import matplotlib.pyplot as plt
5 import math
6 import statistics
7
8 def galton_board(num_balls, num_slots):
9     # slots = {}
10    slots=dict.fromkeys(range(-1*num_slots,num_slots),0)
11    for i in range(num_balls):
12        pos = 0
13        for j in range(num_slots-1):
14            if random.random() < 0.5:
15                pos -= 1
16            else:
17                pos += 1
18            slots[pos]+=1
19    return slots
20
21 # a
22 num_balls = 10000
23 num_slots = 100
24 results = galton_board(num_balls, num_slots)
25 # print(results)
26
27 # Plot the results
28 names = list(results.keys())
29 values = list(results.values())
30
31 plt.bar(range(len(results)),values,tick_label=names)
32 plt.title(f"Galton Board Simulation {num_balls} balls,
33 {num_slots} slots")
34 plt.xlabel("Slot")
35 plt.ylabel("Number of balls")
36 plt.show()
```

```

37
38 # changes in distribution v/s number of levels
39 for i in [2,6,20,50]:
40     results = galton_board(num_balls, i)
41
42     names = list(results.keys())
43     values = list(results.values())
44
45     plt.bar(range(len(results)),values,tick_label=names)
46     plt.title(f"Galton Board Simulation ({num_balls} balls,
47     {i} levels)")
48     plt.xlabel("Slot")
49     plt.ylabel("Number of balls")
50     plt.show()
51
52 # part b
53 means, variances = [], []
54 for i in [2, 6, 20, 50]:
55     results = galton_board(num_balls, i)
56     # means.append(statistics.mean(results))
57     variances.append(statistics.variance(results))
58
59 # print(means)
60 print("The variances of the distributions are: ",variances)
61
62 # part c
63 print("The difference between the even and the odd L:")
64 print("the bars are distributed on even places for odd L")
65 print("The bars are distributed at odd places for even L")

```



Chapter 7

Expectation

7.1 What is expected value?

The expectation of a random variable is the long-term average of the random variable.

Expected value is the probability multiplied by the value of each outcome.

If each outcome of an experiment is equally likely to happen, for example rolling a dice, then probability distribution $f(x; n) = 1/6$, for $x = 1, 2, \dots, 6$.

$$\text{Mean value} = E(X) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$
$$\text{Variance: } \text{Var}(X) = \frac{(1 - 3.5)^2 + \dots + (6 - 3.5)^2}{6}$$
$$= \frac{35}{12}$$

A real example would be how much rain we should expect during the upcoming season, which is the expected value of the discrete random variable rain. Variance then is the departure from the expected mean. As a result, even though 700 mm of rain is what we should expect for the upcoming season, there will still be some departure from our expectations. Say the rain next season is 730 mm instead of the 700 mm we were expecting. The deviation from the mean is the 30 mm of rainfall. To keep the units for the mean and the variance the same and to describe this deviation, which is the square root of variance, we typically use the word standard deviation.

We can determine if you should play the lottery using this approach. Should you purchase a \$10 lottery ticket if there is a 0.0000001% chance that you will win \$10 million? This is a difficult question to answer without considering expected value. One of these tickets costs you \$10 but expected value of winning this lottery is very low.

7.1.1 Formal definitions:

The mean, expected value, or expectation of a random variable X is written as $E(X)$ or μ_X . If we observe N random values of X , then the mean of the N values will be approximately equal to $E(X)$ for large N .

The expectation is defined differently for continuous and discrete random variables.

Definition: Let X be a continuous random variable with pdf $f_X(x)$. The expected value of X is $E(X) = \int_{-\infty}^{\infty} xf(x)dx$ Definition: Let X be a discrete random variable with probability function $f_x(x)$. The expected value of X is $E(X) = \sum_x xf(x) = \sum_x xP(X = x)$

7.1.2 Example:

An investment in Project A will result in a loss of \$26,000 with probability 0.30, break even with probability 0.50, or result in a profit of \$68,000 with probability 0.20. An investment in Project B will result in a loss of \$71,000 with probability 0.20, break even with probability 0.65, or result in a profit of \$143,000 with probability 0.15. Which investment is better?

To calculate $E(X)$ - Project A

Random Variable (X) - The amount of money received from the investment in Project A

X can assume only x_1, x_2, x_3

$X = x_1$ is the event that we have Loss

$X = x_2$ is the event that we are breaking even

$X = x_3$ is the event that we have a Profit

- $x_1 = \$ - 26,000$

- $x_2 = \$0$

- $x_3 = \$68,000$

- $P(X = x_1) = 0.3$

- $P(X = x_2) = 0.5$

- $P(X = x_3) = 0.2$

Random Variable (X) - The amount of money received from the investment in Project B

• X can assume only x_1, x_2, x_3

$X = x_1$ is the event that we have Loss

$X = x_2$ is the event that we are breaking even

$X = x_3$ is the event that we have a Profit

- $x_1 = \$ - 71,000$

- $x_2 = \$0$

- $x_3 = \$143,000$

- $P(X = x_1) = 0.2$

- $P(X = x_2) = 0.65$

- $P(X = x_3) = 0.15$

$$\begin{array}{ll} \text{Project A :} & E(X) = 0.30 \cdot (-\$26,000) + 0.50 \cdot \$0 + 0.20 \cdot \$68,000 \\ & = \$5800 \end{array}$$

$$\begin{array}{ll} \text{Project B :} & E(X) = 0.20 \cdot (-\$71,000) + 0.65 \cdot \\ & = \$7250 \end{array}$$

7.1.3 Theorem 1

Let X be a random variable. Also let $g : R \rightarrow R$ be a $E[g(X)] = \sum_{i=1}^{\infty} g(x_i) P(X = x_i)$ whenever X is a discrete random variable, and $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$ whenever X is a continuous random variable.

7.1.4 The properties of $E(X)$ are as follows.

(i) $E(c) = c$, where c is a constant.

Proof $E(c) = cP(X = c) = c \times 1 = c$

(ii) $E(cX) = cE(X)$, c being a constant.

$$\begin{array}{ll} E(cX) = \sum_i cx_i P(X = x_i) = c \sum_i x_i P(X = x_i) & \text{For the continuous case,} \\ & = cE(X) \end{array}$$

$$\begin{array}{l} E(cX) = \int_{-\infty}^{\infty} cx f_X(x) dx = c \int_{-\infty}^{\infty} x f_X(x) dx \\ \text{case,} \\ = cE(X) \end{array}$$

$$\begin{array}{ll} E(aX + b) = \sum_i (ax_i + b) P(X = x_i) & \text{For the discrete case,} \\ & = a \sum_i x_i P(X = x_i) + b \sum_i P(X = x_i) \\ & = aE(X) + b \end{array}$$

$$\begin{array}{ll} E(aX + b) = \int_{-\infty}^{\infty} (ax + b) f_X(x) dx & \text{For the continuous case,} \\ & = a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} f_X(x) dx \\ & = aE(X) + b \end{array}$$

7.1.5 Expectation of Two-dimensional Random Variable

As in the one-dimensional case we now define the mathematical expectation of a two-dimensional random variable and generalise it for the n . dimensional case. The following is a two-dimensional analogue of Theorem 1.

7.1.5.1 Definition

Let (X, Y) be a two-dimensional random variable and $g : R^2 (= R \times R) \rightarrow R$ be a continuous function. Assume that $E[g(X, Y)]$ exists, Then $E[g(X, Y)]$ is defined as $E[g(X, Y)] = \Sigma g(x, y)f_{X,Y}(x, y)$ if (X, Y) is a discrete random variable, where the summation is over all possible values of (X, Y) ; and $E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dx dy$ for all $(x, y) \in R^2$, if (X, Y) is a continuous random variable, assuming that the series or the double integral converges absolutely.

7.1.5.2 Theorem 2

Let (X, Y) be a two-dimensional random variable, such that $E(X)$ and $E(Y)$ exist. Then $E(X + Y) =$

$$\begin{aligned} E(X + Y) &= \sum_i \sum_j (x_i + y_j) P(X = x_i, Y = y_j) \\ &= \sum_i x_i \left[\sum_j P(X = x_i, Y = y_j) \right] + \sum_j y_j \left[\sum_i P(X = x_i, Y = y_j) \right] \\ &= \sum_i x_i P(X = x_i) + \sum_j y_j P(Y = y_j) \\ &= E(X) + E(Y) \\ E(X + Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y)f_{X,Y}(x, y)dx dy \\ &= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \right] dx + \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} f_{X,Y}(x, y)dx \right] dy \\ \text{Proof: For the continuous case,} &= \int_{-\infty}^{\infty} xf_X(x)dx + \int_{-\infty}^{\infty} yf_Y(y)dy \\ &= E(X) + E(Y) \end{aligned}$$

This linear relation can be generalised for n random variables X_1, X_2, \dots, X_n . Thus, for a n -dimensional where $E(X_1), \dots, E(X_n)$ exist, we have $E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$ If a_1, \dots, a_n are constants, we also have $E(a_1X_1 + \dots + a_nX_n) = a_1E(X_1) + \dots + a_nE(X_n)$

7.1.5.3 Theorem 3

Let X and Y be two independent random variables, for which $E(X)$ and $E(Y)$ exist. Then $E(XY) = E(X)E(Y)$

$$\begin{aligned}
 E(XY) &= \sum_i \sum_j x_i y_j P(X = x_i, Y = y_j) \\
 &= \sum_i \sum_j x_i y_j P(X = x_i) P(Y = y_j) \quad \text{For the continuous case,} \\
 &= \sum_i x_i P(X = x_i) \sum_j y_j P(Y = y_j) \\
 &= E(X)E(Y)
 \end{aligned}$$

$$\begin{aligned}
 E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y) dx dy \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_X(x) f_Y(y) dx dy \\
 &= \int_{-\infty}^{\infty} x f_X(x) dx \int_{-\infty}^{\infty} y f_Y(y) dy = E(X)E(Y)
 \end{aligned}$$

Generalising for n independent random variables X_1, \dots, X_n where $E(X_1), \dots, E(X_n)$ exist, we have

$$E(X_1 X_2 \dots X_n) = E(X_1) E(X_2) \dots E(X_n)$$

7.1.6 Relation between Expectation and Moments

Moments of random variable X are the expectations of the powers of $X - a$, a being a real number. It plays an important role in mathematical statistics. Let r be a non-negative integer.

7.1.6.1 Definition

The moment of order r or the r th moment of X about a fixed point a is defined to be the mean value $E[(X - a)^r]$.

7.1.6.2 Definition

The r th absolute moment of X about a is defined to be the mean value $E[|X - a|^r]$.

7.1.6.3 Definition

The r th central moment of X , denoted by $\mu_r(X)$ or by μ_r is defined as

$$\mu_r = E[(X - E(X))^r] = E[(X - \mu)^r]$$

We have $\mu_0 = 1, \mu_1 = 0$ for all random variables. The central moments μ_r , can be expressed in terms of the moments about origin of the order sr as follows.

By binomial expansion, we have

$$(X - \mu)^r = \sum_{k=0}^r (-1)^{kr} C_k X^{r-k} \mu^k$$

So

i.e.

$$\begin{aligned} E[(X - \mu)^r] &= \sum_{k=0}^r (-1)^{kr} C_k E(X^{r-k}) \mu^k \\ \mu_r &= \sum_{k=0}^r (-1)^{kr} C_k \mu'_{r-k} \mu^k \end{aligned}$$

Since $\mu'_0 = 1$ and $\mu'_1 = \mu$, we have

Similarly,

$$\begin{aligned} \mu_2 &= \sum_{k=0}^2 (-1)^{k2} C_k \mu'_2 - k \mu^k = \mu'_2 - 2\mu'_1 \mu + \mu'_0 \mu^2 \\ &= \mu'_2 - 2\mu^2 + \mu^2 = \mu'_2 - \mu^2 \end{aligned}$$

Again

$$\mu_3 = \mu'_3 - 3\mu'_2 \mu + 2\mu^3$$

$$\mu_r(aX + b) = E[((aX + b) - E(aX + b))^r]$$

$$\begin{aligned} &= E[(aX + b - a\mu - b)^r] = a^r E[(X - \mu)^r] \text{ The following formulae will be helpful for cal-} \\ &= a^r \mu_r(X) \end{aligned}$$

$$\begin{aligned} E[(X - \mu)^2] &= E[X^2 - 2\mu X + \mu^2] = E(X^2) - 2\mu E(X) + \mu^2 \\ \text{calculating the variance.} &\qquad\qquad\qquad Var(X) = E(X^2) - (E(X))^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - (E(X))^2 \end{aligned}$$

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

We also see that

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

If X and Y are independent, $\text{Cov}(X, Y) = 0$.

7.2 Variance

7.2.1 Variance Definition:

Let X be a random variable with probability / density function $f(x)$ and expected value $E(X)$ is μ .

The variance of X is then given

$$\sigma^2 = \text{Var}(X) = E[(X - \mu)^2] = E[X^2 - 2\mu E(X) + \mu^2] = E[X^2] - \mu^2 = \sum_x (x - \mu)^2 f(x)$$

if X is discrete, and

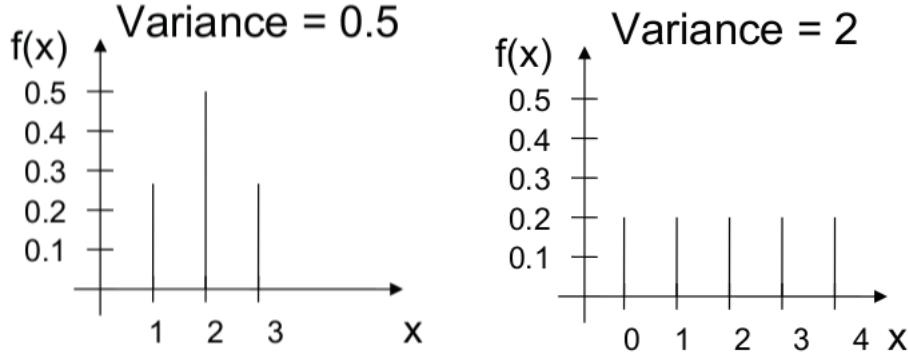
$$\sigma^2 = \text{Var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) d$$

if X is continuous.

The standard deviation is the positive root of the variance: $\sigma = \sqrt{\text{Var}(X)}$

7.2.2 Variance Interpretation

The variance expresses, how dispersed the density / probability function is around the mean.



Rewrite of the variance: $\sigma^2 = \text{Var}(X) = E[X^2] - \mu^2$

7.3 Expectation of $g(X)$

Assume that $g(X)$ is a function of X . Similar to how we may visualise the long-term average of X , we can also imagine the long-term average of $g(X)$. The formula for this average is $E(g(X))$. Consider making N observations of X to obtain the results x_1, x_2, \dots, x_N . To obtain $g(x_1), \dots, g(x_N)$, apply the function g to each of these observations (x_N). The mean of $g(x_1), g(x_2), \dots, g(x_N)$ approaches $E(g(X))$ as N increases to infinity in terms of the number of observations.

Definition: Let X be a continuous random variable, and let g be a function. The expected value

of $g(X)$ is

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Definition: Let X be a discrete random variable with probability function $f_X(x)$. The expected value of $g(X)$ is

$$E(X) = \sum_x g(x)f(x) = \sum_x g(x)P(X = x)$$

7.4 Mean / Expected value Function of a random variables

7.4.1 Definition:

Let X and Y be random variables with joint probability / density function $f(x, y)$. The expected value of $g(X, Y)$ is

$$\mu_{g(X,Y)} = E[g(X, Y)] = \sum_x \sum_y g(X, y)f(X, y)$$

if X and Y are discrete, and

$$\mu_{g(X,Y)} = E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)dxdy$$

if X and Y are continuous.

7.5 Mean / Expected value Function of two random variables

7.5.1 Problem:

Burger King sells both via "drive-in" and "walk-in".

Let X and Y be the fractions of the opening hours that "drive-in" and "walkin" are busy.

Assume that the joint density for X and Y are given by

$$f(x, y) = \begin{cases} 4xy & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The turn over $g(X, Y)$ on a single day is given by

$$g(X, Y) = 6000X + 9000Y$$

What is the expected turn over on a single day?

7.6 Expectation of XY : the definition of $\mathbb{E}(XY)$

Suppose we have two random variables, X and Y . These might be independent, in which case the value of X has no effect on the value of Y . Alternatively, X and Y might be dependent: when we observe a random value for X , it might influence the random values of Y that we are most likely to observe. For example, X might be the height of a randomly selected person, and Y might be the weight. On the whole, larger values of X will be associated with larger values of Y .

To understand what $\mathbb{E}(XY)$ means, think of observing a large number of pairs $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$. If X and Y are dependent, the value x_i might affect the value y_i , and vice versa, so we have to keep the observations together in their pairings. As the number of pairs N tends to infinity, the average $\frac{1}{N} \sum_{i=1}^N x_i \times y_i$ approaches the expectation $\mathbb{E}(XY)$.

For example, if X is height and Y is weight, $\mathbb{E}(XY)$ is the average of (height \times weight). We are interested in $\mathbb{E}(XY)$ because it is used for calculating the covariance and correlation, which are measures of how closely related X and Y are.

7.7 Properties of Expectation

- i) Let g and h be functions, and let a and b be constants. For any random variable X (discrete or continuous),

$$\mathbb{E}\{ag(X) + bh(X)\} = a\mathbb{E}\{g(X)\} + b\mathbb{E}\{h(X)\}$$

In particular,

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

- ii) Let X and Y be ANY random variables (discrete, continuous, independent, or non-independent).

Then

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

More generally, for ANY random variables X_1, \dots, X_n ,

$$\mathbb{E}(X_1 + \dots + X_n) = \mathbb{E}(X_1) + \dots + \mathbb{E}(X_n).$$

- iii) Let X and Y be independent random variables, and g, h be functions. Then

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

Notes: 1. $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ is ONLY generally true if X and Y are INDEPENDENT.

2. If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. However, the converse is not generally

true: it is possible for $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ even though X and Y are dependent.

7.8 Probability as an Expectation

Let A be any event. We can write $\mathbb{P}(A)$ as an expectation, as follows. Define the indicator function:

$$I_A = \begin{cases} 1 & \text{if event } A \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

Then I_A is a random variable, and

$$\begin{aligned} \mathbb{E}(I_A) &= \sum_{r=0}^1 r\mathbb{P}(I_A = r) \\ &= 0 \times \mathbb{P}(I_A = 0) + 1 \times \mathbb{P}(I_A = 1) \\ &= \mathbb{P}(I_A = 1) \\ &= \mathbb{P}(A) \end{aligned}$$

Thus

$$\mathbb{P}(A) = \mathbb{E}(I_A) \text{ for any event } A$$

7.9 Variance, covariance, and correlation

The variance of a random variable X is a measure of how spread out it is. Are the values of X clustered tightly around their mean, or can we commonly observe values of X a long way from the mean value? The variance measures how far the values of X are from their mean, on average.

Definition: Let X be any random variable. The variance of X is

$$\text{Var}(X) = \mathbb{E}\left((X - \mu_X)^2\right) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

The variance is the mean squared deviation of a random variable from its own mean.

If X has high variance, we can observe values of X a long way from the mean. If X has low variance, the values of X tend to be clustered tightly around the mean value.

Example: Let X be a continuous random variable with p.d.f.

$$f_X(x) = \begin{cases} 2x^{-2} & \text{for } 1 < x < 2 \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{E}(X)$ and $\text{Var}(X)$

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx = \int_1^2 x \times 2x^{-2} dx = \int_1^2 2x^{-1} dx \\
&= [2 \log(x)]_1^2 \\
&= 2 \log(2) - 2 \log(1) \\
&= 2 \log(2)
\end{aligned}$$

For $\text{Var}(X)$, we use

$$\text{Var}(X) = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2$$

Now

$$\begin{aligned}
E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_1^2 x^2 \times 2x^{-2} dx = \int_1^2 2dx \\
&= [2x]_1^2 \\
&= 2 \times 2 - 2 \times 1 \\
&= 2.
\end{aligned}$$

Thus

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 \\
&= 2 - \{2 \log(2)\}^2 \\
&= 0.0782
\end{aligned}$$

7.9.1 Covariance

Covariance is a measure of the association or dependence between two random variables X and Y . Covariance can be either positive or negative. (Variance is always positive.)

Definition: Let X and Y be any random variables. The covariance between X and Y is given by

$$\text{cov}(X, Y) = \mathbb{E}\{(X - \mu_X)(Y - \mu_Y)\} = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

where $\mu_X = \mathbb{E}(X), \mu_Y = \mathbb{E}(Y)$.

1. $\text{cov}(X, Y)$ will be positive if large values of X tend to occur with large values of Y , and small values of X tend to occur with small values of Y . For example, if X is height and Y is weight of a randomly selected person, we would expect $\text{cov}(X, Y)$ to be positive.
2. $\text{cov}(X, Y)$ will be negative if large values of X tend to occur with small values of Y , and small values of X tend to occur with large values of Y . For example, if X is age of a randomly selected person, and Y is heart rate, we would expect X and Y to be negatively correlated (older people have slower heart rates).

2. If X and Y are independent, then there is no pattern between large values of X and large values of Y , so $\text{cov}(X, Y) = 0$. However, $\text{cov}(X, Y) = 0$ does NOT imply that X and Y are independent, unless X and Y are Normally distributed.

7.9.2 Properties of Variance

- i) Let g be a function, and let a and b be constants. For any random variable X (discrete or continuous),

$$\text{Var}\{ag(X) + b\} = a^2 \text{Var}\{g(X)\}.$$

In particular, $\text{Var}(aX + b) = a^2 \text{Var}(X)$.

- ii) Let X and Y be independent random variables. Then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

- iii) If X and Y are NOT independent, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{cov}(X, Y).$$

Correlation (non-examinable)

The correlation coefficient of X and Y is a measure of the linear association between X and Y . It is given by the covariance, scaled by the overall variability in X and Y . As a result, the correlation coefficient is always between -1 and +1, so it is easily compared for different quantities.

Definition: The correlation between X and Y , also called the correlation coefficient, is given by

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

The correlation measures linear association between X and Y . It takes values only between -1 and +1, and has the same sign as the covariance.

The correlation is ± 1 if and only if there is a perfect linear relationship between X and Y , i.e. $\text{corr}(X, Y) = 1 \iff Y = aX + b$ for some constants a and b .

The correlation is 0 if X and Y are independent, but a correlation of 0 does not imply that X and Y are independent.

7.10 Conditional Expectation and Conditional Variance

Throughout this section, we will assume for simplicity that X and Y are discrete random variables. However, exactly the same results hold for continuous random variables too.

Suppose that X and Y are discrete random variables, possibly dependent on each other. Suppose

that we fix Y at the value y . This gives us a set of conditional probabilities $\mathbb{P}(X = x | Y = y)$ for all possible values x of X . This is called the conditional distribution of X , given that $Y = y$.

Definition: Let X and Y be discrete random variables. The conditional probability function of X , given that $Y = y$, is:

$$\mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(X = x \text{ AND } Y = y)}{\mathbb{P}(Y = y)}.$$

We write the conditional probability function as:

$$f_{X|Y}(x | y) = \mathbb{P}(X = x | Y = y).$$

Note: The conditional probabilities $f_{X|Y}(x | y)$ sum to one, just like any other probability function:

$$\sum_x \mathbb{P}(X = x | Y = y) = \sum_x \mathbb{P}_{\{Y=y\}}(X = x) = 1,$$

using the subscript notation $\mathbb{P}_{\{Y=y\}}$ of Section 2.3. We can also find the expectation and variance of X with respect to this conditional distribution. That is, if we know that the value of Y is fixed at y , then we can find the mean value of X given that Y takes the value y , and also the variance of X given that $Y = y$.

Definition: Let X and Y be discrete random variables. The conditional expectation of X , given that $Y = y$, is

$$\mu_{X|Y=y} = \mathbb{E}(X | Y = y) = \sum_x x f_{X|Y}(x | y)$$

$\mathbb{E}(X | Y = y)$ is the mean value of X , when Y is fixed at y .

7.10.1 Conditional expectation as a random variable

The unconditional expectation of X , $\mathbb{E}(X)$, is just a number: e.g. $\mathbb{E}X = 2$ or $\mathbb{E}X = 5.8$.

The conditional expectation, $\mathbb{E}(X | Y = y)$, is a number depending on y .

If Y has an influence on the value of X , then Y will have an influence on the average value of X . So, for example, we would expect $\mathbb{E}(X | Y = 2)$ to be different from $\mathbb{E}(X | Y = 3)$.

We can therefore view $\mathbb{E}(X | Y = y)$ as a function of y , say $\mathbb{E}(X | Y = y) = h(y)$. To evaluate this function, $h(y) = \mathbb{E}(X | Y = y)$, we:

- i) fix Y at the chosen value y ;
- ii) find the expectation of X when Y is fixed at this value. However, we could also evaluate the function at a random value of Y :
- i) observe a random value of Y ;

- ii) fix Y at that observed random value;
- iii) evaluate $\mathbb{E}(X | Y = \text{observed random value})$.

We obtain a random variable: $\mathbb{E}(X | Y) = h(Y)$.

The randomness comes from the randomness in Y , not in X .

Conditional expectation, $\mathbb{E}(X | Y)$, is a random variable with randomness inherited from Y , not X .

Example: Suppose $Y = \begin{cases} 1 & \text{with probability } 1/8 \\ 2 & \text{with probability } 7/8 \end{cases}$

and $X | Y = \begin{cases} 2Y & \text{with probability } 3/4 \\ 3Y & \text{with probability } 1/4 \end{cases}$

Conditional expectation of X given $Y = y$ is a number depending on y :

If $Y = 1$, then: $X | (Y = 1) = \begin{cases} 2 & \text{with probability } 3/4 \\ 3 & \text{with probability } 1/4 \end{cases}$

$$\text{so } \mathbb{E}(X | Y = 1) = 2 \times \frac{3}{4} + 3 \times \frac{1}{4} = \frac{9}{4}.$$

If $Y = 2$, then: $X | (Y = 2) = \begin{cases} 4 & \text{with probability } 3/4 \\ 6 & \text{with probability } 1/4 \end{cases}$

$$\text{so } \mathbb{E}(X | Y = 2) = 4 \times \frac{3}{4} + 6 \times \frac{1}{4} = \frac{18}{4}.$$

Thus $\mathbb{E}(X | Y = y) = \begin{cases} 9/4 & \text{if } y = 1 \\ 18/4 & \text{if } y = 2 \end{cases}$

So $\mathbb{E}(X | Y = y)$ is a number depending on y , or a function of y .

7.10.2 Conditional expectation of X given random Y is a random variable:

From above, $\mathbb{E}(X | Y) = \begin{cases} 9/4 & \text{if } Y = 1 (\text{probability } 1/8) \\ 18/4 & \text{if } Y = 2 (\text{probability } 7/8) \end{cases}$

So $\mathbb{E}(X | Y) = \begin{cases} 9/4 & \text{with probability } 1/8 \\ 18/4 & \text{with probability } 7/8 \end{cases}$

Thus $\mathbb{E}(X | Y)$ is a random variable.

The randomness in $\mathbb{E}(X | Y)$ is inherited from Y , not from X .

Conditional expectation is a very useful tool for finding the unconditional expectation of X (see below). Just like the Partition Theorem, it is useful because it is often easier to specify conditional probabilities than to specify overall probabilities.

7.10.3 Conditional variance

The conditional variance is similar to the conditional expectation.

- $\text{Var}(X | Y = y)$ is the variance of X , when Y is fixed at the value $Y = y$.
- $\text{Var}(X | Y)$ is a random variable, giving the variance of X when Y is fixed at a value to be selected randomly.

Definition: Let X and Y be random variables. The conditional variance of X , given Y , is given by

$$\text{Var}(X | Y) = \mathbb{E}(X^2 | Y) - \{\mathbb{E}(X | Y)\}^2 = \mathbb{E}\{(X - \mu_{X|Y})^2 | Y\}$$

Like expectation, $\text{Var}(X | Y = y)$ is a number depending on y (a function of y), while $\text{Var}(X | Y)$ is a random variable with randomness inherited from Y .

7.11 Risk and Return: Example of expectations and Variance

7.11.1 Expected Return

- The future is uncertain.
- Investors do not know with certainty whether the economy will be growing rapidly or be in recession.
- Investors do not know what rate of return their investments will yield.
- Therefore, they base their decisions on their expectations concerning the future.
- The expected rate of return on a stock represents the mean of a probability distribution of possible future returns on the stock.
- The table below provides a probability distribution for the returns on stocks A and B

State	Probability	Return On Stock A	Return On Stock B
1	20%	5%	50%
2	30%	10%	30%
3	30%	15%	10%
4	20%	20%	-10%

- The state represents the state of the economy one period in the future i.e. state 1 could represent a recession and state 2 a growth economy.
- The probability reflects how likely it is that the state will occur.

- The sum of the probabilities must equal 100%.
- The last two columns present the returns or outcomes for stocks A and B that will occur in each of the four states.
- Given a probability distribution of returns, the expected return can be calculated using the following equation:

$$E[R] = \sum_{i=1}^N p_i R_i$$

Where:

$E[R]$ = the expected return on the stock

N = the number of states

p_i = the probability of state i

R_i = the return on the stock in state i .

- In this example, the expected return for stock A would be calculated as follows:

$$E[R]_A = .2(5\%) + .3(10\%) + .3(15\%) + .2(20\%) = 12.5\%$$

$$E[R]_B = .2(50\%) + .3(30\%) + .3(10\%) + .2(-10\%) = 20\%$$

So we see that Stock B offers a higher expected return than Stock A. However, that is only part of the story; we haven't considered risk.

7.11.2 Measures of Risk

- Risk reflects the chance that the actual return on an investment may be different than the expected return.
- One way to measure risk is to calculate the variance and standard deviation of the distribution of returns.
- Probability Distribution:

State	Probability	Return On Stock A	Return On Stock B
1	20%	5%	50%
2	30%	10%	30%
3	30%	15%	10%
4	20%	20%	-10%

- $E[R]_A = 12.5\%$

- $E[R]_B = 20\%$

- Given an asset's expected return, its variance can be calculated using the following equation:

$$Var(R) = \sum_{i=1}^N (p_i(R_i - E[R])^2$$

Where:

N = the number of states

p_i = the probability of state i

R_i = the return on the stock in state i

$E[R]$ = the expected return on the stock

- The standard deviation is calculated as the positive square root of the variance:

$$SD(R) = \sigma = \sqrt{\sigma^2}$$

- The variance and standard deviation for stock A is calculated as follows:

$$\sigma_A^2 = .2(.05 - .125)^2 + .3(.1 - .125)^2 + .3(.15 - .125)^2 + .2(.2 - .125)^2 = .002625$$

$$\sigma_A = (.002625)^{0.5} = .0512 = 5.12\%$$

$$\sigma_B^2 = .2(.50 - .20)^2 + .3(.30 - .20)^2 + .3(.10 - .20)^2 + .2(-.10 - .20)^2 = .042$$

$$\sigma_B = (.042)^{0.5} = .2049 = 20.49\%$$

- Although Stock B offers a higher expected return than Stock A, it also is riskier since its variance and standard deviation are greater than Stock A's.
- This, however, is still only part of the picture because most investors choose to hold insecurities as part of a diversified portfolio.

7.11.3 Portfolio Risk and Return

- Most investors do not hold stocks in isolation.
- Instead, they choose to hold a portfolio of several stocks.
- When this is the case, a portion of an individual stock's risk can be eliminated, i.e., diversified away.
- From our previous calculations, we know that:
 - the expected return on Stock A is 12.5%
 - the expected return on Stock B is 20%

- the variance on Stock A is .00263
- the variance on Stock B is .04200
- the standard deviation on Stock A is 5.12%
- the standard deviation on Stock B is 20.49%
- The Expected Return on a Portfolio is computed as the weighted average of the expected returns on the stocks which comprise the portfolio.
- The weights reflect the proportion of the portfolio invested in the stocks.
- This can be expressed as follows:

$$E[R_p] = \sum_{i=1}^N (w_i E[R_i])$$

Where:

- $E[R_p]$ = the expected return on the portfolio
- N = the number of stocks in the portfolio
- w_i = the proportion of the portfolio invested in stock i
- $E[R_i]$ = the expected return on stock i

- For a portfolio consisting of two assets, the above equation can be expressed as:

$$E[R_p] = w_1 E[R_1] + w_2 E[R_2]$$

If we have an equally weighted portfolio of stock A and stock B (50% in each stock), then the expected return of the portfolio is:

$$E[R_p] = .50(.125) + .50(.20) = 16.25\%$$

- The variance/standard deviation of a portfolio reflects not only the variance/standard deviation of the stocks that make up the portfolio but also how the returns on the stocks which comprise the portfolio vary together.
- Two measures of how the returns on a pair of stocks vary together are the covariance and the correlation coefficient.
 - Covariance is a measure that combines the variance of a stock's returns with the tendency of those returns to move up or down at the same time other stocks move up or down.

- Since it is difficult to interpret the magnitude of the covariance terms, a related statistic, the correlation coefficient, is often used to measure the degree of co-movement between two variables. The correlation coefficient simply standardizes the covariance.
- The Covariance between the returns on two stocks can be calculated as follows:

$$Cov(R_A, R_B) = \sum_{i=1}^N p_i(R_{Ai} - E[R_A])(R_{Bi} - E[R_B])$$

Where:

$Cov(R_A, R_B)$ = the covariance between the returns on stocks A and B

N = the number of states

p_i = the probability of state i

R_{Ai} = the return on stock A in state i

$E[R_A]$ = the expected return on stock A

R_{Bi} = the return on stock B in state i

$E[R_B]$ = the expected return on stock B

- The Correlation Coefficient between the returns on two stocks can be calculated as follows:

$$\rho_{A,B} = Corr(R_A, R_B) = \frac{Cov(R_A, R_B)}{SD(R_A)SD(R_B)}$$

- The covariance between stock A and stock B is as follows:

$$Cov(R_A, R_B) = .2(.05 - .125)(.5 - .2) + .3(.1 - .125)(.3 - .2) + .3(.15 - .125)(.1 - .2) + .2(.2 - .125)(-.1 - .2) = -.0105$$

The correlation coefficient between stock A and stock B is as follows:

$$\rho_{A,B} = \frac{-0.0105}{(.0512)(.2049)} = -1.00$$

- Using either the correlation coefficient or the covariance, the Variance on a Two-Asset Portfolio can be calculated as follows:

$$s_p^2 = (w_A)^2 s_A^2 + (w_B)^2 s_B^2 + 2w_A w_B \rho_{A,B} \sigma_A \sigma_B$$

OR

$$s_p^2 = (w_A)^2 s_A^2 + (w_B)^2 s_B^2 + 2w_A w_B Cov(R_A, R_B)$$

- The Standard Deviation of the Portfolio equals the positive square root of the variance.

- Let's calculate the variance and standard deviation of a portfolio comprised of 75

$$\sigma_p^2 = (.75)^2(.0512)^2 + (.25)^2(.2049)^2 + 2(.75)(.25)(-1)(.0512)(.2049) = .00016$$

$$\sigma_p = \sqrt{.00016} = .0128 = 1.28\%$$

- Notice that the portfolio formed by investing 75% in Stock A and 25% in Stock B has a lower variance and standard deviation than either Stocks A or B and the portfolio has a higher expected return than Stock A.
- This is the purpose of diversification; by forming portfolios, some of the risk inherent in the individual stocks can be eliminated.
- Risk of returns by standard deviation (variance)
- Dependence of returns by covariance (linear correlation)

7.12 Midterm question paper with solutions

1. Let A and B be two events with $P(A) > 0$, $P(B | A) = 0.3$, and $P(A \cap B^c) = 0.2$. Then find $P(A)$.

Answer: $P(B | A) = 0.3 \Rightarrow P(A \cap B) = 0.3P(A)$. And we know that $P(A) = P(A \cap B) + P(A \cap B^c) \Rightarrow P(A) = 0.3P(A) + 0.2$. So, $P(A) = 2/7 = 0.28$

2. A system with m components functions if and only if at least one of m components functions. Suppose all the m components of the system functions independently, each with probability $3/4$. If the probability of functioning of the system is $63/64$, then find the value of m .

Answer: $P(\text{at least one component is working}) = 63/64$. Now,

$$1 - P(\text{no component is working}) = 63/64. \text{ So, } 1 - (1/4)^m = 63/64$$

$$\text{We get } 1 - 63/64 = (1/4)^m \Rightarrow 1/64 = (1/4)^m \Rightarrow (1/4)^3 = (1/4)^m$$

. So, $m=3$

3. Consider a telephone operator who, on average, handles five calls every three minutes. What is the probability there will be no call in the next minute?

Answer: Use Poisson distribution to solve this problem. Five calls every three meets so rate is $5/3$. $P(\text{no call in the next minute}) = P(X = 0) = \frac{e^{-5/3}(5/3)^0}{0!} = e^{-5/3}$
 $= 0.189$

4. Let X be the random variable with density function $f(x) = 7e^{-7x}, 0 < x < \infty$. Let $Y = 4X + 3$, then find the density function of Y.

Answer: We know that $F(y) = P(Y \leq y) = P(4X + 3 \leq y) = P\left(X \leq \frac{y-3}{4}\right) = F_X\left(\frac{y-3}{4}\right)$. Now taking derivative on both sides we get, $f(y) = \frac{1}{4}f_x\left(\frac{y-3}{4}\right)$ So, $f(y) = \frac{7}{4}e^{-7(y-3)/4}; 3 < y < \infty$

5. Consider the discrete bivariate random vector with the joint probability mass function is given by $f(10, 1) = f(20, 1) = f(20, 2) = 1/10$, $f(10, 2) = f(10, 3) = 1/5$, and $f(20, 3) = 3/10$. Show that random variables X and Y are not independent.

Answer: $f(10, 3) \neq f_X(10)f_Y(3)$

6. Let the joint probability mass function of (X, Y) is given by $f(0, 10) = f(0, 20) = 2/18$, $f(1, 10) = f(1, 30) = 3/18$, $f(1, 20) = 4/18$, and $f(2, 30) = 4/18$ and zero elsewhere. Then find $f(y = 10 | x = 0)$.

Answer: $f(y = 10 | x = 0) = \frac{f(x=0,y=10)}{f_X(x=0)}$. So, $f_X(x = 0) = f(x = 0, y = 10) + f(x = 0, y = 20) + f(x = 0, y = 30) = 4/18$

And $f(x = 0, y = 10) = 2/18$. Hence, we get $f(y = 10 | x = 0) = \frac{2/18}{4/18} = 1/2$.

7. Let \mathcal{A} be the partitions of the sample space, and let B be any set. Then, for each $i = 1, 2, 3, \dots$ write the Bayes' rule.

Answer: $P(A_i | B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}$

8. Write the necessary and sufficient condition for cumulative distribution function.

- Answer:** a) $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$
 b) $F(x)$ is a nondecreasing function of x .
 c) $F(x)$ is right continuous.

9. Suppose we are experiencing an extreme heatwave and there are two weather phenomena that can occur: Heatstroke (H) and Dehydration (D). Let's assume that the probability of experiencing a Heatstroke is 80% and the probability of experiencing Dehydration is 65%. We also know that the probability of experiencing both Heatstroke and Dehydration is 60%. If someone is experiencing Heatstroke, what is the probability that they are also experiencing Dehydration?

Answer: We know that $P(H) = 0.8$, $P(D) = 0.65$, and $P(H \text{ and } D) = 0.60$

$$\text{And } P(D | H) = P(H \text{ and } D)/P(H)$$

Substituting the given probabilities, we get: $P(D | H) = 0.6/0.8 = 0.75$

10. Suppose we have a team of 10 scientists who are conducting research in Antarctica during an extreme cold wave. The team needs to select two scientists from their group to be the lead researcher and the assistant lead researcher. How many ways can this be done?

Answer: Let n_1 = the number of ways the lead researcher can be chosen = 10. Let n_2 = the number of ways the assistant lead researcher can be chosen once the chair has been chosen = 9. Then $N = n_1 * n_2 = (10)(9) = 90$

11. What is the probability that a committee of 10 people chosen from a group consisting of 40 principals, 35 teachers, and 25 students, will include three principals, five teachers, and two students?

Answer: Let X be the event that a committee of 10 people chosen from a group consisting of 40 principals, 35 teachers, and 25 students, will include three principals, five teachers, and two students. Here, X follows hypergeometric distribution and its probability can be given by:

$$P(X) = \frac{(40C3)(35C5)(25C2)}{(100C10)} = 0.0556$$

This means there is a 0.0556 chance that precisely 3 principals, five teachers, and two students will be chosen for the committee.

12. A catalyst producer produces a device for testing defects in a certain electrocatalyst (EC). The catalyst producer claims that the test is 97% reliable if the EC is defective and 99% reliable when it is flawless. However, 4% of said EC may be expected to be defective upon delivery. What is the probability that EC found flawless given that it is tested defective?

Answer: Let A : EC is defective; \bar{A} : the EC is flawless;

B : the EC is tested to be defective; \bar{B} : the EC is tested to be flawless.

The probabilities would be

B/A : EC is (known to be) defective, and tested defective, $P(B/A) = 0.97$

\bar{B}/A : EC is (known to be) defective, but tested flawless, $P(\bar{B}/A) = 1 - P(B/A) = 0.03$,

B/\bar{A} : EC is (known to be) defective, but tested defective, $P(B/\bar{A}) = 1 - P(\bar{B}/\bar{A}) = 0.01$

$P(\text{EC found flawless given that it is tested defective})$

$$= P(\bar{A}/B) = \frac{P(B/\bar{A})P(\bar{A})}{P(B/\bar{A})P(\bar{A}) + P(B/A)P(A)} = \frac{0.01 * 0.96}{0.01 * 0.96 + 0.97 * 0.04} = 0.1983$$

13. Suppose that 5 people, including you and a friend, line up at random. Let the random variable X denote the number of people standing between you and a friend. Determine the probability mass function of X in tabular form. Also, verify that the p.m.f. is a valid p.m.f.

Answer: Given, X denotes the number of people standing between you and a friend.

Then, $X = 0, 1, 2, 3$

$$P(X = 0) = \frac{4 * 2! * 3!}{5!} = 4/10 = 2/5$$

(where 4 is no. of positions you and your friend can stand, $2!$ is no. of ways you and your friend can be arranged, $3!$ is no. of ways remaining people will be arranged.)

$$P(X = 1) = \frac{3 * 2! * 3!}{5!} = 3/10$$

$$P(X = 2) = \frac{2 * 2! * 3!}{5!} = 2/10 = 1/5$$

$$P(X = 3) = \frac{1 * 2! * 3!}{5!} = 1/10$$

Now

$$P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 2/5 + 3/10 + 1/5 + 1/10 = 1$$

So, p.m.f is valid p.m.f.

14. Let Jim and his fiance stay together. Let

C = Event that Jim had Covid

T = Event that Jim tests positive

F = Event that his fiance had Covid

$$P(C) = 0.1, \quad P(T | \bar{C}) = 0.005, \quad P(\bar{T} | C) = 0, \quad P(F | \bar{C}) = 0, \quad P(F | C) = 0.95$$

Find the probability that Jim had Covid given that Jim tests positive and his fiance had not covid.

Answer: P(Jim had Covid given that Jim tests positive and his fiance had not covid)

$$\begin{aligned} P(C | T, \bar{F}) &= \frac{P(C, T, \bar{F})}{P(C, T, \bar{F}) + P(\bar{C}, T, \bar{F})} \\ &= \frac{0.1 * 1 * 0.05}{0.1 * 1 * 0.05 + 0.9 * 1 * 0.005} \\ &= \frac{0.005}{0.005 + 0.0045} = 52.6\% \end{aligned}$$

There is 52.6% that Jim had Covid if he tests positive and his fiance did not have covid.

15. How many different ways can the letters of the word TRIANGLE be arranged a) If the order of the vowels IAE cannot be changed, though their placement may (IAETRNGL and TRIANGEL are acceptable but EIATRNGL and TRIENGLA are not)? b) If the order of the vowels IAE can be changed, though their placement may not?

Answer: a) Step one is to choose the places that the vowels go. Here we are picking three places out of eight, and the order that we do this is not important. This is a combination and there are a total of $8C3 = 56$ ways to perform this step. The remaining five letters may be arranged in $5! = 120$ ways. This gives a total of $56 \times 120 = 6720$ arrangements.

b) We arrange three letters in $3! = 6$ ways and the other five letters in $5! = 120$ ways. The total number of ways for this arrangement is $6 \times 120 = 720$.

16. The joint probability of random variable X, Y is given below (a) Find $f_x(x)$.

$$f_{x,y}(x, y) = \begin{cases} xye^{-\frac{(x^2+y^2)}{2}} & x \geq 0, y \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Find $f_x(x)$

(b) Find $f_{Y/X}(y/x)$

Answer: Solution 1: (a) $f_x(x) = \int_{-\infty}^{\infty} f_{x,y}(x, y) dy$

$$\begin{aligned} &= \int_0^{\infty} xye^{-\frac{(x^2+y^2)}{2}} dy \\ &= x^{-x^2/2} \int_0^{\infty} ye^{-\frac{y^2}{2}} dy \\ &= xe^{-x^2/2} \left(\int_0^{\infty} ye^{-\frac{y^2}{2}} dy = 1 \right) \end{aligned}$$

(b) $f_{Y/X}(y/x) = \frac{f_{x,y}(x,y)}{f_x(x)}$

$$\begin{aligned}
&= \frac{xye^{-\frac{(x^2+y^2)}{2}}}{xe - (x^2/2)} \\
&= ye^{-y^2/2}
\end{aligned}$$

17. The two random variables X and Y have a joint CDF.

$$F_{x,y}(x, y) = \begin{cases} (1 - \frac{1}{x^2}) (1 - \frac{1}{y^2}) & x \geq 1, y \geq 1 \\ 0 & \text{elsewhere} \end{cases}$$

- (a) Find the marginal CDF of X and Y.
- (b) Are X and Y independent?
- (c) Find the probability for $\{X \leq 5, Y \leq 5\}$.
- (d) Using the results obtained in above parts, find probability for $\{X > 3, Y > 3\}$

Answer: (a) $F_X(x) = F_{X,Y}(x, \infty)$

$$= (1 - 1/x^2)$$

(b) $F_Y(y) = F_{X,Y}(\infty, y)$

$$= (1 - 1/y^2)$$

(c) To prove random variable X and Y independent we need to show

$$\begin{aligned}
F_{X,Y}(x, y) &= F_X(x)F_Y(y) \\
F_{X,Y}(x, y) &= (1 - 1/x^2)(1 - 1/y^2) \\
F_X(x) &= (1 - 1/x^2) \\
F_Y(y) &= (1 - 1/y^2)
\end{aligned}$$

Which is true in this case so X and Y are independent

- (d) $P(x \leq 5, y \leq 5) = F_{x,y}(5, 5) = (1 - 1/5^2)(1 - 1/5^2) = 0.9216$
- (e) $P(x > 3, y > 3) = 1 - F_{x,y}(3, 3) = 1 - (1 - 1/3^2)(1 - 1/3^2) = 0.209$

18. Let Y be a random variable having the density function as

$$f(y, y_0, \beta) = \frac{\beta y_0^\beta}{y^{(\beta+1)}} \text{ if } y > y_0.$$

Where $\beta > 0, y_0 > 0$. If $X = \log\left(\frac{Y}{y_0}\right)$, then find range of X, density function of X

i.e., $f(x)$, $P(X > 3)$, and $P(X = 3)$.

Answer: (a) We are given that $X = \log\left(\frac{Y}{y_0}\right) \Rightarrow Y = y_0 e^X$. Since

$Y > 0 \Rightarrow y_0 e^X > 0$. So, $X > 0$

(b) $F_X(x) = P(X \leq x) = P\left(\log\left(\frac{Y}{y_0}\right) \leq x\right) = P(Y \leq y_0 e^x) = F_Y(y_0 e^x)$. Now

taking derivative with respect to x on both sides we get

$$f_X(x) = y_0 e^x f_Y(y_0 e^x) \Rightarrow f_X(x) = \beta e^{-\beta x}$$

$$(c) P(X > 3) = 1 - P(X \leq 3) = 1 - \int_0^3 \beta e^{-\beta x} dx = e^{-3\beta}$$

(d) $P(X = 3) = 0$, probability at a single point is zero for continuous distribution.

7.13 Practice Problems

Question-1: Suppose that n students are selected at random without replacement from a class containing T students, of whom A are boys and $T - A$ are girls. Let X denote the number of boys that are obtained. Find the sample size n such that the variance of X is maximum.

Solution-1: Given that n students are selected at random without replacement from a class of T students, of whom A are boys and $T - A$ are girls. The random variable X denotes the number of boys obtained. Now, we have to find a sample size n such that the variance is maximum.

Here, the core is that n students are drawn without replacement out of T students, of which A are boys and $T - A$ are girls. The number of boys is our success rate. Hence, the random variable X takes on the hyper-geometric distribution with parameters A , $T - A$, and n .

The variance of X is given as:

$$\begin{aligned} \text{Var}(X) &= n * \frac{A}{T} * \left(1 - \frac{A}{T}\right) * \frac{T-n}{T-1} \\ &= \frac{nA(T-A)(T-n)}{T^2(T-1)} \\ &= \frac{A(T-A)(nT-n^2)}{T^2(T-1)} \end{aligned}$$

To get the value of n such that the variance is maximum is obtained by differentiating $\text{Var}(X)$ wrt. n and equating it to 0.

$$\begin{aligned} \frac{A(T-A)(T-2n)}{T^2(T-1)} &= 0 \\ \Rightarrow n &= \frac{T}{2} \end{aligned}$$

Since, n can be integer only, the required n value will be $\frac{T}{2}$ if T is an even number and if T is odd, then the required n value will be $\frac{T-1}{2}$ and $\frac{T+1}{2}$.

Question-2: A new battery supposedly with a charge of 1.5 volts actually has a voltage with a uniform distribution between 1.43 and 1.60 volts. Then,

- (a) What is the expectation and the standard deviation of the voltage?
- (b) What is its cumulative distribution function?
- (c) If a box contains 50 batteries, what is the expectation and variance of the number of batteries in the box with a voltage less than 1.5 volts?

Solution-2: Given that a new battery supposedly with a charge of 1.5 volts actually has a voltage with a uniform distribution between 1.43 and 1.60 volts. So the random variable, X , denoting the voltage of the battery is $X \sim \text{Uni}(1.43, 1.60)$.

- (a) The expectation of the voltage is given as

$$E[X] = \frac{\text{sum of limits}}{2} = \frac{1.43 + 1.60}{2} = 1.515$$

and the standard deviation of the voltage

$$\begin{aligned}\sigma &= \sqrt{\frac{(\text{difference between the limits})^2}{12}} \\ &= \frac{1.60 - 1.43}{\sqrt{12}} = 0.0491\end{aligned}$$

- (b) Its cumulative distribution function is given as

$$\begin{aligned}F(x) &= \int_{1.43}^x \frac{1}{1.60 - 1.43} dx \\ &= \frac{x - 1.43}{1.60 - 1.43} \\ &= \frac{x - 1.43}{0.17} \text{ for } 1.43 \leq x \leq 1.60\end{aligned}$$

(c) Given that a box contains 50 batteries. Let X be the random variable denoting the number of batteries with a voltage less than 1.5 volts. Then, X would take on the Binomial distribution with parameters $n = 50$ and a success probability equals the probability that a battery will have a voltage less than 1.5 volts.

The probability that a battery will have a voltage less than 1.5 volts is given as $F(1.5) = \frac{1.5 - 1.43}{0.17} = \frac{0.07}{0.17} = 0.412$. Hence, the random variable $X \sim \text{Bin}(50, 0.412)$. So, the expected value is $E[X] = np = 50 * 0.412 = 20.6$ and the variance of X is $\text{Var}(X) = np(1 - p) = 50 * 0.412 * 0.588 = 12.11$.

Question-3: Suppose we observe a random variable X and then based on it, we make an attempt to predict the value of another random variable Y . Let $g(X)$ be the predictor of Y . Then, find the best possible predictor of Y .

Solution-3: Given that we observe a random variable X and then based on it, we make an attempt to predict the value of another random variable Y and that predictor of Y is $g(X)$.

Now, we need to find the best possible predictor of Y , i.e., we need to find $g(X)$ that tends close to Y . One way to do this is to choose g which minimizes $E[(Y - g(X))^2]$.

Consider, $E[(Y - g(X))^2 | X]$

Adding and subtracting $E[Y | X]$, we get

$$\begin{aligned} &= E[(Y - E[Y | X] + E[Y | X] - g(X))^2 | X] \\ &= E[(Y - E[Y | X])^2 | X] + E[(E[Y | X] - g(X))^2 | X] + 2E[(Y - E[Y | X])(E[Y | X] - g(X)) | X] \end{aligned}$$

Here, note that given X , $E[Y | X] - g(X)$, being a function of X , is a constant. Thus, we can write

$$\begin{aligned} &= E[(Y - E[Y | X])(E[Y | X] - g(X)) | X] \\ &= (E[Y | X] - g(X))E[(Y - E[Y | X]) | X] \\ &= (E[Y | X] - g(X))(E[(Y | X) - E[Y | X]]) \\ &= 0 \end{aligned}$$

Substituting this in the above main equation, we get

$$E[(Y - g(X))^2 | X] \geq E[(Y - E[Y | X])^2 | X]$$

Now, since given X , all the above are just constants, on applying expectation to it once again we get the following,

$$E[(Y - g(X))^2] \geq E[(Y - E[Y | X])^2].$$

Question-4: Passengers try to get a seat reservation in a train running between two stations. They try until they are successful. If there the chance of getting reservation in an attempt by a passenger is 30% then what is the average number of attempts that passengers need to get a seat reserved is?

Solution-4: This problem can be modeled using a geometric distribution, where the probability of success (getting a seat reservation) in one attempt is $p = 0.3$.

The expected number of attempts needed to get a seat reservation can be calculated as:

$$E(X) = 1/p.$$

Substituting $p = 0.3$, we get: $E(X) = 1/0.3 = 3.33$ Therefore, on average, a passenger needs to make about 3.33 attempts to get a seat reservation on the train.

Question-5: Your probability class has 300 students and each student has probability $1/3$ of getting an A , independently of any other student. What is the mean of X , the number of students that get an A ?

Solution-5: Let $X_i = 1$ if the i^{th} student gets an A , 0 otherwise.

Thus X_1, X_2, \dots, X_n are Bernoulli random variables with common mean $p = 1/3$ and variance $p(1 - p) = (1/3)(2/3) = 2/9$.

Their sum $X = X_1 + X_2 + \dots + X_n$ is the number of students that get an A . Since X is the number of "successes" in n independent trials, it is a binomial random variable with parameters n and p .

Using the linearity of X as a function of the X_i , we have $E[X] = X300i = 1$.

$$E[X] = (1/3 + 1/3 + 1/3 + \dots + 1/3) = 300/3 = 100.$$

Question-6: Riya passes through four traffic lights on her way to work, and each light is equally likely to be green or red, independently of the others.

(a) What is the PMF, the mean, and the variance of the number of red lights that she encounters?

(b) Suppose that each red light delays Riya by exactly two minutes. What is the variance of Riya's commuting time?

Solution-6: (a) Let X be the number of red lights that Riya encounters. The PMF of X is binomial with $n = 4$ and $p = 1/2$. The mean and the variance of X are $E[X] = np = 2$ and $\text{var}(X) = np(1 - p) = 4 \cdot (1/2) \cdot (1/2) = 1$.

(b) The variance of Riya's commuting time is the same as the variance of the time by which Riya is delayed by the red lights. This is equal to the variance of $2X$, which is $4 \text{ var}(X) = 4$.

Question-7: Suppose that the number of visitors to a website follows a Poisson distribution with a mean of 10 visitors per hour. What is the probability that the website will have at least one visitor in the next 5 minutes?

Solution-7: We can approach this problem by first finding the average number of visitors in 5 minutes since the Poisson distribution is based on the number of events per unit of time. There are 12 five-minute intervals in an hour, so the average number of visitors in a 5-minute interval is:

$$10 \text{ visitors/hour} / 12 \text{ intervals/hour} = 0.833 \text{ visitors/interval}$$

We can now use the Poisson distribution formula to calculate the probability of at least one visitor in the next 5 -minute interval:

$P(X \geq 1) = 1 - P(X = 0)$, where X is the random variable representing the number of visitors in a 5 -minute interval.

The Poisson probability mass function is given by:

$P(X = k) = e^{-\lambda} \lambda^k / k!$ where λ is the average number of visitors in a 5 -minute interval, which we calculated to be 0.833.

$$\text{Therefore, } P(X = 0) = e^{-0.833} 0.833^0 / 0! \approx 0.434.$$

So, the probability of at least one visitor in the next 5 -minute interval is:

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - 0.434 = 0.566.$$

Therefore, there is a 56.6% chance that the website will have at least one visitor in the next 5 minutes.

Question-8: Suppose that a manufacturing company produces electronic components, and the proba-

bility that a component is defective is 0.05. The company performs quality checks on each component, and if a component is found to be defective, it is discarded. The company continues to perform quality checks until the first non-defective component is found. Let X be the number of defective components that are discarded before the first non-defective component is found. What is the probability that at least two defective components are discarded before the first non-defective component is found?

Solution-8: The number of defective components that are discarded before the first non-defective component is found follows a geometric distribution with parameter $p = 0.05$. That is, the probability that $X = k$ is given by:

$$P(X = k) = (1 - p)^k * p, \text{ where } k = 0, 1, 2, \dots$$

To find the probability that at least two defective components are discarded before the first non-defective component is found, we need to sum the probabilities for $k = 2, 3, 4, \dots$. That is:

$$\begin{aligned} P(X \geq 2) &= P(X = 2) + P(X = 3) + P(X = 4) + \dots \\ &= (1 - p)^2 * p + (1 - p)^3 * p + (1 - p)^4 * p + \dots \\ &= p(1 - p)^2 [1 + (1 - p) + (1 - p)^2 + \dots] \\ &= p(1 - p)^2 [1/(1 - (1 - p))] \\ &= p(1 - p)^2 / p \\ &= (1 - p)^2 \end{aligned}$$

Substituting $p = 0.05$, we get:

$$P(X \geq 2) = (1 - 0.05)^2 = 0.9025.$$

Therefore, the probability that at least two defective components are discarded before the first non-defective component is found is 0.9025.

Question-9: You write a code over and over, and each time there is probability p that it works correctly, independently from previous attempts. What is the mean of X , the number of tries until the code works correctly?

Solution-9: We will apply the total expectation theorem, with $A_1 = \{X = 1\} = \{\text{first try is a success}\}$, $A_2 = \{X > 1\} = \{\text{first try is a failure}\}$, and end up with a much simpler calculation.

If the first try is successful, we have $X = 1$, and $E[X | X = 1] = 1$.

If the first try fails ($X > 1$), we have wasted one try, and we are back where we started. So, the expected number of remaining tries is $E[X]$, and $E[X | X > 1] = 1 + E[X]$.

Thus, $E[X] = P(X = 1)E[X | X = 1] + P(X > 1)E[X | X > 1] = p + (1 - p)(1 + E[X])$, from which we obtain $E[X] = 1/p$.

Question-10: The joint probability density function of two random variables is

$$f_{X,Y}(x, y) = k(x^2 + y)$$

where $2 < x < 3$ and $0 < y < 2$. Find the mean value of $E(x)$?

Solution-10:

$$\int_{y=0}^{y=2} \int_{x=2}^{x=3} f_{X,Y}(x, y) dx dy = 1$$

$$\int_{y=0}^{y=2} \int_{x=2}^{x=3} k(x^2 + y) dx dy = 1$$

By solving $k = 3/44$. Now find marginal pdf $f_X(x) = \int_{y=0}^{y=2} f_{X,Y}(x, y) dy$

$$f_X(x) = 2k(x^2 + 1)$$

$$E[X] = \int_{x=2}^{x=3} x f_X(x) dx = \int_{x=2}^{x=3} 2k(x^2 + 1) x dx = 225/88.$$

Question-11: The Rayleigh distribution function is given by

$$f_X(x) = \begin{cases} \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}; & x > 0 \\ 0; & \text{otherwise} \end{cases}$$

Calculate mean value $E(X)$ for the given random variable.

Solution-11: We have to find $E[X] = \int_{x=0}^{x=\infty} \frac{x^2}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx$.

To solve that we will be using the property of Gaussian random variable. We know that pdf of

Gaussian $g_X(x) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{x^2}{2\sigma^2}}$ with 0 mean and variance σ^2 .

$\int_{x=-\infty}^{x=\infty} x^2 g_X(x) dx = \sigma^2$ (as mean is 0).

$2 \int_{x=0}^{x=\infty} \frac{x^2}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{x^2}{2\sigma^2}} dx = \sigma^2$ as the above function was even.

By rearranging we can get

$$\int_{x=0}^{x=\infty} \frac{x^2}{\sigma^2} \frac{-x^2}{2\sigma^2} dx = \left(\frac{2\pi\sigma^2 \frac{1}{2}}{2} \right) = \sigma \sqrt{\frac{\pi}{2}}.$$

Question-12: Consider a random variable Y defined as

$$Y = \sum_{i=0}^n X_i$$

where random variable $X_1, X_2, X_3, \dots, X_n$ are statistically independent and defined as

$$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } (1-p) \end{cases}$$

(a) Find the mean of a random variable Y .

(b) The expected value of Y^2 .

Solution-12: (a) $E[Y] = E[\sum_0^n X_i] = \sum_0^n E[X_i]$

$$E[X_i] = p^1 + (1-p)^0 = p.$$

$$E[Y] = \sum_0^n p = (n+1)p.$$

$$(b) E[Y^2] = E[(\sum_0^n X_i)^2] = E[X_0 + X_1 + \dots + X_n]$$

$$= E[X_0^2 + X_1^2 + \dots + X_n^2 + 2X_0X_1 + 2X_0X_2 + \dots + 2X_0X_n + \dots + 2X_nX_1 + 2X_nX_2 + \dots + 2X_{n-1}X_n]$$

$$= E\left[\sum_{j=0}^n \sum_{i=0}^n X_i X_j\right] = \sum_{j=0}^n \sum_{i=0}^n E[X_i X_j] = \sum_{j=0}^n \sum_{i=0}^n E[X_i] E[X_j] = \sum_{j=0}^n \sum_{i=0}^n (p(n+1))^2$$

$$= (p(n+1))^2 (n+1)(n+1) = p^2(n+1)^4.$$

Question-13: The statistically independent random variable X and Y have mean values $E[X] = 2$ and $E[Y] = 4$. They have second order moment $E[X^2] = 8$ and $E[Y^2] = 25$. consider random variable $W = 3X - Y$. Find

(a) The mean value $E[W]$.

(b) The second order moment of W .

(c) The variance of W .

Solution-13: (a) $E[W] = 3E[X] - E[Y] = 3 * 2 - 4 = 2$.

(b) $E[W^2] = E[9X^2 + Y^2 - 6XY] = 9E[X^2] + E[Y^2] - 6E[X]E[Y] = (9 * 8) + 25 - (6 * 2 * 4) = 49$.

(c) $E[W^2] - E[W]^2 = 49 - 4 = 45$.

Question-14: A random variable X represents the value of coins given in change when purchases are made at the Reliance Fresh store. Suppose the probability of 1 paise, 5 paise, 10 paise and 25 paise being present are 0.35, 0.25, 0.20 and 0.15 respectively. What is the mean of X ?

Solution-14: $E[X] = 1 * 0.35 + 5 * 0.25 + 10 * 0.20 + 25 * 0.15 = 7.35$.

Question-15: Consider a one-dimensional random walk described as follows. An inebriated man walking in a narrow hallway takes steps of equal length l . He steps forward with probability $p = 3/4$ or backward with probability $q = 1/4$. Let X be the distance from the starting point after 100 steps.

a) What will be the mean value of X ?

b) What will be the variance of X ?

Solution-15: (a) Let's first consider the expected value of one step. The probability of moving forward is $p = 3/4$ and the probability of moving backward is $q = 1/4$. Therefore, the expected value of one step is:

$$E(\text{step}) = pl - ql = (3/4) * l - (1/4) * l = 1/2.$$

Now, after 100 steps, the expected value of the distance from the starting point can be calculated as:

$$E(X) = E(\text{number of steps})E(\text{step}) = 100(1/2) = 50l$$

Therefore, the mean value of X is 501.

(b) The variance of X can be calculated as follows:

Let's first calculate the variance of one step. The variance of a random variable is defined as:

$$\text{Var}(\text{step}) = E(\text{step}^2) - (E(\text{step}))^2$$

The probability of moving forward is $p = 3/4$ and the probability of moving backward is $q = 1/4$.

Therefore, the expected value of one step squared is:

$$E(\text{step}^2) = (p\text{l})^2 + (q\text{l})^2 = ((3/4)^*)^2 + ((1/4)^*1)^2 = 5/16 * 1^2$$

The expected value of one step is $E(\text{step}) = 1/2$ (as calculated above).

Therefore, the variance of one step is:

$\text{Var}(\text{step}) = E(\text{step}^2) - (E(\text{step}))^2 = (5/16 * 1^2) - (1/2)^2 = 1^2/16$. Now, after 100 steps, the variance of the distance from the starting point can be calculated as:

$$\text{Var}(X) = \text{Var}(\text{number of steps}) \text{Var}(\text{step}) = 100(1^2/16) = 25*1^2.$$

7.14 Lab Assignment 5

Question-1: A data file named `random_num` has been provided containing an array of 10000 random real numbers. Write a Python code to: a) Find the mean and standard deviation.

b) Plot the histogram (take bin size 0.02) of the given data and mention the most probable bin.

c) Fit gamma, log normal, beta, exponential, and normal distribution on given data using the hint given below. Based on the `ks_pvalue` report the distributions that our data follows.

d) From the histogram plot we can see that there are two distinct subgroups present in the data.

For each subgroup find the mean and variance and also fit the distributions given in part c. For each subgroup based on the `ks_pvalue` report the distributions the subgroup follows.

Hint: To fit the distribution to a data set use following packages and functions

```
1 from fitter import Fitter, get_common_distributions, get_distributions
2 f = Fitter(data, distributions= a list containg distribution's name)
3 f.fit()
4 f.summary()
```

For gamma, log normal, beta, exponential and normal we use following names in the code: “gamma”, “lognorm”, “beta”, “expon”, “norm” respectively. If `ks_pvalue` is less than 0.05 (level of significance) for any distribution, then we reject that distribution (means data do not follow that distribution). For example, if $ks_pvalue = 0.0006 < 0.05$ and our distribution is “norm” then we can say that data do not follow the normal distribution. If $ks_pvalue > 0.05$ for a particular distribution, then data follows that distribution.

Solution:

```

1 import numpy as np
2 import matplotlib.pyplot as plt
3 from fitter import Fitter, get_common_distributions, get_distributions
4
5 #load data
6 data = np.loadtxt('random_num.txt')

1 #part 1
2 print('The mean of the data is ', round(np.mean(data),5))
3 print('The variance of the data is ', round(np.var(data),5))

```

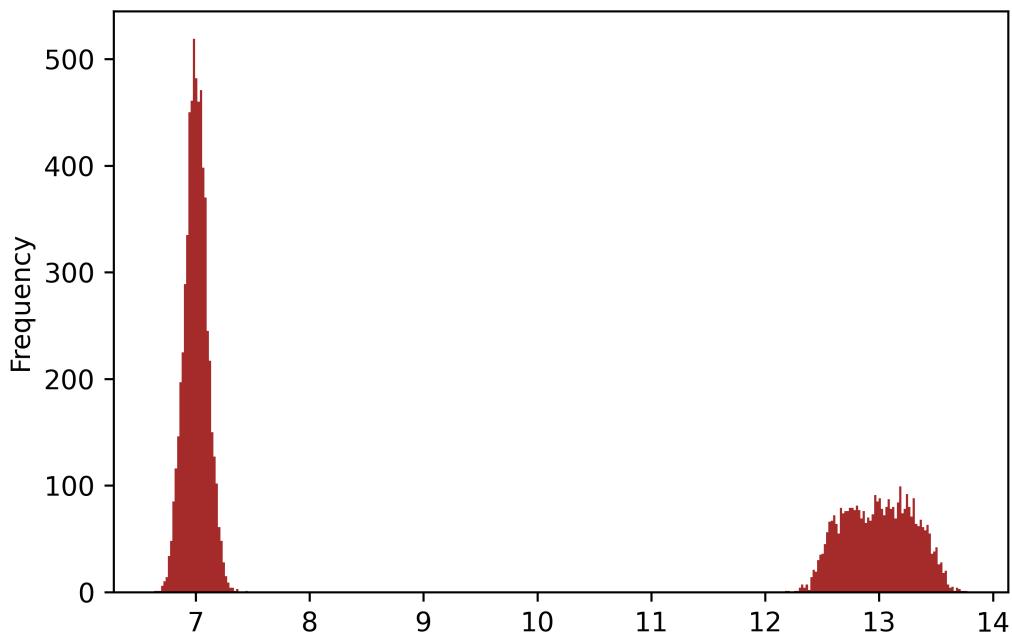
The mean of the data is 9.31984

The variance of the data is 8.56798

```

1 #Part 2
2 print('\nPart 2')
3 binwidth = 0.02
4 plt.figure()
5 n, bins, patches = plt.hist(data, bins=np.arange(min(data), max(data) + binwidth,
       binwidth), color="brown")
6 plt.ylabel('Frequency')
7 plt.show()
8 print('The most probable bin is {} and its range is [{},{}].'.format(np.argmax(n
       )+1,bins[np.argmax(n)],bins[np.argmax(n)+1]))

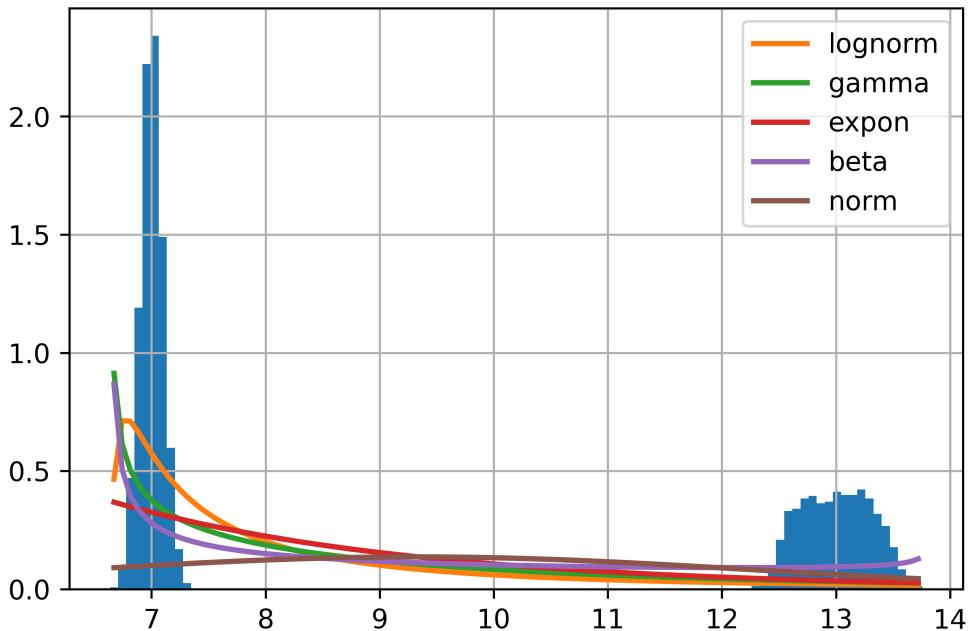
```



The most probable bin is 2 and its range is [6.975692064262501,6.9956920642625]

```
1 #Part 3
2 ff=Fitter(data, distributions= ['gamma',
3                               'lognorm',
4                               "beta",
5                               "expon",
6                               "norm"])
7
8 ff.fit()
9 print(ff.summary())
10 if len(ff.summary()[ff.summary() ['ks_pvalue']]>0.05])< 1:
11     print('Data does not follows any of the given distributions')
12 else:
13     print('Data follows the following distributions')
14     print(ff.summary()[ff.summary() ['ks_pvalue']]>0.05])
15 plt.show()
```

	sumsquare_error	aic	...	ks_statistic	ks_pvalue
lognorm	10.715943	540.725776	...	0.271907	0.0
gamma	12.962323	495.540185	...	0.308715	0.0
expon	13.163634	466.764898	...	0.408352	0.0
beta	13.353331	429.434384	...	0.351068	0.0
norm	15.356282	460.865435	...	0.369693	0.0



```

1 #Part 4
2 data1 = data[data<9.31]
3 print('Mean of left subgroup is ', np.mean(data1))
4 print('Mean of left subgroup is ', np.var(data1))
5
6 ff=Fitter(data1, distributions= [ 'gamma',
7                         'lognorm',
8                         "beta",
9                         "expon",
10                        "norm"])
11
12 ff.fit()
13 print(ff.summary())
14 if len(ff.summary()[ff.summary()['ks_pvalue']>0.05])< 1:
15     print('Left subgroup does not follow any of the given distributions')
16 else:
17     print('Left subgroup follows the following distributions')
18     print(ff.summary()[ff.summary()['ks_pvalue']>0.05])
19 plt.show()
20
21
22 data2 = data[data>9.31]
23 print('Mean of right subgroup is ', np.mean(data1))

```

```

24 print('Mean of right subgroup is ', np.var(data1))

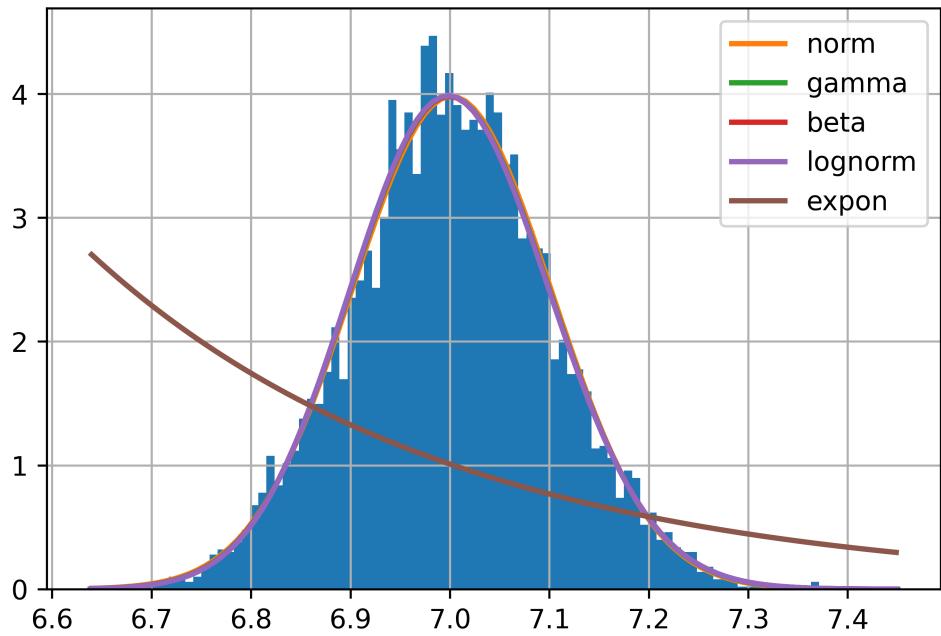
25
26 ff=Fitter(data2, distributions= [ 'gamma',
27                         'lognorm',
28                         "beta",
29                         "expon",
30                         "norm"])
31
32 ff.fit()
33 print(ff.summary())
34 if len(ff.summary()[ff.summary()['ks_pvalue'] > 0.05]) < 1:
35     print('Right subgroup does not follow any of the given distributions')
36 else:
37     print('Right subgroup follows the following distributions')
38     print(ff.summary()[ff.summary()['ks_pvalue'] > 0.05])
39 plt.show()

```

Mean of left subgroup is 7.001718764615539

Mean of left subgroup is 0.010048325963023933

	sumsquare_error	aic	...	ks_statistic	ks_pvalue
norm	2.987818	299.481857	...	0.010528	0.501478
gamma	3.031803	293.551272	...	0.011291	0.411765
beta	3.058419	293.800088	...	0.011713	0.366437
lognorm	3.059593	291.810474	...	0.011600	0.378253
expon	259.575517	26.238785	...	0.377360	0.000000



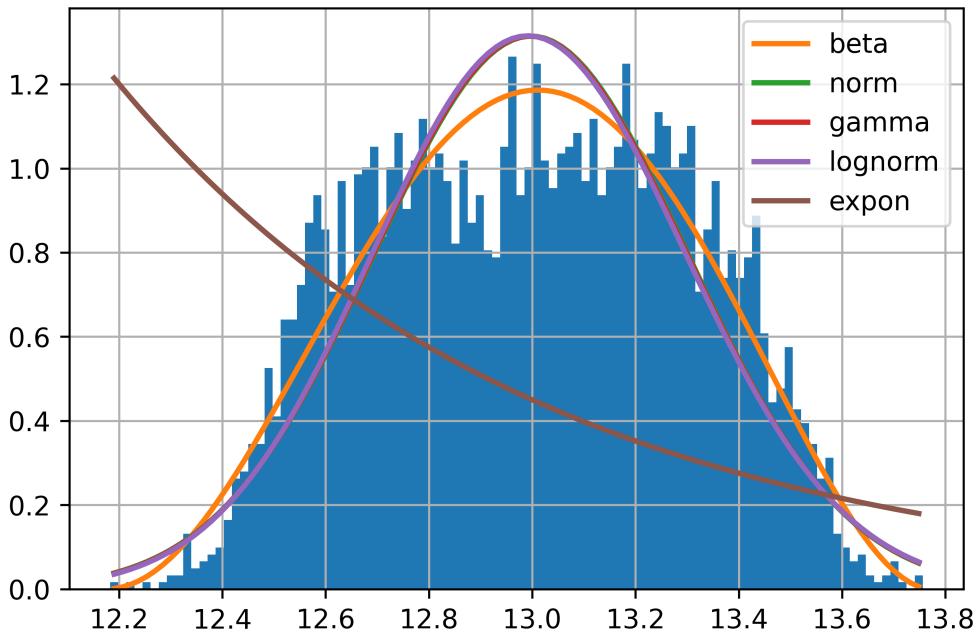
Left subgroup follows the following distributions

	sumsquare_error	aic	...	ks_statistic	ks_pvalue
norm	2.987818	299.481857	...	0.010528	0.501478
gamma	3.031803	293.551272	...	0.011291	0.411765
beta	3.058419	293.800088	...	0.011713	0.366437
lognorm	3.059593	291.810474	...	0.011600	0.378253

Mean of right subgroup is 12.997904116408616

Mean of left subgroup is 0.09211451373360274

	sumsquare_error	aic	...	ks_statistic	ks_pvalue
beta	1.859625	200.904178	...	0.032163	6.568683e-04
norm	3.755303	174.312086	...	0.047543	4.941490e-08
gamma	3.763071	176.511989	...	0.046934	7.716719e-08
lognorm	3.785494	176.935672	...	0.046983	7.446776e-08
expon	34.223579	156.224811	...	0.294381	2.235018e-297



Right subgroup does not follow any of the given distributions

Question-2:

Covid-19 proved to be a deadly pandemic which caused millions of deaths and destroyed the economy of various nations. Lockdowns were imposed to contain the spread of the virus. Mobility was significantly reduced during this period.

The Community Mobility Reports provided by google show movement trends by region, across different categories of places. These reports are created with aggregated, anonymized sets of data from users who have turned on the Location History setting, which is off by default. The data shows how visitors to (or time spent in) categorized places change compared to our baseline days. A baseline day represents a normal value for that day of the week. The baseline day is the median value from the 5-week period Jan 3 – Feb 6, 2020. For each region-category, the baseline isn't a single value—it's 7 individual values. The same number of visitors on 2 different days of the week, result in different percentage changes.

You are given a dataset which contains the mobility data of India in 2021. You are expected to do the following for Mumbai City:

1. Plot the monthly graph (Month Name on X-axis) for each category of mobility (grocery, retail, transit, parks, residential).
- What do you infer from these graphs?
2. Calculate the variance for each of the 6 mobilities for 2 time periods: During the lockdown period (consider only the month of April and May (till 20/05/2021)).

3. For the entire year (365 days). Compare these variances for each category. What do you infer ? C. Let's assume that the data given by Google is probabilistic in nature and each mobility has a certain probability associated with it. So on a particular day for a region we compute the expected mobility using the formula :

Grocery/Pharma	Retail	Transport	Parks	Residential	Workplace
p=0.2	p=0.2	p=0.05	p=0.02	p=0.5	p=0.03

Compute the expected mobility for each day during the lockdown period and compare these values with the original mobilities for each category during the lockdown period. What do you infer?

Solution:

```

1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import statistics
5 from tabulate import tabulate
6
7 # df20 = pd.read_csv("2020_IN_Region_Mobility_Report.csv")
8 df21 = pd.read_csv("2021_IN_Region_Mobility_Report.csv")

```

Part A

```

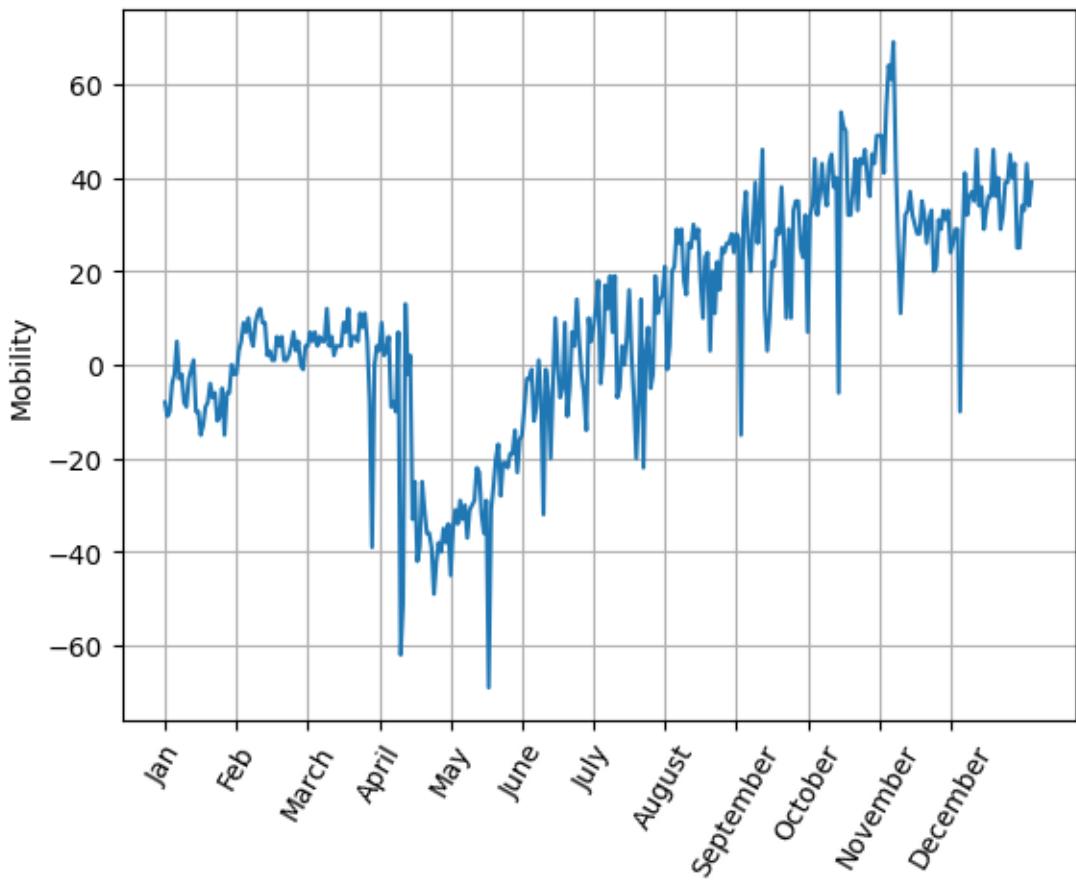
1 grocery = list(df21['grocery_and_pharmacy_percent_change_from_baseline'])
   [128037:128402]
2 retail = list(df21['retail_and_recreation_percent_change_from_baseline'])
   [128037:128402]
3 transit = list(df21['transit_stations_percent_change_from_baseline'])
   [128037:128402]
4 parks = list(df21['parks_percent_change_from_baseline']) [128037:128402]
5 workplaces = list(df21['workplaces_percent_change_from_baseline'])
   [128037:128402]
6 residential = list(df21['residential_percent_change_from_baseline'])
   [128037:128402]
7
8 months=['Jan', 'Feb', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'November', 'December']
9 xlen=[x for x in range(0,335,30)]
10 plt.plot([x for x in range(len(residential))], grocery)
11 plt.xticks(xlen,months, rotation=60)

```

```

12 plt.grid()
13 plt.ylabel('Mobility')
14 plt.show()

```



Part B

```

1 grocery_lock = statistics.variance(grocery[90:140])
2 grocery_full = statistics.variance(grocery)
3
4 retail_lock = statistics.variance(retail[90:140])
5 retail_full = statistics.variance(retail)
6
7 transit_lock = statistics.variance(transit[90:140])
8 transit_full = statistics.variance(transit)
9
10 parks_lock = statistics.variance(parks[90:140])
11 parks_full = statistics.variance(parks)
12
13 residential_lock = statistics.variance(residential[90:140])
14 residential_full = statistics.variance(residential)
15

```

```

16 workplace_lock = statistics.variance(workplaces[90:140])
17 workplace_full = statistics.variance(workplaces)
18
19
20 data = [["Grocery", grocery_lock, grocery_full],
21         ["Retail", retail_lock, retail_full],
22         ["Transport", transit_lock, transit_full],
23         ["Parks", parks_lock, parks_full],
24         ["Residential", residential_lock, residential_full],
25         ["Workplace", workplace_lock, workplace_full]]
26
27 print(tabulate(data, headers=["Category", "Lockdown Period", "Whole Year"]))

```

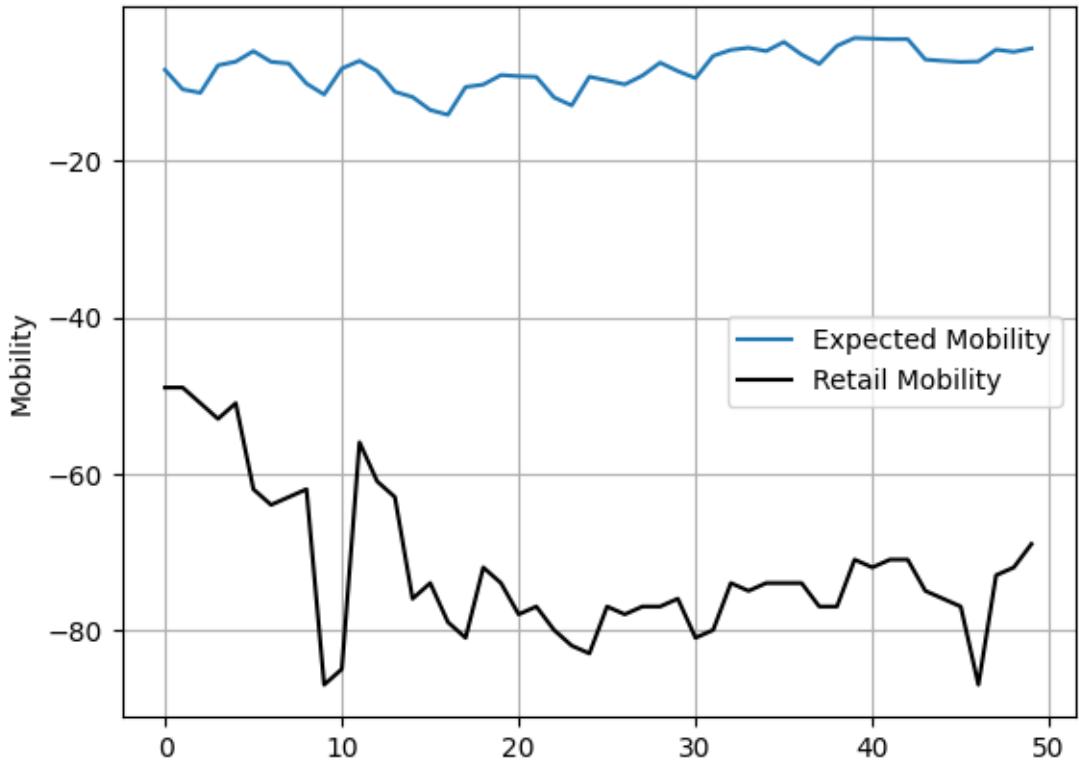
Category	Lockdown Period	Whole Year
Grocery	338.689	577.514
Retail	94.5473	274.726
Transport	105.265	308.207
Parks	69.4106	186.997
Residential	18.869	29.9222
Workplace	127.536	216.54

Part C

```

1 mob = []
2 for i in range(len(grocery[90:140])):
3     E_mob = 0.2*grocery[i] + 0.2*retail[i] + transit[i]*0.05 + parks[i]*0.02 +
4         residential[i]*0.5 + workplaces[i]*0.03 ##### Calculating the expected
5         mobility for each day
6     mob.append(E_mob)
7
8 plt.plot([x for x in range(len(mob))],mob)
9 plt.plot([x for x in range(len(mob))], retail[90:140], color='black') # Change
10    the y values to compare with all the mobilities
11 plt.grid()
12 plt.ylabel("Mobility")
13 plt.legend(['Expected Mobility', 'Retail Mobility'])
14 plt.title('Comparison of expected mobility during lockdown')
15 plt.show()

```



Question-3:

Consider the following data for two stocks A and B :

Scenario	Probability	Asset A return (%)	Asset B return (%)
1	0.4	-10	5
2	0.2	10	20
3	0.4	20	10

1. Theoretically compute each stock's expected return, variance, and standard deviation.
2. For each stock randomly sample N return values following the above-given probability distribution. For simulation take $N = 100, 200, 500$, and 1000 . For each value of N , from the sampled return values compute mean, variance, and standard deviation and compare them with the theoretical results for each stock.
3. Suppose that the investor makes a portfolio that invests w_1 proportion of the wealth in stock A and the remaining (let's say w_2) in stock B . For $(w_1 = 0.4, w_2 = 0.6)$, $(w_1 = 0.5, w_2 = 0.5)$ and $(w_1 = 0.8, w_2 = 0.2)$ theoretically compute the expected portfolio return and risk. Based on the expected risk and return value which (w_1, w_2) combination you will prefer for investment?
4. For your preferred (w_1, w_2) combination compute the empirical portfolio return and risk from

the simulated return data in step 2 for $N = 1000$. Compare the theoretical and empirical results.

Soution:

Par 1: The expected return on stock A is

$$\mu_A = 0.4 \times (-10) + 0.2 \times 10 + 0.4 \times (20) = 6.0\%$$

The expected return on stock B is

$$\mu_B = 0.4 \times 5 + 0.2 \times (20) + 0.4 \times (10) = 10.0\%$$

The variance of stock A is

$$\sigma_A^2 = 0.4(-0.1 - 0.06)^2 + 0.2(0.1 - 0.06)^2 + 0.4(0.2 - 0.06)^2 = 0.0184$$

The variance of stock B is

$$\sigma_B^2 = 0.4(0.05 - 0.1)^2 + 0.2(0.2 - 0.1)^2 + 0.4(0.1 - 0.1)^2 = 0.003$$

The variance of stock A is

$$\sigma_A^2 = 0.4(-0.1 - 0.06)^2 + 0.2(0.1 - 0.06)^2 + 0.4(0.2 - 0.06)^2 = 0.0184$$

The standard deviation of stock A is

$$\sigma_A = \sqrt{\sigma_A^2} \times 100 = 13.56\%$$

The standard deviation of stock B is

$$\sigma_B = \sqrt{\sigma_B^2} \times 100 = 5.47\%$$

```

1 import numpy as np
2 import matplotlib.pyplot as plt
3 import math
4
5
6 ##### Part 1#####
7 def Expected_asset_return(x,p):
8     R = 0

```

```

9     for i in range(len(x)):
10        R = R + p[i] * x[i]
11    return R
12
13 def Asset_variance(x,p):
14     R = Expectated_asset_return(x,p)
15     V = 0
16     for i in range(len(x)):
17         V = V + p[i]*(x[i] - R)**2
18     return V
19
20 r_A = [-0.1,0.1,0.2]
21 r_B = [0.05,0.2,0.1]
22 p = [0.4, 0.2, 0.4]
23
24 u_A = Expectated_asset_return(r_A, p)
25 v_A = Asset_variance(r_A, p)
26 sd_A = Asset_variance(r_A, p)**0.5
27
28 u_B = Expectated_asset_return(r_B, p)
29 v_B = Asset_variance(r_B, p)
30 sd_B = Asset_variance(r_B, p)**0.5
31
32 print('Part (1)')
33 print('The Expected return on Asset A is {}% and B is {}%'.format(round(u_A,3)*100, round(u_B,3)*100))
34 print('Variance of Asset A is {} and B is {}'.format(round(v_A,4),round(v_B,4)))
35 print('Standard deviation of Asset A is {}% and B is {}%'.format(round(sd_A,4)*100, round(sd_B,4)*100))

```

The Expected return on Asset A is 6.0% and B is 10.0%

Variance of Asset A is 0.0184 and B is 0.003

Standard deviation of Asset A is 13.56% and B is 5.48%

Part 2

```

1 def Samples(x,p,N):
2     return np.random.choice(x,N,p)
3
4 for N in [100,200,500,1000]:

```

```

5   SA = Samples(r_A ,p,N)
6   SB = Samples(r_B ,p, N)
7   print( 'For N= ', N)
8   print( 'The simulated mean of stock A is {}% and stock B is {}%'.format(
9     round(np.mean(SA)*100,3), round(np.mean(SB)*100,3)))
10  print( 'The simulated variance of stock A is {} and stock B is {}'.format(
11    round(np.var(SA),5), round(np.var(SB),5)))
12  print( 'The simulated standard deviation of stock A is {}% and stock B is
13    {}%'.format(round(np.var(SA)**0.5,5)*100, round(np.var(SB)**0.5,5)*100))
14  print()

```

For N= 100

The simulated mean of stock A is 5.2% and stock B is 11.0%

The simulated variance of stock A is 0.0145 and stock B is 0.0033

The simulated standard deviation of stock A is 12.04% and stock B is 5.745%

For N= 200 The simulated mean of stock A is 5.1% and stock B is 11.55%

The simulated variance of stock A is 0.0164 and stock B is 0.00378

The simulated standard deviation of stock A is 12.806000000000001% and stock B is 6.152%

For N= 500

The simulated mean of stock A is 6.06% and stock B is 11.35% The simulated variance of stock A is 0.01635 and stock B is 0.0037 The simulated standard deviation of stock A is 12.786% and stock B is 6.085%

For N= 1000

The simulated mean of stock A is 7.39% and stock B is 11.355%

The simulated variance of stock A is 0.01543 and stock B is 0.00379

The simulated standard deviation of stock A is 12.421% and stock B is 6.155%

Part 3

The formula to compute the expected return of the portfolio (w_1, w_2) is

$$\mu_P = w_1 \times \mu_A + w_2 \times \mu_B \quad (7.1)$$

and variance is

$$\sigma_P^2 = w_1^2 \sigma_A^2 + w_2^2 \sigma_B^2 + 2w_1 w_2 \rho_{AB} \times \sigma_A \sigma_B \quad (7.2)$$

where $\rho_{AB} = \text{cov}(A, B) / (\sigma_A \times \sigma_B)$ the correlation coefficient. On using the correlation coefficient formula we get

$$\rho_{AB} = 0.5383 \quad (7.3)$$

For (w1 = 0.4, 0.6):

The expected return of the portfolio is :

$$\mu_P = 0.4 \times 6 + 0.6 \times 10 = 8.4\% \quad (7.4)$$

Variance

$$\sigma_P^2 = (0.4)^2 \times 0.0184 + (0.6)^2 \times 0.003 + 2 \times 0.4 \times 0.6 \times 0.5383 \times 0.1356 \times 0.0547 = 0.00594 \quad (7.5)$$

Standard Deviation:

$$\sigma_A = \sqrt{0.00594} \times 100 = 7.71\% \quad (7.6)$$

Similarity we can compute for other combinations, and it comes out the following:

For (w1=0.5, w2=0.5)

$$\mu_P = 8.0\%$$

$$\sigma_P^2 = 0.00735$$

$$\sigma_A = 8.573\%$$

For (w1=0.8, w2 = 0.2)

$$\mu_P = 6.8\%$$

$$\sigma_P^2 = 0.01318$$

$$\sigma_A = 11.479\%$$

```

1 def Expected_portfolio_return(w, u):
2     Rp = 0
3     for i in range (len(w)):
4         Rp = Rp + w[i]*u[i]
5     return Rp
6
7

```

```

8 def corr_coeff(r, p):
9     u = np.zeros(len(r))
10    for i in range(len(u)):
11        u[i] = Expected_asset_return(r[i], p)
12
13    v = np.zeros(len(r))
14    for i in range(len(v)):
15        v[i] = Asset_variance(r[i], p)
16
17
18
19    corr= np.zeros((len(r),len(r)))
20
21    for i in range(corr.shape[0]):
22        for j in range(corr.shape[1]):
23            for k in range(len(p)):
24                corr[i,j] =corr[i,j] + p[k] *(r[i][k] - u[i])*(r[j][k] - u[j])
25
26        corr[i,j] = corr[i,j]/(math.sqrt(v[i])*math.sqrt(v[j]))
27
28    return corr
29
30
31
32 def Portfolio_variance(w, v, corr_coeff_matrix):
33     v_P = 0
34     for i in range(len(w)):
35         v_P = v_P + w[i]**2 * v[i]
36     v_P1 = 0
37     for i in range(corr_coeff_matrix.shape[0]):
38         for j in range(i+1, corr_coeff_matrix.shape[0]):
39             v_P1 = v_P1 + 2 *w[i]*w[j] *corr_coeff_matrix[i,j]* math.sqrt(v[i])
40             * math.sqrt(v[j])
41             v_P = v_P + v_P1
42             if np.round(v_P1,5) <= 0:
43                 v_P1 = 0
44             return v_P
45 r = [r_A, r_B]

```

```

46
47 corr_coeff_matrix = corr_coeff(r, p)
48
49 print('Correlation coefficient matrix is: \n', corr_coeff_matrix)
50
51
52
53 Risk_return= []
54
55 W= [[0.1 ,0.9],[0.2 ,0.8],[0.3 ,0.7],[0.4 ,0.6],[0.5 ,0.5],[0.6 ,0.4],[0.7 ,0.3]
56 ,[0.8 ,0.2],[0.9 ,0.1]]
57 for w in W:
58     u = [u_A, u_B]
59
60     v = [v_A, v_B]
61
62     u_P = Expected_portfolio_return(w, u)
63     v_P = Portfolio_variance(w, v, corr_coeff_matrix)
64     sd_P = math.sqrt(v_P)
65     Risk_return.append([sd_P,u_P])
66
67     print( '\nFor (w1, w2) = ',w)
68     print( 'The expected return of the portfolio is {}%'.format(round(u_P
69 *100,3)))
70     print( 'Variance of the portfolio is {}'.format(round(v_P,5)))
71     print( 'Standard deviation of the portfolio is {}%'.format(round(sd_P
72 *100,3)))
73
74 Risk_return = np.array(Risk_return)
75 plt.plot(Risk_return[:,0]*100,Risk_return[:,1]*100)
76 plt.xlabel( 'Return(%)')
77 plt.ylabel( 'Risk(%)')

```

Correlation coefficient matrix is:

```

[[1,0.5383819]
[0.5383819,1]]

```

For (w1, w2) = [0.1, 0.9]

The expected return of the portfolio is 9.6%

Variance of the portfolio is 0.00333

Standard deviation of the portfolio is 5.774%

For $(w_1, w_2) = [0.2, 0.8]$

The expected return of the portfolio is 9.2%

Variance of the portfolio is 0.00394

Standard deviation of the portfolio is 6.274%

For $(w_1, w_2) = [0.3, 0.7]$

The expected return of the portfolio is 8.8%

Variance of the portfolio is 0.00481

Standard deviation of the portfolio is 6.933%

For $(w_1, w_2) = [0.4, 0.6]$

The expected return of the portfolio is 8.4%

Variance of the portfolio is 0.00594

Standard deviation of the portfolio is 7.71%

For $(w_1, w_2) = [0.5, 0.5]$ The expected return of the portfolio is 8.0%

Variance of the portfolio is 0.00735

Standard deviation of the portfolio is 8.573%

For $(w_1, w_2) = [0.6, 0.4]$ The expected return of the porftolio is 7.6% Variance of the porftolio is

0.00902 Standard deviation of the porftolio is 9.499%

For $(w_1, w_2) = [0.7, 0.3]$

The expected return of the portfolio is 7.2%

Variance of the portfolio is 0.01097

Standard deviation of the portfolio is 10.472%

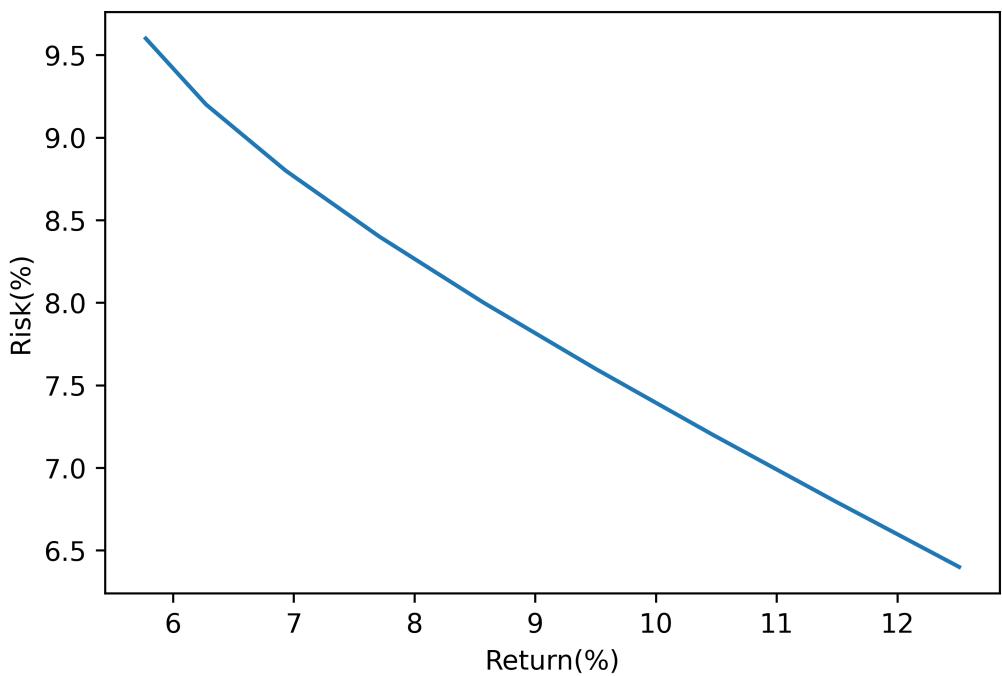
For $(w_1, w_2) = [0.8, 0.2]$ The expected return of the porftolio is 6.8Variance of the porftolio is

0.01318 Standard deviation of the porftolio is 11.479

For $(w_1, w_2) = [0.9, 0.1]$ The expected return of the portfolio is 6.4%

Variance of the portfolio is 0.01565

Standard deviation of the portfolio is 12.512%



Part 4

```

1 N=10000
2 SA = Samples(r_A,p,N)
3 SB = Samples(r_B,p, N)
4
5 w1, w2 = 0.1,0.9
6
7 P = w1*SA + w2*SB
8
9 print('The mean of simulated portfolio is {}'.format(round(np.mean(P)*100, 4)))
10 print('The variance of simulaed portfolio is {}'.format(round(np.var(P),4)))
11 print('The standard deviation of simulated portfolio is {}'.format(round(np.var(P)
    **0.5*100,4)))

```

The mean of simulated portfolio is 11.2051%.

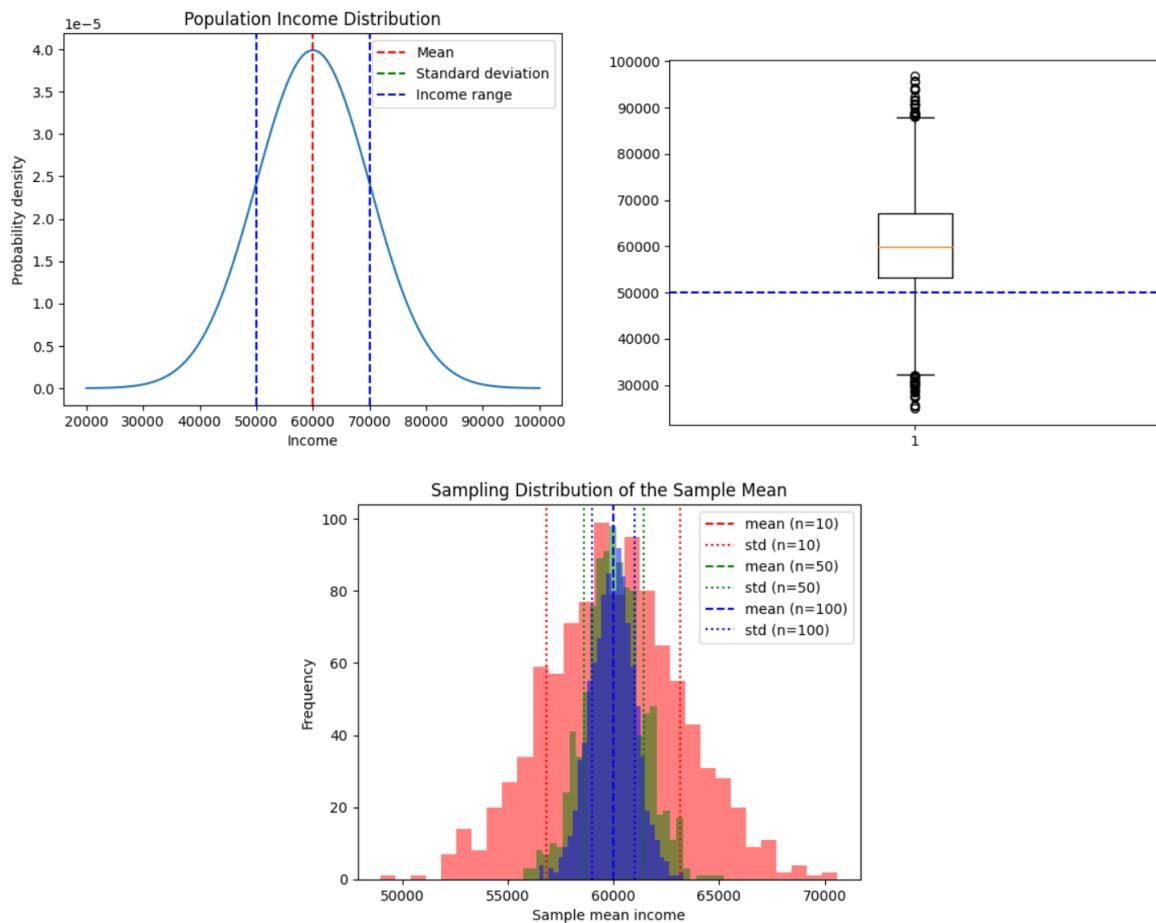
The variance of simulaed portfolio is 0.0033

The standard deviation of simulated portfolio is 5.767%

7.15 Lab Assignment 7

1. Suppose that the annual incomes of a population of 10,000 people follow a normal distribution with a mean of \$ 60,000 and a standard deviation of \$ 10,000.

- (a) What is the probability that a randomly selected person from this population has an income between \$ 50,000 and \$ 70,000 per year?
- (b) What is the probability that the average income of a random sample of 100 people from this population is between \$ 55,000 and \$ 65,000 per year?
- (c) Create a histogram of the income distribution of the population, and add vertical lines to indicate the mean, standard deviation, and the income range in part a.
- (d) Create a box plot of the income distribution of the population, and add a horizontal line to indicate the income range in part a.
- (e) Create a histogram of the sampling distribution of the sample mean for sample sizes of 10, 50, and 100, and add vertical lines to indicate the mean and standard deviation of the sampling distribution. How do you infer your observation with regards to the Central Limit Theorem? [Use: plt.axvline to draw the lines]



```
1 import numpy as np
```

```

2 import matplotlib.pyplot as plt
3 from scipy.stats import norm
4
5 # Define the population parameters
6 mu = 60000
7 sigma = 10000
8 n_population = 10000
9
10 # Define the income range in part 1
11 lower_income = 50000
12 upper_income = 70000
13
14 # Define the sample size in part 2
15 n_sample = 100
16
17 # Define the income range in part 2
18 lower_mean_income = 55000
19 upper_mean_income = 65000
20
21 # a. Probability of a randomly selected person having an
22 income in the range [50,000, 70,000]
23 z_lower = (lower_income - mu) / sigma
24 z_upper = (upper_income - mu) / sigma
25 prob = norm.cdf(z_upper) - norm.cdf(z_lower)
26 print(f"Probability of income in range [{lower_income},
27 {upper_income}] = {prob:.2f}")
28
29 # b Probability of sample mean income in the range [55,000,
30 65,000]
31 sem = sigma / np.sqrt(n_sample)
32 z_lower = (lower_mean_income - mu) / sem
33 z_upper = (upper_mean_income - mu) / sem
34 prob = norm.cdf(z_upper) - norm.cdf(z_lower)
35 print(f"Probability of sample mean income in range
36 [{lower_mean_income}, {upper_mean_income}] = {prob:.2f}")
37
38 # c Histogram of the income distribution of the population
39 x = np.linspace(mu - 4*sigma, mu + 4*sigma, 1000)
40 y = norm.pdf(x, mu, sigma)
41 plt.plot(x, y)
42 plt.axvline(x=mu, color='r', linestyle='--', label='Mean')
43 plt.axvline(x=mu-sigma, color='g', linestyle='--',
44 label='Standard deviation')
45 plt.axvline(x=mu+sigma, color='g', linestyle='--')
46 plt.axvline(x=lower_income, color='b', linestyle='--',
47 label='Income range')

```

```

48 plt.axvline(x=upper_income, color='b', linestyle='--')
49 plt.legend()
50 plt.xlabel('Income')
51 plt.ylabel('Probability density')
52 plt.title('Population Income Distribution')
53 plt.show()
54
55 # d Box plot of the income distribution of the population
56 incomes = np.random.normal(mu, sigma, n_population)
57 plt.boxplot(incomes)
58 plt.axhline(y=lower_income, color='b', linestyle='--',
59 label='Income range')
60
61 # e Histogram of the sampling distribution of the sample mean
62 for sample sizes of 10, 50, and 100
63 sample_sizes = [10, 50, 100]
64 colors = ['r', 'g', 'b']
65 for i, n in enumerate(sample_sizes):
66     means = []
67     for j in range(1000):
68         sample = np.random.choice(incomes, n)
69         means.append(np.mean(sample))
70     plt.hist(means, bins=30, alpha=0.5, color=colors[i])
71     mean_sampling_dist = mu
72     std_sampling_dist = sigma / np.sqrt(n)
73     plt.axvline(x=mean_sampling_dist, color=colors[i],
74     linestyle='--', label=f'mean (n={n})')
75     plt.axvline(x=mean_sampling_dist - std_sampling_dist,
76     color=colors[i], linestyle=':', label=f'std (n={n})')
77     plt.axvline(x=mean_sampling_dist + std_sampling_dist,
78     color=colors[i], linestyle=':')
79 plt.legend()
80 plt.xlabel('Sample mean income')
81 plt.ylabel('Frequency')
82 plt.title('Sampling Distribution of the Sample Mean')
83 plt.show()

```

2. Suppose we have a sequence of independent and identically distributed random variables $X_1, X_2, X_3, \dots, X_n$, where each X_i has a gamma distribution with shape parameter $\alpha = 2$ and scale parameter $\beta = 1/2$. We want to investigate the weak law of large numbers for this distribution using simulation. Write a Python code to generate n independent and identically distributed samples from this distribution, and compute the sample mean S_n for each n . Plot a histogram of the sample means for various values of n , such as $n = 10, 50, 100, 500$, and 1000 .
 - (a) Based on your plots, explain whether the sample means appear to converge to the expected

value of X as n increases, and discuss how the sample size affects the accuracy of the estimate of the expected value.

- (b) Compute the sample variance for each n , and plot a histogram of the sample variances for the same values of n . Based on your plots, explain whether the sample variances appear to converge to the true variance of X as n increases, and discuss how the sample size affects the accuracy of the estimate of the variance.
- (c) Calculate the standard error of the sample means for each value of n , and plot a histogram of the standard errors. Based on your plots, discuss how the standard errors change with the sample size, and how they relate to the accuracy of the sample mean estimates.

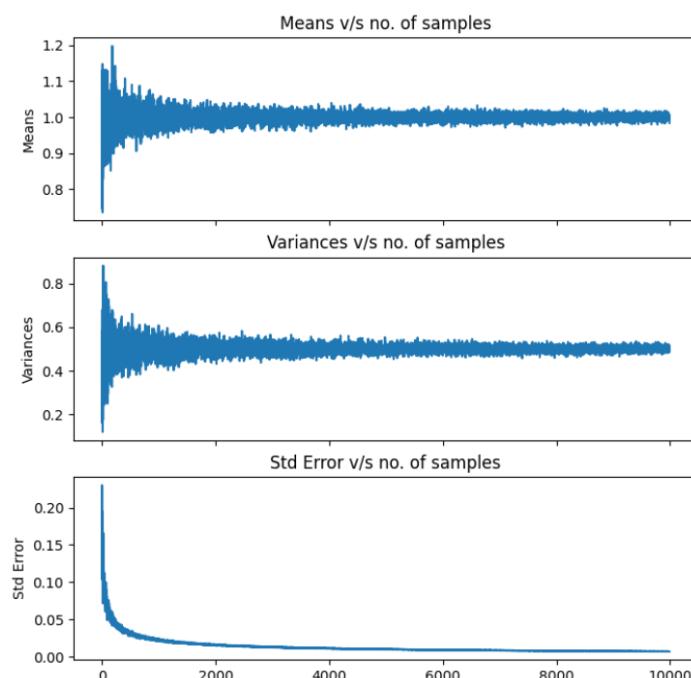


Fig. 7.1: Plot: Q2

```

1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 # Set the parameters of the gamma distribution
5 alpha = 2
6 beta = 1/2
7
8 # Set the values of n to simulate
9 # ns = [10, 50, 100, 500, 1000]
10 ns = 10000
11

```

```

12 # Initialize the arrays to store the sample means, variances,
13 and standard errors
14 means = []
15 variances = []
16 std_errors = []
17
18 # Simulate the samples for each value of n
19 for n in range(10, ns):
20     # Generate n samples from the gamma distribution
21     samples = np.random.gamma(alpha, beta, n)
22
23     # Compute the sample mean and store it in the means array
24     sample_mean = np.mean(samples)
25     means.append(sample_mean)
26
27     # Compute the sample variance and store it in the variances array
28     sample_variance = np.var(samples)
29     variances.append(sample_variance)
30
31     # Compute the standard error of the sample mean and store it in the
32     # std_errors array
33     std_error = np.sqrt(sample_variance/n)
34     std_errors.append(std_error)
35
36 # Plot the histograms of the sample means, variances, and
37 standard errors
38 fig, axs = plt.subplots(3, sharex=True, figsize=(8, 8))
39
40 axs[0].plot(means)
41 axs[0].set_ylabel("Means")
42 axs[0].set_title("Means v/s no. of samples")
43
44 axs[1].plot(variances)
45 axs[1].set_ylabel("Variances")
46 axs[1].set_title("Variances v/s no. of samples")
47
48 axs[2].plot(std_errors)
49 axs[2].set_ylabel("Std Error")
50 axs[2].set_title("Std Error v/s no. of samples")
51 plt.show()

```

3. You are given a dataset of exam scores from a population of 75 students. The mean score of the population is unknown, but you are interested in estimating it using a random sample of 5,10,15,25 scores from the population. Use pandas to open the ‘marks.csv’ dataset.

- (a) Write a Python function that takes in the dataset and calculates the point estimate of the population mean using the sample mean. [Hint: Calculate the means of 100 samples of the given sizes and compute the overall mean of the 100 means]
- (b) Find the Population variance and standard deviation using the sample variance and standard distribution.
- (c) Plot sampling distribution of sample means and print the sampling error for sample sizes($n=5,10,15, 25$)

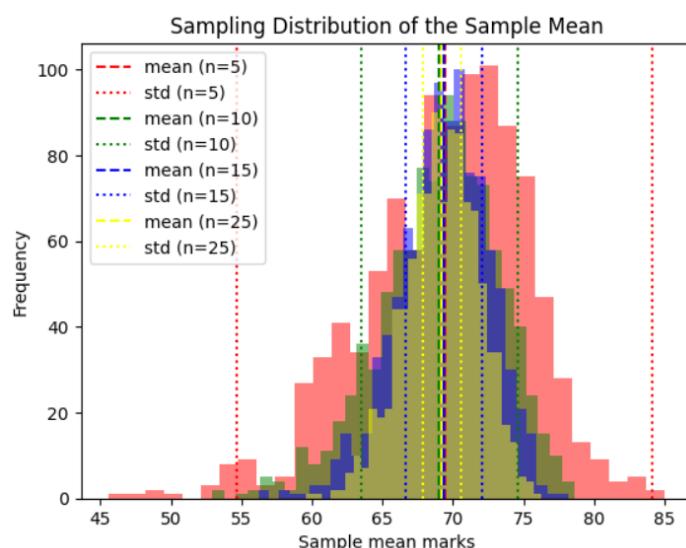


Fig. 7.2: Plot: Q3

```

1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4
5 # Load the dataset
6 df = pd.read_csv('marks.csv')
7
8 # Calculate the sample size
9 n = [5, 10, 15, 25]
10
11 # Calculate the sample mean
12 sample_mean = np.mean(df['marks'].sample(i) for i in n)
13
14 # Function to calculate the point estimate of the population
15 mean
16 for i in n:
17     def point_estimate(df, i):

```

```

18     sample_mean = np.mean(df['marks'].sample(i))
19     sample_variance = np.var(df['marks'].sample(i))
20     std_error = np.sqrt(sample_variance)
21
22     return sample_mean, sample_variance, std_error
23
24 # Call the functions and print the results
25 print('Point Estimate:', point_estimate(df, n))
26
27 sample_sizes = [5, 15, 25]
28 colors = ['r', 'g', 'b']
29 for i, n in enumerate(sample_sizes):
30     means = []
31     for j in range(1000):
32         sample = np.random.choice(df['marks'], n)
33         means.append(np.mean(sample))
34     plt.hist(means, bins=30, alpha=0.5, color=colors[i])
35     mean_sampling_dist = mu
36     std_sampling_dist = sigma / np.sqrt(n)
37     plt.axvline(x=mean_sampling_dist, color=colors[i],
38                 linestyle='--', label=f'mean (n={n})')
39     plt.axvline(x=mean_sampling_dist - std_sampling_dist,
40                 color=colors[i], linestyle=':', label=f'std (n={n})')
41     plt.axvline(x=mean_sampling_dist + std_sampling_dist,
42                 color=colors[i], linestyle=':')
43 plt.legend()
44 plt.xlabel('Sample mean income')
45 plt.ylabel('Frequency')
46 plt.title('Sampling Distribution of the Sample Mean')
47 plt.show()

```

Chapter 8

Estimation Theory

8.1 Estimation

So, far we have understood finding, a best-fit distribution on a given data can help in predicting the random variable. Once we know the distribution, question arise of estimating its parameter. For eg., if it is normal, parameter μ and σ^2 need to be estimated.

Therefore, it is a parameter estimation problem, and it can be dealt with method like maximum likelihood estimation (MLE). This is achieved by maximizing a likelihood function so that, under the unused statistical model, the observed data is most probable.

Suppose we have a random sample X_1, \dots, X_n whose best-fit prob. distributor has some unknown parameter θ . Our aim is to find a good estimator of which maximizes the likelihood of getting the data we observed.

8.1.1 Difference between likelihood and probability.

$$L(\mu, \sigma; \text{data}) = P(\text{data} ; \mu, \sigma)$$

$P(\text{data}, \mu, \sigma)$ is probability of finding an event, given the model parameter. $L(\mu, \sigma; \text{data})$ is its likelihood of the parameters μ and σ given that we have observed a data.

The basic idea.

To estimate the model parameters with given data. Definition Let x_1, x_2, \dots, x_n be a sample from dis-

tribution that is having unknown parameter like $\theta_1, \theta_2, \dots, \theta_n$ with probability function $f(x_i; \theta_1, \theta_2, \dots, \theta_n)$

Then (1) the joint probability density (or distribution) function of x_1, x_2, \dots, x_n would be

$$L(\theta_1, \theta_2, \dots, \theta_n) = \prod_{i=1}^n f(x_i, \theta_1, \dots, \theta_n)$$

(2) if $(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n)$ maximizes the likelihood function, the $\hat{\theta}_i$ is the maximum likelihood estimation of θ_i .

8.1.2 Normal Parameter

Question: Let $x \sim N(\mu, \sigma^2)$. Find maximum likelihood estimator of mean μ and variance σ^2 .

Answer: The prob. density function here is a function of $\theta_1 = \mu$ and $\theta_2 = \sigma^2$, such that

$$f(x_i; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{(x_i - \theta_1)^2}{2\theta_2}\right] \text{ for } -\infty < \theta_1 < \infty \text{ and } 0 < \theta_2 < \infty.$$

Now, the likelihood function would be

$$\begin{aligned} L(\theta_1, \theta_2) &= \prod_{i=1}^n f(x_i, \theta_1, \theta_2) \\ &= \theta_2^{-n/2} (2\pi)^{-n/2} \exp\left[-\frac{1}{2\theta_2} \sum_{i=1}^n (x_i - \theta_1)^2\right] \end{aligned}$$

Therefore, the log likelihood function:

$$\log L(\theta_1, \theta_2) = -\frac{n}{2} \log \theta_2 - \frac{x}{2} \log(2\pi) - \frac{\sum (x_i - \theta_1)^2}{2\theta_2}$$

Now, take the partial derivative of the log likelihood w.r.t θ_1 , and equate to

$$\begin{aligned} \frac{\partial \log(\theta_1, \theta_2)}{\partial \theta_1} &= -2 \frac{\sum (x_i - \theta_1)(-1)}{2\theta_2} \\ &= \frac{\sum (x_i - \theta_1)}{\theta_2} = 0. \end{aligned}$$

Now, by multiply by θ_2 , we get

$$\sum x_i - n\theta_1 = 0 \cdot \hat{\theta}_1 = \sum \frac{x_i}{n} = \hat{\mu} = \tilde{x}$$

Now, for θ_2 , take the partial derivative w.r.t θ_2 and set to 0 .

$$\frac{\partial L(\theta_1, \theta_2)}{\partial \theta_2} = \frac{-n}{2\theta_2} + \frac{\sum (x_i - \theta_1)^2}{2\theta_2^2} = 0.$$

multiples by $2\theta_2^2$

$$\begin{aligned}\frac{\partial(\theta_1, \theta_2)}{\partial \theta_2} &= \left(-\frac{n}{2\theta_2^2} + \frac{\sum(n_1 - \theta_1)^2}{2\theta_2^2} \right) = 0 \\ &= -n\theta_2 + \sum(x_i - \theta_1)^2 = 0.\end{aligned}$$

Now, solve for θ_2

$$\hat{\theta}_2 = \hat{\sigma}_2 = \frac{\sum(x_1 - \bar{x})^2}{n}$$

Therefore, we have shown that the maximum likelihood estimate of normal distribution one:

$$\hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$$

$$\hat{\sigma}^2 = \sum \frac{(x_i - \bar{x})^2}{n}$$

Now, the question arises that if these are the unbiased estimator of μ and σ^2 .

Let's see what is an unbiased estimator.

Question: If $x \sim N(\mu, \sigma^2)$ then prove

$$\begin{aligned}\hat{\mu} &= \frac{\sum x_i}{n} = \bar{x} \\ \tilde{\sigma}^2 &= \frac{\sum(x_i - \bar{x})^2}{n-1}\end{aligned}$$

are the unbiased of estimates.

Prof: Recall if $x_i \sim N(\mu, \sigma)$

$$\begin{aligned}\text{then } E(x_i) &= \mu \\ v(x_i) &= \sigma^2 \\ \therefore E(\bar{x}) &= E\left(\frac{1}{n} \sum x_i\right) \\ &= \frac{1}{n} \times n\mu = \mu. \\ \therefore E(\bar{x}) &= \mu.\end{aligned}$$

Therefore, the maximum likelihood estimator of μ is unbiased.

8.1.3 Sample variance

$$\begin{aligned}s^2 &= \hat{\sigma}^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \\ \Sigma(x_i - \bar{x})^2 &= \sum (x_i - \mu + \mu - \bar{x})^2 \\ &= \Sigma(x_i - \mu)^2 - n(\bar{x} - \mu)^2 \\ &= V(x_i) - x / (\text{Var}/\bar{x}) \\ &= n\sigma^2 - n\frac{\sigma^2}{n} \\ &= (n-1)\sigma^2\end{aligned}$$

Note that if we estimate variance by n instead of $n - 1$, we don't obtain an unbiased estimator. Therefore, what is the bias of the sample variance when we divide by n .

$$\begin{aligned}
S_{\text{bias}}^2 &= \frac{1}{n} (x_i - \bar{x})^2 \\
&= \frac{n-1}{n} \cdot \frac{1}{n-1} (x_i - \bar{x})^2 \\
&= \frac{n-1}{n} s^2 \\
E(s_{\text{biased}}^2) &= \frac{n-1}{n} E(s^2) \\
&= \left(1 - \frac{1}{n}\right) \sigma^2 \\
\text{by def. bias} &= E(\hat{\theta}) \equiv \theta \\
&= \left(1 - \frac{1}{n}\right) \sigma^2 - \sigma^2 \\
&= -\frac{1}{n} \sigma^2
\end{aligned}$$

As n tends to infinity, bias approaches to zero.

8.1.4 Criteria for selection of estimator:

(1) Unbiasedness:

A statistic of is $\hat{\theta}$ is said to be unbiased estimator iff expected value of the random variable equals to the parameter θ .

$$E(x) = \theta$$

The bias of an estimator is defined as the deviation of the deviation of the expectation from the true value.

$$E(\hat{\theta}) - \theta = \text{bias.}$$

The smaller the bias in a estimate the move preferable.

(2) Consistency. An estimator is consistent if estimate $\hat{\theta}$ it constructs is guarantees to convey to the true parameter value 0 w.r.t. quanlify of data to which it applied increases.

$$P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

(3) Variance (efficiency)

There may exist several loss and unbiased estimates for same parameter. For example in sample from no dirt $N(\mu, \sigma^2)$ where σ^2 is known, sample mean (\bar{x}), sample (\bar{x}') are both consistent estimate

of μ , which is same as population median. Thus, there is necessary of some further criterion which enable us to choose between the estimates which common property of consistency. A statistic $\hat{\theta}_1$ is said to be more efficient unbiased estimator of the parameter θ than $\hat{\theta}_2$ if

- (a) both are unbiased estimator
- (b) variance of $\hat{\theta}_1 < \text{Var } \hat{\theta}_2$.

Let X_1, X_2, \dots, X_n and $N(\mu, \sigma^2)$ with σ^2 known. For large n sample mean (\bar{X}) such that sample median (\tilde{X}) are both consistent estimate of μ .

$$\begin{aligned} v(\bar{x}) &= \sigma^2/n \\ V(\tilde{x}) &= \pi\sigma^2/2n \\ &= 1.57\sigma^2/n \end{aligned}$$

$$v(\bar{X}) < v(\tilde{x})$$

We conclude that for normal distribution sample mean is μ more efficient estimator of μ than the sample media for large samples.

(3) Sample standard deviation is always biased estimator of the population standard.

Prof by contradiction.

Suppose if possible $E(s) = \sigma$

$$\begin{aligned} \text{var}(S) &= E(s^2 - E(S))^2 \\ &= \sigma^2 - \sigma^2 \\ &= 0 \end{aligned}$$

This is contradiction as $\text{var}(S)$ means S is a constant but is a variable.

Here $E(s) \neq \sigma$.

8.2 Statistical Inference ...

Statistical inference is the process by which we infer population properties from sample properties.

There are two types of statistical inference:

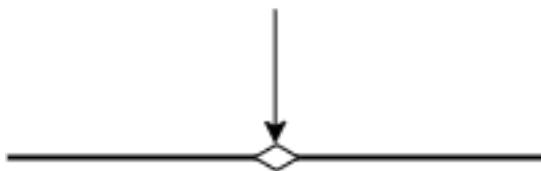
- Estimation
- Hypotheses Testing

The concepts involved are actually very similar, which we will see in due course. Below, we provide a basic introduction to estimation.

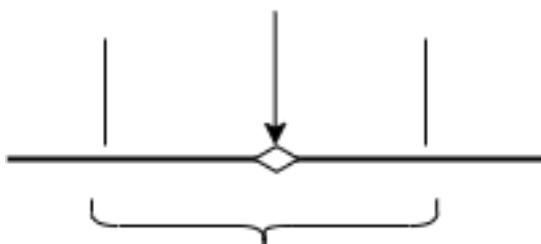
8.3 Estimation ...

The objective of estimation is to approximate the value of a population parameter on the basis of a sample statistic. For example, the sample mean \bar{X} is used to estimate the population mean μ . There are two types of estimators:

- Point Estimator

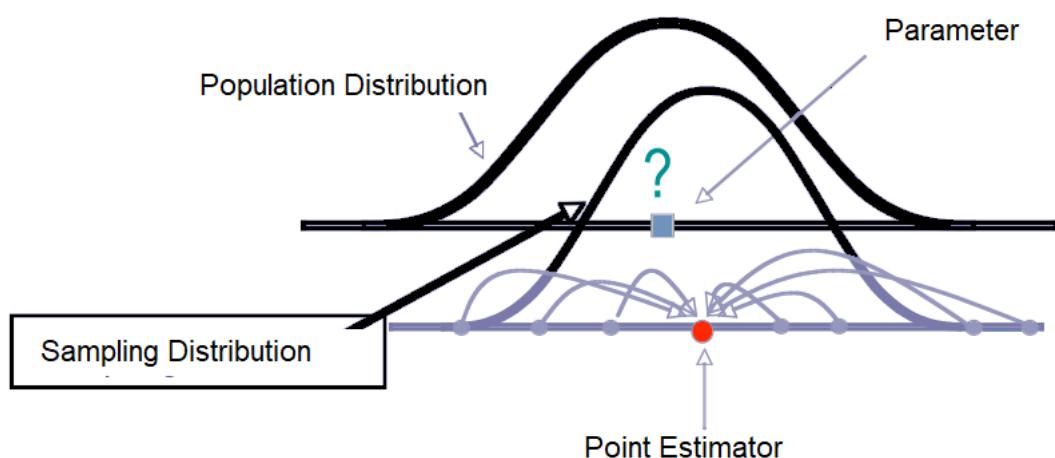


- Interval Estimator



8.3.1 Point Estimator

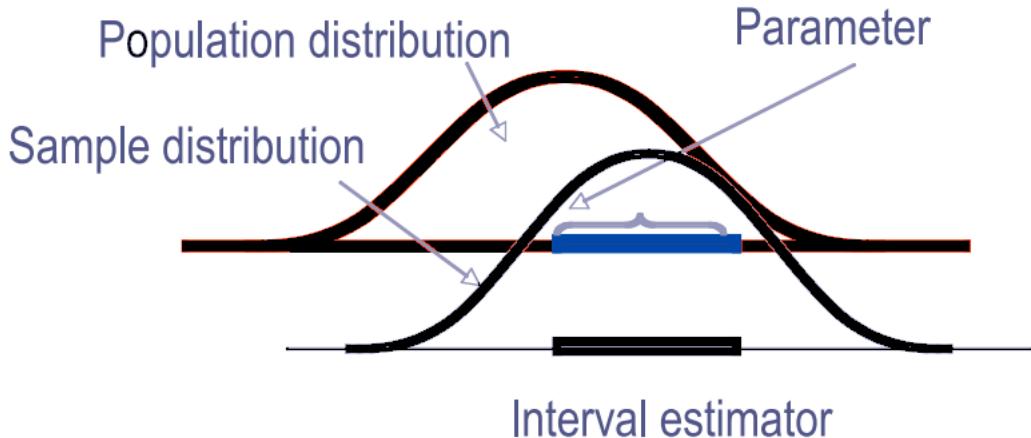
A point estimator draws inferences about a population by estimating the value of an unknown parameter using a single value or point.



Recall that for a continuous variable, the probability of assuming any particular value is zero. Hence, we are only trying to generate a value that is close to the true value. Point estimators typically do not reflect the effects of larger sample sizes, while interval estimator do.

8.3.2 Interval Estimator

An interval estimator draws inferences about a population by estimating the value of an unknown parameter using an interval. Here, we try to construct an interval that "covers" the true population parameter with a specified probability.



Interval estimator

As an example, suppose we are trying to estimate the mean summer income of students. Then, an interval estimate might say that the (unknown) mean income is between \$380 and \$420 with probability 0.95.

8.4 Quality of Estimators

The desirability of an estimator is judged by its characteristics. Three important criteria are:

- Unbiasedness
- Consistency
- Efficiency

8.4.1 Unbiasedness

An unbiased estimator of a population parameter is an estimator whose expected value is equal to that parameter. Formally, an estimator $\hat{\mu}$ for parameter μ is said to be unbiased if:

$$E(\hat{\mu}) = \mu.$$

Example: The sample mean \bar{X} is an unbiased estimator for the population mean μ , since

$$E(\bar{X}) = \mu.$$

It is important to realize that other estimators for the population mean exist: maximum value in a sample, minimum value in a sample, average of the maximum and the minimum values in a sample
...

Being unbiased is a minimal requirement for an estimator. For example, the maximum value in a sample is not unbiased, and hence should not be used as an estimator for μ .

8.4.2 Consistency

An unbiased estimator is said to be consistent if the difference between the estimator and the target population parameter becomes smaller as we increase the sample size. Formally, an unbiased estimator $\hat{\mu}$ for parameter μ is said to be consistent if $V(\hat{\mu})$ approaches zero as $n \rightarrow \infty$.

Note that being unbiased is a precondition for an estimator to be consistent.

Example 1: The variance of the sample mean \bar{X} is σ^2/n , which decreases to zero as we increase the sample size n . Hence, the sample mean is a consistent estimator for μ .

8.4.3 Efficiency

Suppose we are given two unbiased estimators for a parameter. Then, we say that the estimator with a smaller variance is more efficient.

Example 1: For a normally distributed population, it can be shown that the sample median is an unbiased estimator for μ . It can also be shown, however, that the sample median has a greater variance than that of the sample mean, for the same sample size. Hence, \bar{X} is a more efficient estimator than sample median.

Example 2: Consider the following estimator. First, a random portion of a sample is discarded from an original sample; then, the mean of the retained values in the sample is taken as an estimate for μ . This estimator is unbiased, but is not as efficient as using the entire sample. The intuitive reasoning is that we are not fully utilizing available information, and hence the resulting estimator has a greater variance.

8.5 Estimating μ When σ^2 is Known ...

Constructing point estimates using the sample mean \bar{X} is the "best" (according to our criteria above) estimator for the population mean μ .

Suppose the variance of a population is "known." How does one construct an interval estimate for μ ?

The key idea is that from the central limit theorem, we know that when n is sufficiently large, the standardized variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

follows the standard normal distribution. It is important to realize that this is true even though we do not know the value of μ . The value of σ , however, is assumed to be given (this assumption, which could be unrealistic, will be relaxed later).

8.6 It follows that for a given α , we have

$$P\left(\mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} \leq \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Since our "unknown" is actually μ , the above can be rearranged into:

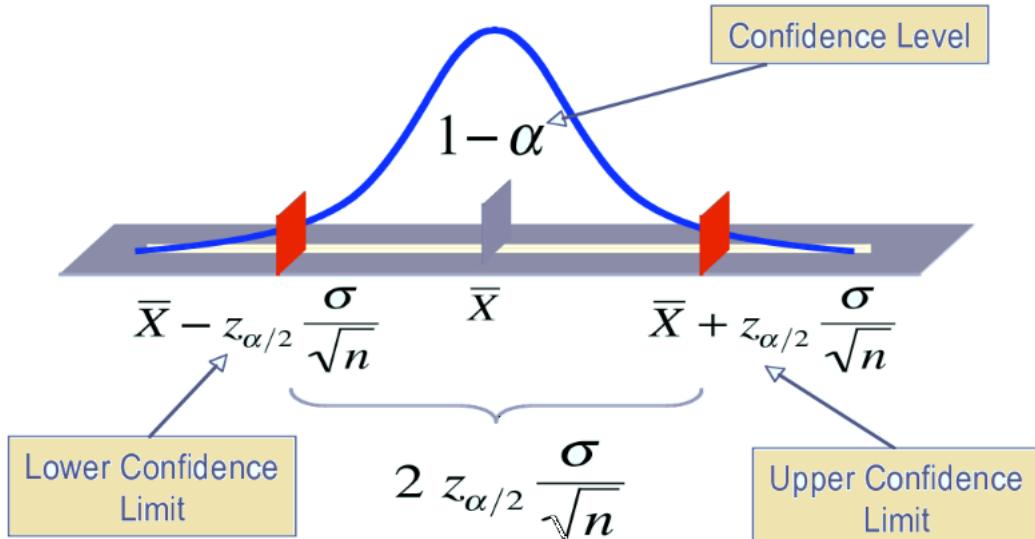
$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

That is, the probability for the interval

$$\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

to contain, or to cover, the unknown population mean μ is $1 - \alpha$; and we now have a so-called confidence interval for μ . Note that the interval estimator (2) is constructed from \bar{X} , $z_{\alpha/2}$, σ , and n , all of which are known. The user-specified value $1 - \alpha$ is called the confidence level or coverage probability.

Pictorially, we have



8.7 Interpretation:

If the interval estimator (2) is used repeatedly to estimate the mean μ of a given population, then $100(1 - \alpha)\%$ of the constructed intervals will cover μ .

The often-heard media statement "19 times out of 20" refers to a confidence level of 0.95. Such a statement is good, since it emphasizes the fact that we are correct only 95% of the time.

Example: Demand during Lead Time

A computer company delivers computers directly to customers who order via the Internet. To reduce inventory cost, the company employs an inventory model. The model requires information about the mean demand during delivery lead time between a central manufacturing facility and local warehouses. Past experience indicates that lead-time demand is normally distributed with a standard deviation of 75 computers per lead time (which is also random).

Construct the 95% confidence interval for the mean demand. Demand data for a sample of 25 lead-time periods are given in the file Xm10-01.xls.

Solution: Since $1 - \alpha = 0.95$, we have $\alpha = 0.05$ and hence $\alpha/2 = 0.025$, for which $z_{0.025} = 1.96$. From the given data file, we obtain the sample mean $\bar{X} = 370.16$. The confidence interval is therefore (see (2))

$$\left(370.16 - 1.96 \frac{75}{\sqrt{25}}, 370.16 + 1.96 \frac{75}{\sqrt{25}} \right)$$

or simply (340.76, 399.56).

8.8 Width of Confidence Interval ...

Suppose we are told that with 95% confidence that the average starting salary of accountants is between \$15,000 and \$100,000. Clearly, this provides little information, despite the high "confidence" level.

Now, suppose instead: With 95% confidence that the average starting salary of accountants is between \$42,000 and \$45,000.

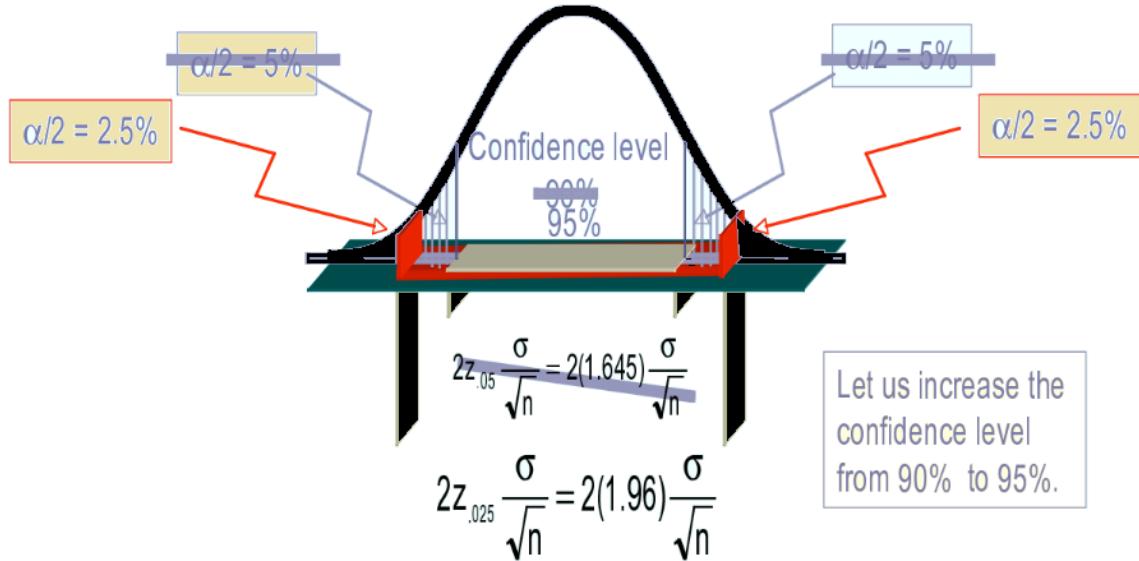
The second statement of course offers more precise information. Thus, for a given α , the width of a confidence interval conveys the extent of precision of the estimate. To reduce the width, or to increase precision, we can increase the sample size. In general, recall that the upper and lower confidence limits are:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

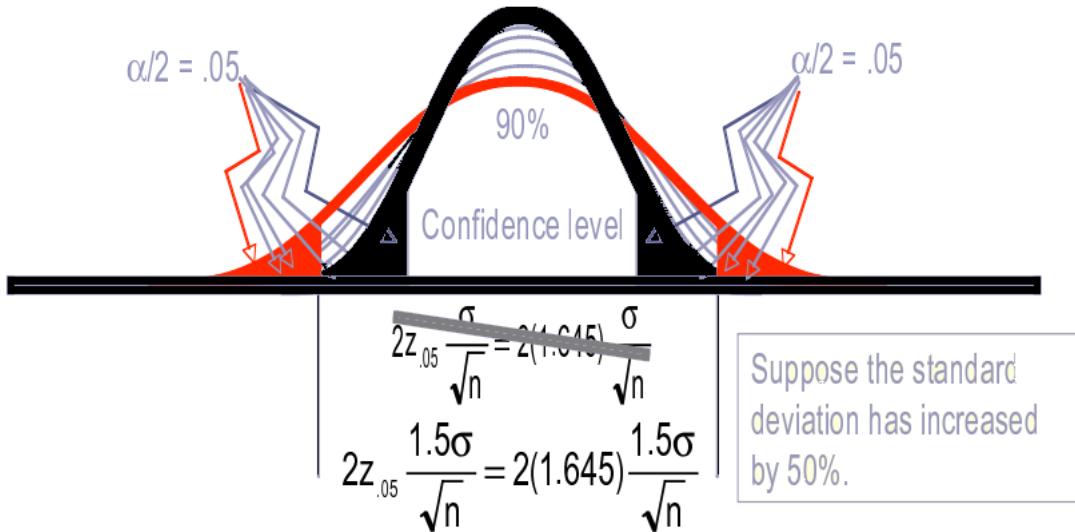
Hence, the width of the confidence interval is $2z_{\alpha/2}\sigma/\sqrt{n}$. It follows that precision depends on α , σ , and n .

8.9 Details ...

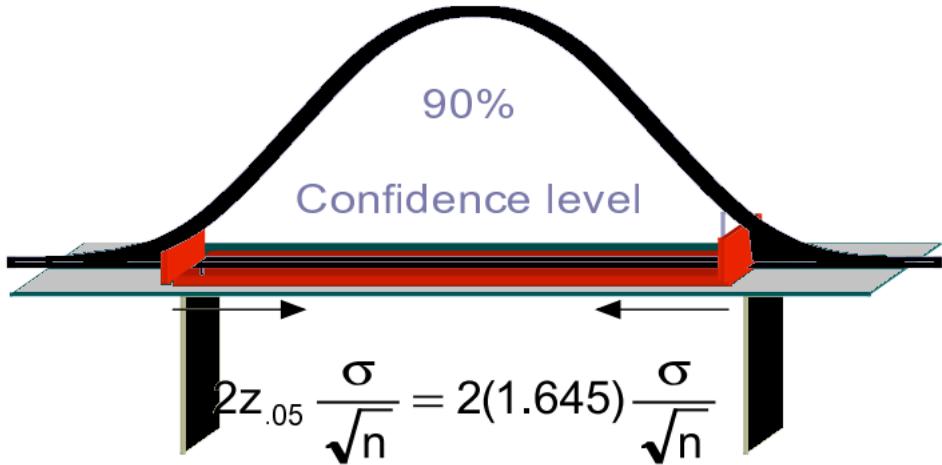
A smaller α implies a wider interval:



A larger σ implies a wider interval:



A larger n implies a narrower interval:



8.10 Selecting the Sample Size ...

To control the width of the confidence interval, we can choose a necessary sample size. Formally, suppose we wish to "estimate the mean to within w units." This means that we wish to construct an interval estimate of the form $\bar{X} \pm w$.

By solving the equation

$$w = z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

we obtain the required sample size.

$$n = \left(\frac{z_{\alpha/2}\sigma}{w} \right)^2$$

Example: Tree Diameters

A lumber company must estimate the mean diameter of trees in an area of forest to determine whether or not there is sufficient lumber to harvest. They need to estimate this to within 1 inch at a confidence level of 99%. Suppose the tree diameters are normally distributed with a standard deviation of 6 inches. What sample size is sufficient to guarantee this?

Solution: The required precision is ± 1 inch. That is, $w = 1$. For $\alpha = 0.01$, we have $z_{\alpha/2} = z_{0.005} = 2.575$. Therefore,

$$n = \left(\frac{z_{\alpha/2}\sigma}{w} \right)^2 = \left(\frac{2.575 \cdot 6}{1} \right)^2 = 239.$$

Thus, we need to sample at least 239 trees to achieve a 99% confidence interval of $\bar{X} \pm 1$.

Chapter 9

Sampling Theory and Hypothesis Testing

9.1 Sampling

Population :

Totality of the observations with which we are concerned i.e. a group that includes all the cases (individuals, objects, or groups) *i.e* population is the entire group that we want to draw conclusions about.

For example: If there are 600 students in the school, or stream flow in a certain stream over infinite timeline.

Sample :

A relatively small subset from a population or specific group of individuals that we want to collect data from. For example: A sample of 30 students or stream flow in the stream over the last 30 years.

9.1.1 Statistical Inference

To make decision or to draw conclusions about a population by utilizing the information contained in a sample.

Example: In attempting to determine the average length of life a certain brand of light bulb, it would be impossible to test all such bulbs . So we draw a sample and estimate its value. That is a parameter estimation problem (mean life of light bulb). However in many situations we require to decide whether to accept or reject a statement about some parameter. The statement is called a hypothesis, and the decision making procedure about the hypothesis is called hypothesis testing. Here we will assume parameter value and try to see if it is correct.

Statistical inference is divided into two major areas:

9.1.2 Parameter estimation

The objective of estimation is to determine the **approximate value** of a population parameter on the basis of a sample statistic.

What is normal body temperature? Take a sample and estimate its value.

Hypothesis testing

Given point estimator(s) from samples, we may wish to infer about the results, or if any statistical differences exist.

What is normal body temperature? Is it actually 37.6°C (on average)?

9.1.3 Notation

The table for notation is given below :

Sample and Population Notations		
Measure	Sample Notation	Population Notation
Mean	\bar{Y}	μ_Y
Proportion	p	π
Standard deviation	S_Y	σ_Y
Variance	S_Y^2	σ_Y^2

Sampling :

- **Parameter**

A measure (for example, mean or standard deviation) used to describe a population distribution.

- **Statistic**

A measure (for example, mean or standard deviation) used to describe a sample distribution.

Examples of statistics :

- **Sample mean**

The average cost of a gallon of gas in the US is \$2.65 .

- **Difference in means**

The difference in the average gas price in Los Angeles (\$2.96) compared with Des Moines, Iowa

(\$2.32) is 64 cents .

- **Proportion**

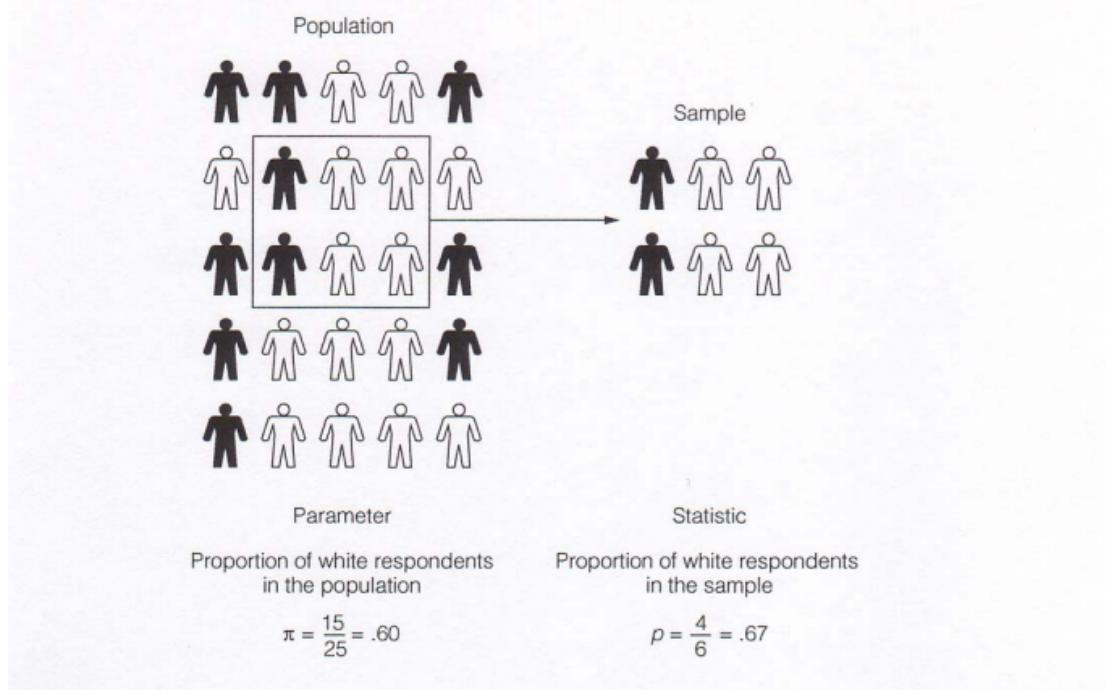
67% of high school students in the U.S. exercise regularly.

- **Difference in proportions**

The difference in the proportion of Democrats who approve of Obama (83%) versus Republicans who do (14%) is 69%.

Sampling: Parameter & Statistic

Figure 11.1 The Proportion of White Respondents in a Population and in a Sample



9.1.4 Random Sampling

- **Simple Random Sample**

A sample designed in such a way as to ensure that (1) every member of the population has an equal chance of being chosen and (2) every combination of N members has an equal chance of being chosen.

This can be done using a computer, calculator, or a table of random numbers.

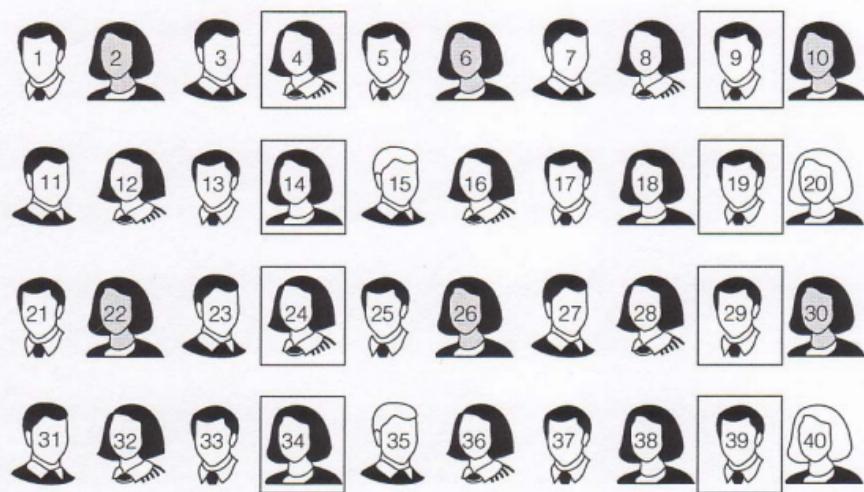
- **Systematic random sampling**

A method of sampling in which every Kth member (K is a ratio obtained by dividing the population size by the desired sample size) in the total population is chosen for inclusion in

the sample after the first member of the sample is selected at random from among the first K members of the population.

Figure 11.2 **Systematic Random Sampling**

From a population of 40 students, let's select a systematic random sample of 8 students. Our skip interval will be 5 ($40 \div 8 = 5$). Using a random number table, we choose a number between 1 and 5. Let's say we choose 4. We then start with student 4 and pick every 5th student:



Our trip to the random number table could have just as easily given us a 1 or a 5, so all the students do have a chance to end up in our sample.

Sampling Distributions

- **Sampling error**

The discrepancy between a sample estimate of a population parameter and the real population parameter.

- **Sampling distribution**

A theoretical distribution of all possible sample values for the statistic in which we are interested.

- **Sampling distribution of the mean**

A theoretical probability distribution of sample means that would be obtained by drawing from the population all possible samples of the same size.

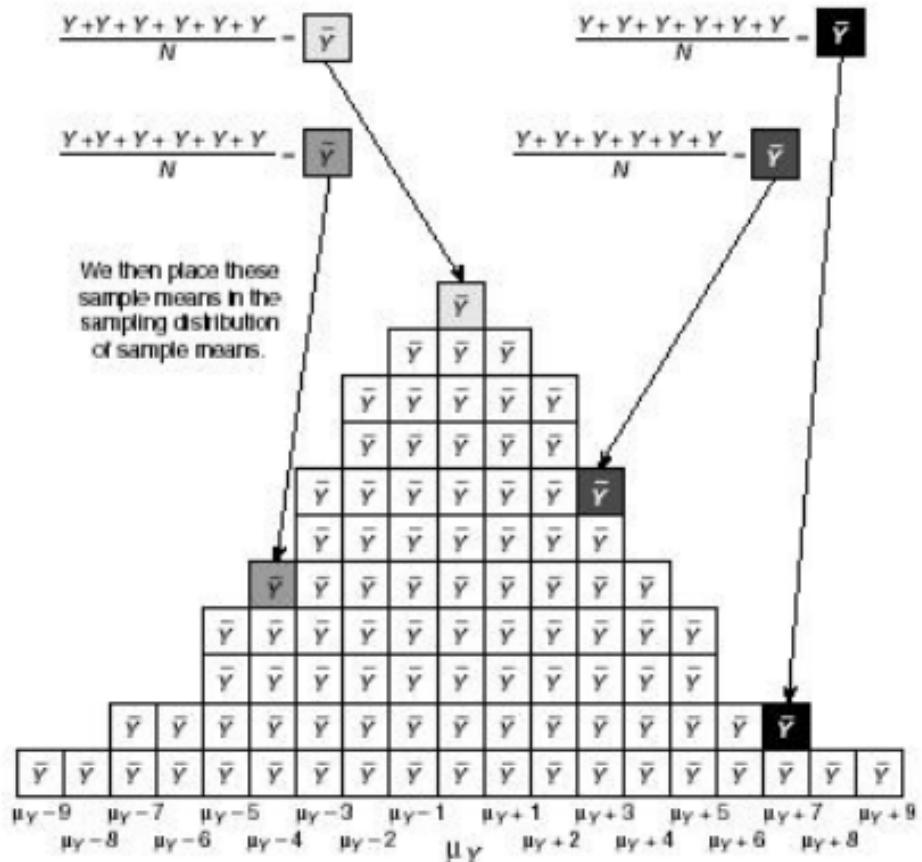
If we repeatedly drew samples from a population and calculated the sample means, those sample means would be normally distributed (as the number of samples drawn increases.)

- **Standard error of the mean**

The standard deviation of the sampling distribution of the mean. It describes how much dispersion there is in the sampling distribution of the mean.

Figure 10.5 Generating the Sampling Distribution of the Mean

From a population (with a population mean of μ_Y) we start drawing samples and calculating the means for those samples:



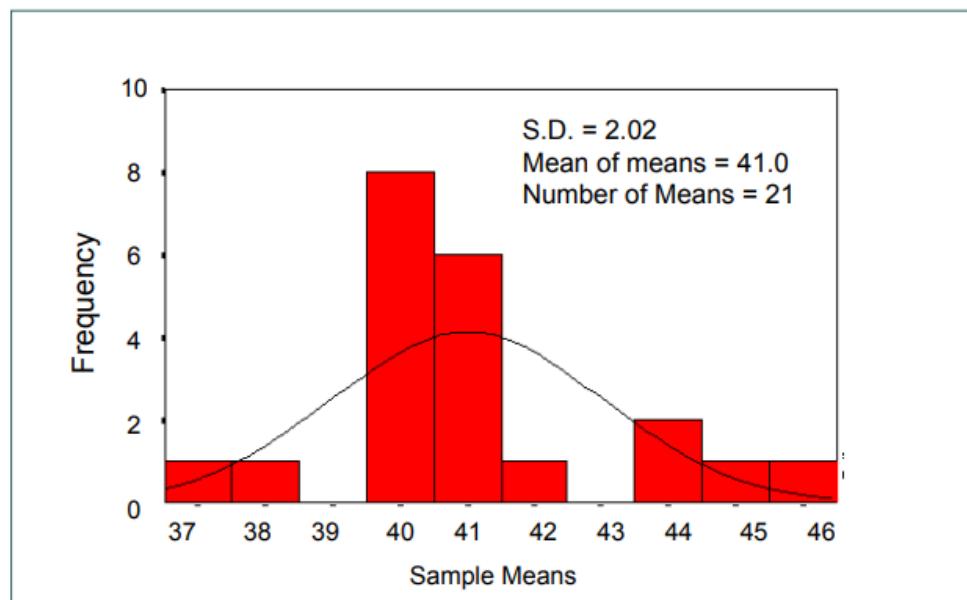


Fig. 9.1: Distribution of Sample Means with 21 Samples

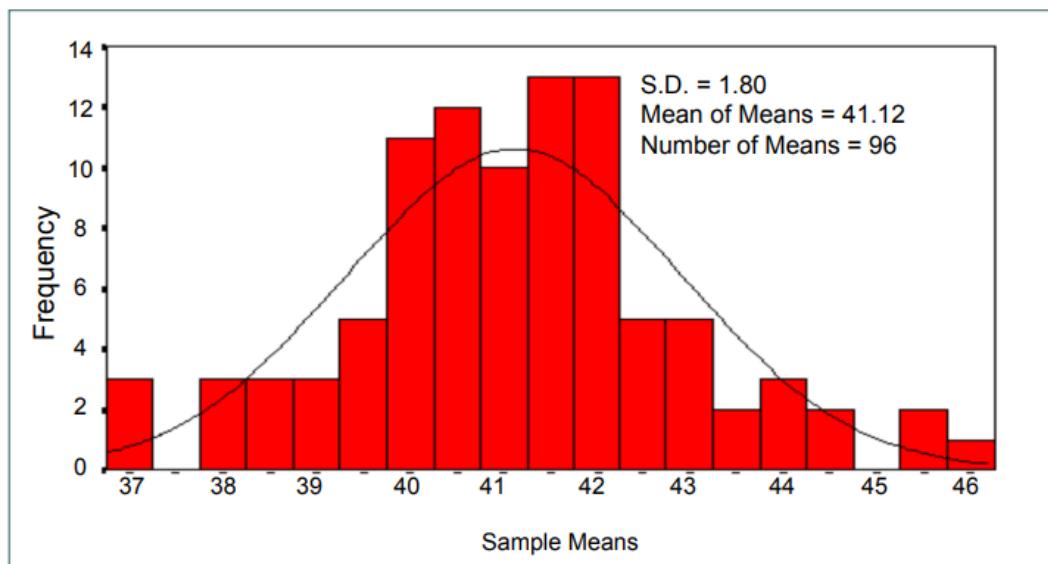


Fig. 9.2: Distribution of Sample Means with 96 Samples

9.1.5 Law of Large Numbers

If one draws independent samples from a population with mean μ , then as the number of observations increases, the sample mean \bar{x} gets closer and closer to the population mean μ .

Example: Population: seasonal home-run totals for 7032 baseball players from 1901 to 1996

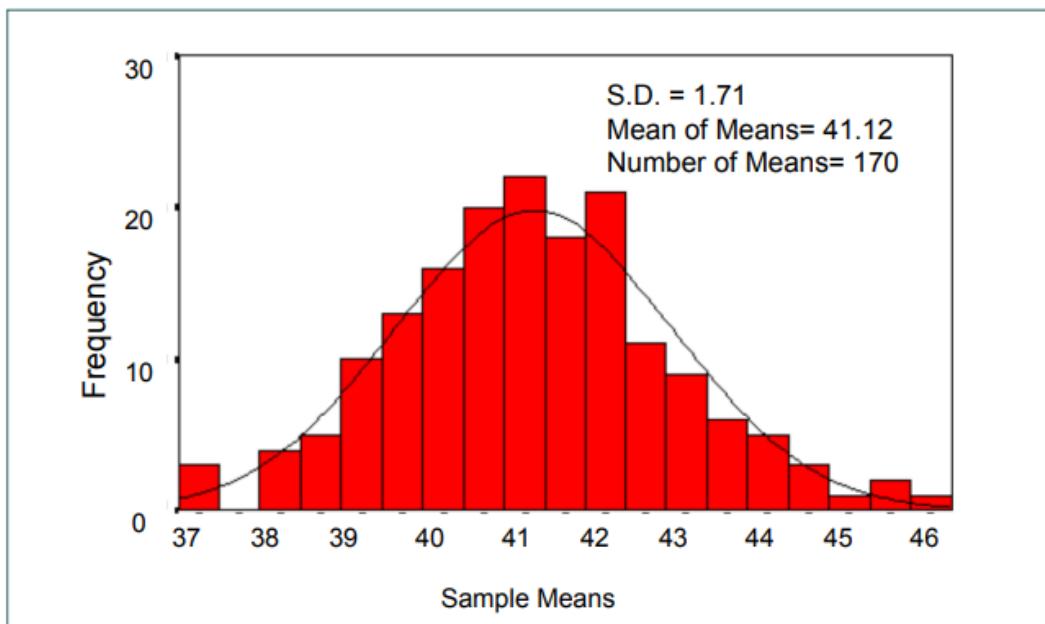


Fig. 9.3: Distribution of Sample Means with 170 Samples

Sample Size	Mean	Variance
100 samples of size $n = 1$	3.69	46.8
100 samples of size $n = 10$	4.43	4.43
100 samples of size $n = 100$	4.42	0.43
100 samples of size $n = 1000$	4.42	0.06
Population Parameter	$\mu = 4.42$	

9.1.6 The Central Limit Theorem

If all possible random samples of size N are drawn from a population with mean μ and a standard deviation s , then as N becomes larger, the sampling distribution of sample means becomes approximately normal, with mean μ and standard deviation $= \frac{s}{\sqrt{N}}$.

If X_1, X_2, \dots, X_n are observation from a population with mean μ and variance σ^2 , then the Central Limit Theorem says,

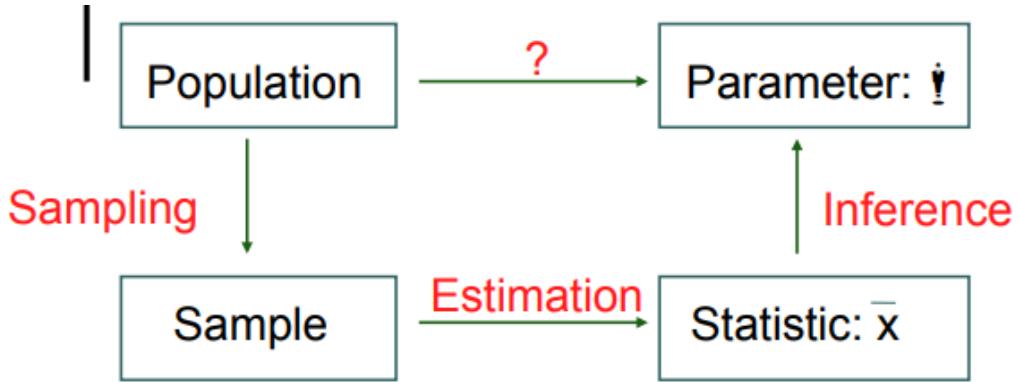
$$\hat{\mu} = \bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$E(\hat{X}) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} n\mu = \mu$$

$$Var(\bar{X}) = Var\left(\sum_{i=1}^n \frac{X_i}{n}\right) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \implies \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

9.2 Inference with Sample Mean



Sample mean is our estimate of population mean.

Example: (Central limit theorem)

Question: An electronics company manufactures resistors that have a mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance is normal. Find the probability that a random sample of $n=25$ resistors will have an average resistance less than 95 ohms.

Note that the sampling distribution of X is normal, with mean $\mu_{\bar{X}} = 100$ ohms and a standard deviation of

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

Therefore, the desired probability corresponds to the shaded area in the figure. Standardizing the point $\bar{X} = 95$ in figure, we find that

$$z = \frac{95 - 100}{2} = -2.5$$

and therefore,

$$\mathcal{P}(\bar{x} \leq 95) = \mathcal{P}(Z \leq -2.5) = 0.0062$$

9.3 Estimating the difference between two mean

Properties of the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$, the difference two sample means:

When independent random samples of n_1 and n_2 observations have been selected from population with means and variances s_1^2 and s_2^2 respectively, the sampling distribution of the difference will have

the following properties: 1) The mean and standard deviation of $(\bar{x}_1 - \bar{x}_2)$ will be

$$m_{(\bar{x}_1 - \bar{x}_2)} = m_1 - m_2$$

and

$$s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example:

Question: Two independent experiments are being run in which two different types of Paints are compared. Eighteen specimens are painted using type A and the drying time, In hours, is recorded on each. The same is done with type B. The population standard deviations are both known to be 1.0. Assuming that the mean drying time is equal for the two types of paint, find $\mathcal{P}(X_A - X_B > 1)$, where X_A and X_B are average drying times for samples of size $n_A = n_B = 18$.

9.4 Confidence interval (CI)

Definition -

Probability that a population parameter will fall between a set of values for a certain proportion of times.

Formula for CI is given below -

$$CI = \bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

where -

\bar{x} = Sample mean

z = Confidence interval value

σ = sample standard deviation

n = sample size

Need of CI -

Confidence interval provides information about the relationship between the actual statistic for a given simulation and the expected value of the parameter. Basically it represent how "good" an estimate is. We can understand it with the example if in a sample of 20 chocolate bars the amount of calories has been measured and it is found that the sample mean is 224 calories. Then how certain are we that the population mean is close to 224? Here confidence interval is helpful.

Classification of confidence interval is shown in above figure. First one is based on population mean which is also of two types based on the knowledge of σ (known and unknown) and the second one is population proportion. -The central point is the true population mean and is the most likely sample mean.

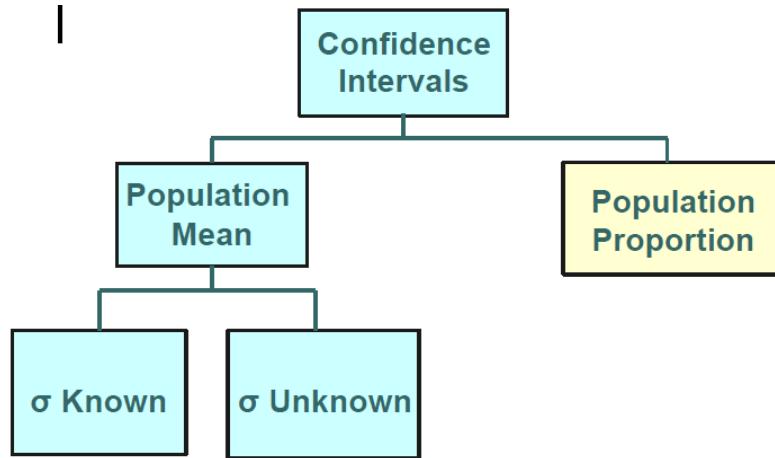


Fig. 9.4: Classification of Confidence interval

9.4.1 Population Mean

Distribution of confidence interval is shown in above figures. From these figures we can infer following things-

- The larger the standard error the flatter the curve.
- 95% of all possible sample means fall within a range of about two standard errors from the mean (1.96 to be precise).
- A confidence level of $\alpha = 95\%$ implies that 95% of all the samples would give an interval that includes μ , or whatever other parameter is estimated.

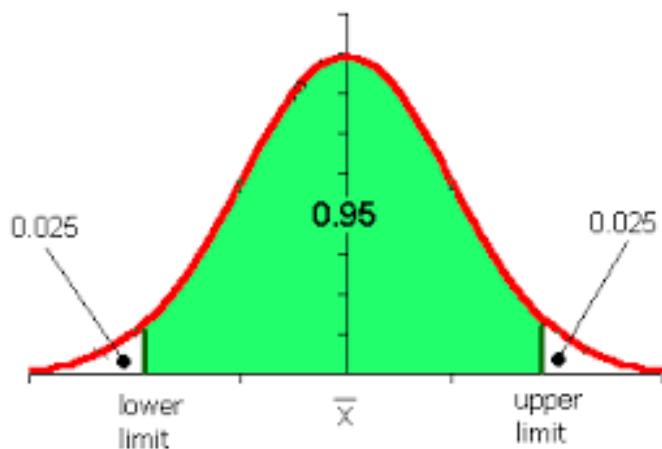


Fig. 9.5: Confidence interval distribution

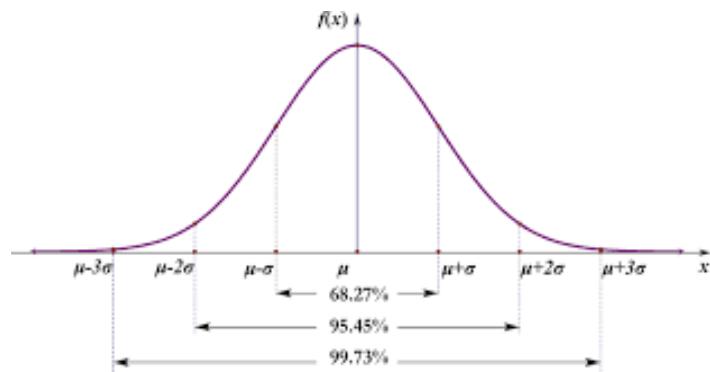


Fig. 9.6: Confidence interval distribution

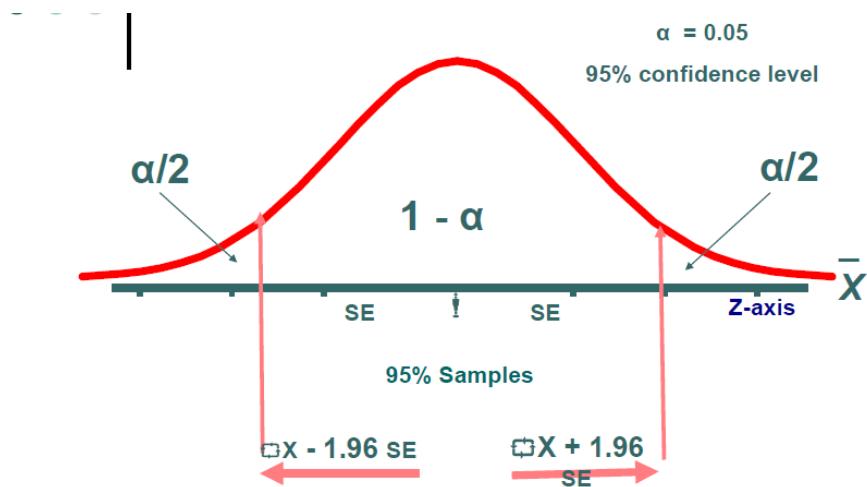


Fig. 9.7: Confidence interval distribution

Desired Confidence Interval	Z Score
90%	1.645
95%	1.96
99%	2.576

Fig. 9.8: Confidence level and corresponding z - values

9.4.1.1 σ - known

Two-Sided z-Interval

If an experimenter wishes to construct a confidence interval for a population mean based on a sample of size n with a sample mean and using an assumed known value for the population standard deviation, then the appropriate confidence interval is

$$\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

which is known as a two-sided z-interval or variance known confidence interval.

Note -All of these confidence intervals were calculated using the normal distribution, which is justified by the central limit theorem

- However, by doing this, we are making the assumptions that the sample size is large and the population variance is known!
- We will see later how to calculate confidence intervals when these assumptions are not true.

Confidence interval for known variance is a results of the Central Limits Theorem. The underlying assumptions:

- sample size $n \geq 30$, or
- the corresponding random var. is (approx.) normally distributed $(1 - \alpha)100\%$ confidence interval :
- typical values of α : $\alpha = 10\%$, $\alpha = 5\%$, $\alpha = 1\%$
- what happens to the length of CI when the confidence level increases?
- what happens to the length of the CI when n increases?

- The length of two-sided interval is

$$L = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

-As the standard error \hat{m} of decreases, so that $\hat{m} = \bar{x}$ becomes a more “accurate” estimate of m .

-The length of a confidence interval also depends upon the confidence level. As the confidence level increases, the length of the confidence interval also increase.

-For a fixed critical point, a confidence interval length L is inversely proportional to the square root of the sample size n .

Ex: Question -

-A sample of 41 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is .35 ohms. Determine a 95% confidence interval for the true mean resistance of the population.

Solution:

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 2.20 \pm 1.96 \left(\frac{0.35}{\sqrt{41}} \right) = 2.20 \pm 0.1071$$

Margin of Error(e) :

The amount added and subtracted to the point estimate to form the confidence interval.

$$\boxed{\text{Parameter} = \text{Estimate} \pm \text{Margin of Error}}$$

Example: Margin of error for estimating μ , σ known:

$$\boxed{e = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}$$

then,

$$\boxed{CI = \bar{x} \pm e}$$

Question :

Before we sample our Chips Ahoy cookies bags, we want to decide the minimum number of bags needed for a certain margin of error (saves on cookies). If we want a confidence level of 95% (so $Z^* = 1.96$) and we want a margin of error less than 100 chips, then how many minimum number of bags of chips are needed? ($\sigma = 18.1$)

Solution :

We know that

$$e = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \text{ then } n = \left(z_{\frac{\alpha}{2}} \frac{\sigma}{e} \right)^2 = \left(\frac{1.96 * 18.1}{100} \right)^2 = 5.3$$

so we need to sample at least 6 bags of chips.

Determination of Sample Size

The required sample size can be found to reach a desired margin of error (e) and level of confidence ($1 - \alpha$) Required sample size, σ known:

$$\boxed{n = \left(\frac{z_{\frac{\alpha}{2}} * \sigma}{E} \right)^2}$$

Question : If $\sigma = 45$, what sample size is needed to be 90% confident of being correct within ± 5 ?

Solution : $\sigma = 45$, $z^* = 1.645$, $E = 5$ then,

$$n = \left(\frac{1.645 * 45}{100} \right)^2 = 219.19$$

Hence $n = 220$.

Estimation of standard errors

Since the standard error is generally not known, we usually work with the estimated standard error:

$$S_{\bar{x}} = \frac{S_x}{\sqrt{N}}$$

9.4.1.2 σ - unknown

Large Sample CI for Population Mean

Key assumption: if n is sufficiently large, then CLT allows us to approximate a $100(1-\alpha)\%$ CI for a sample mean. (replace s with sample std dev, S)

$$z = \frac{\bar{X} - m}{\frac{S}{\sqrt{n}}}$$

$$\bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{S}} \leq m \leq \bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{S}}$$

Large Sample CI is generally safe for $n > 40$

Calculation of Test Statistics

For one sample tests, use Z test statistic if population is Normal, m is known, or if sample size is large

$$Z_c = \frac{\hat{X} - m}{S_{\hat{X}}}$$

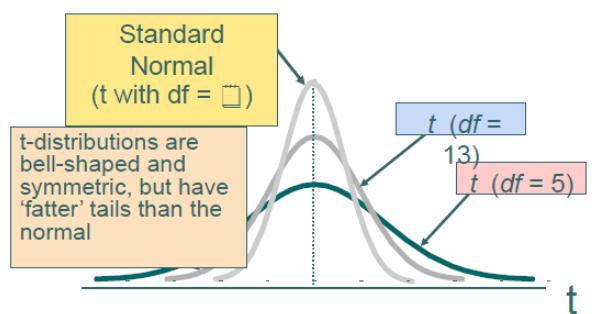
For one sample tests, use T static if population distribution is not known or if sample size is small (less than 30)

$$t_c = \frac{\hat{X} - \mu}{S_{\hat{X}}}$$

Test Statistic If the distribution is normal (or the sample size is larger than 30) and the standard deviation is known. Then

$$Z = \frac{\hat{x} - m}{\frac{s}{\sqrt{n}}}$$

Note: $t \rightarrow Z$ as n increases



t - distribution values

Number of components that are free to vary about a parameter $D_f = \text{Sample size} - \text{Number of parameters estimated}$ D_f is $n - 1$ for one sample test of mean

Confidence Level	t (10 d.f.)	t (20 d.f.)	t (30 d.f.)	Z
.80	1.372	1.325	1.310	1.28
.90	1.812	1.725	1.697	1.64
.95	2.228	2.086	2.042	1.96
.99	3.169	2.845	2.750	2.58

n	a			
	0.05	0.025	0.01	0.005
1	6.314	12.706	31.821	63.656
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
15	1.753	2.131	2.602	2.947
20	1.725	2.086	2.528	2.845
40	1.684	2.021	2.423	2.704
Infinity	1.645	1.960	2.326	2.576

Fig. 9.9: t - distribution table

One Sample Intervals

\bar{X} and S are based on sample data from a normal population.

CI for m :

$$\bar{X} \pm t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}}$$

We now use critical values from the t-distribution instead of the critical values from the Normal Distribution.

Question : Let \bar{x} be the mean and s the sample standard deviation of a sample of n observations from a normal distributed population with mean and unknown variance.

A $(1 - \alpha)100\%$ confidence interval is given by

$$\hat{x} - t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \leq m \leq \hat{x} + t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}}$$

• For $30 \leq n$ the standard normal distribution can be used instead of the t distribution.

Example

A random sample of $n = 25$ has $\bar{X} = 50$ and $S = 8$. Set up a 95% confidence interval estimate for m .

Solution

$$\begin{aligned}\hat{x} - t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} &\leq m \leq \hat{x} + t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \\ 0 - 2.0639 \frac{8}{\sqrt{25}} &\leq m \leq 50 + 2.0639 \frac{8}{\sqrt{25}} \\ 46.69 &\leq m \leq 53.3\end{aligned}$$

Example

Consider the following sample of fat content (in percentage) of $n = 10$ randomly selected hot dogs: 25.2, 21.3, 22.8, 17.0, 29.8, 21.0, 25.5, 16.0, 20.9, 19.5, construct the 95% confidence interval for population mean.

Solution

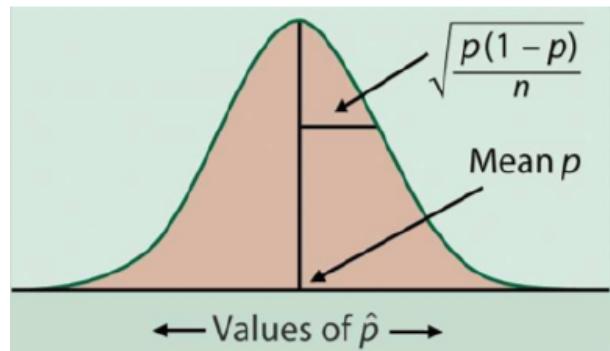
$$\begin{aligned}\hat{x} \pm t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} &= 21.90 \pm 1.96 \frac{4.134}{\sqrt{10}} \\ &= 21.90 \pm 2.56 = (18.94, 24.46)\end{aligned}$$

9.4.2 Proportion

For small samples, use the **Binomial distribution** to calculate probabilities for the sample count or sample proportion.

- Definition of “small”: $np \leq 10$ or $n(1-p) \leq 10$
- For large samples, we use the **Normal approximation** to the Binomial distribution for the sample count or sample proportion.

$$\text{mean}(\hat{p}) = p; \text{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$



Standard Error

$$SE(\text{Mean}) = \frac{S}{\sqrt{n}}; SE(p) = \sqrt{\frac{p(1-p)}{n}}$$

Point and interval estimation for Proportion

If $\mathcal{X} \sim \mathcal{B}(n, p)$, then $\hat{p} = \frac{X}{n}$

$\mathcal{X} \sim \mathcal{B}(n, p) \implies \mathcal{E}(\mathcal{X}) = np$,

hence $\mathcal{E}(\hat{p}) = \mathcal{E}\left(\frac{\mathcal{X}}{n}\right) = \frac{1}{n}\mathcal{E}(\mathcal{X}) = \frac{1}{n}np = p$

It means that \hat{p} is an unbiased estimate

$$\mathcal{V}(X) = np(1-p); \text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

Confidence Intervals for Proportions

Based on the sampling distribution, our confidence interval for the population proportion p is

$$p \pm \mathcal{Z}_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

However, this interval is not very useful, since it still depends on the unknown population proportion p .

We fix this by using our sample proportion in place of p , so our $100^*(1-\alpha)\%$ confidence interval for

$$\hat{p} \pm \mathcal{Z}_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Examples Question : In a random sample of $n=500$ families owing television sets in the city of Hamilton, Canada, it is found that $x=340$ subscribed to HBO. Find a 95% confidence interval for the actual proportion of families in this city who subscribe to HBO?

Ans : $0.64 \leq p \leq 0.72$

Question : How large sample is required if we want to be 95% confident that our estimate of p is within 0.02?

Ans : 2090

Question : A random sample of 400 voters showed 32 preferred candidate A. Set up a 95% confidence interval estimate for p .

Ans : $0.053 \leq p \leq 0.107$

9.5 What is Hypothesis Testing?

Making statistical decisions based on experimental data involves the statistical process of hypothesis testing. In essence, hypothesis testing is an assumption we make regarding the population parameter.

*Examples: According to climate scientists, recent years have seen an increase in global warming.
Hypothesis testing uses experimental data to demonstrate whether a claim is true or false.*

To prove such instances, we need a statistical method to prove the assumption we are asserting.

9.5.1 Why do we use it?

Hypothesis testing is an essential procedure in statistics. A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. When we say that a finding is statistically significant, it's thanks to a hypothesis test.

9.5.2 What are the basics of Hypothesis?

*The basic of hypothesis is **normalisation** and **standard normalisation**. all our hypothesis is revolve around basic of these 2 terms. let's see these.*

*In fig1, you can see there are different normal curve all those normal curve can have different mean's and variances where as in fig2, if you notice the graph is properly distributed with **mean =0** and **variance =1** always. concept of z-score comes in picture when we use **standardised normal data**.*

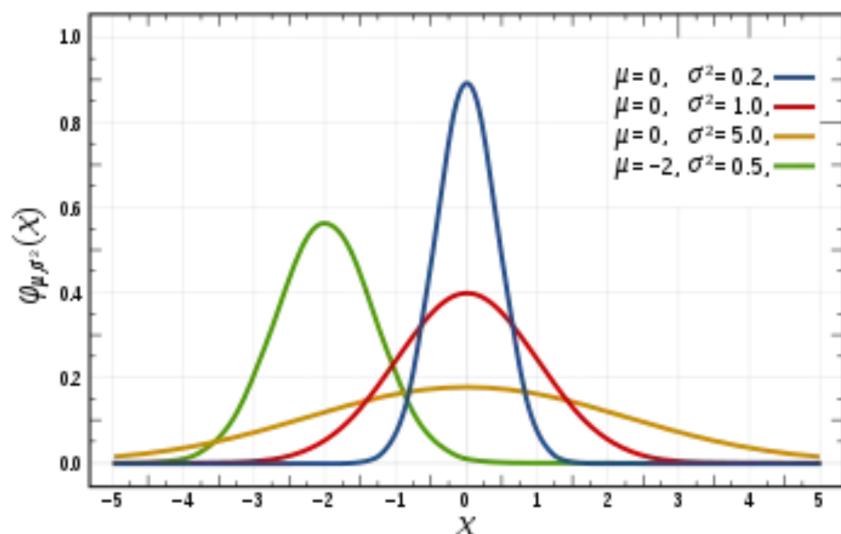


Fig. 9.10: Normal Curve images with different mean and variance

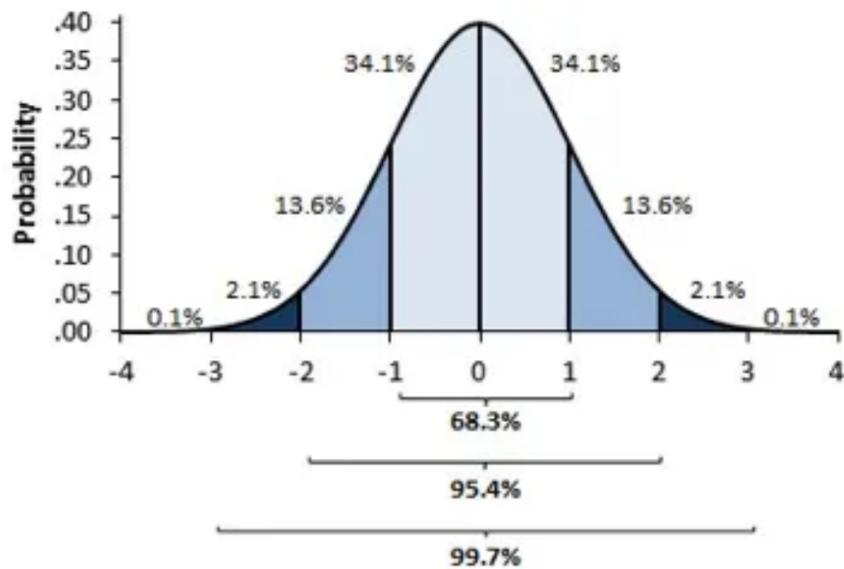


Fig. 9.11: Standardised Normal curve image and separation on data in percentage in each section

9.5.3 Which are important parameter of hypothesis testing?

1. Null hypothesis :- In inferential statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured phenomena.

Example : a company production is = 50 unit/per day etc.

2. Alternative hypothesis :- It is contrary to the null hypothesis. It is usually taken to be that the observations are the result of a real effect.

Example : a company production is !=50 unit/per day etc.

Condition of Null Hypothesis

	True	False
Possible Action		
Fail to reject H_0	Correct action	Type II error
Reject H_0	Type I error	Correct action

Conclusion:- If H_0 is rejected, we conclude that H_A is true. If H_0 is not rejected, we conclude that H_0 may be true.

Type I error:- When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by α . In hypothesis testing, the normal curve that shows the critical region is called the α region.

Type II error:- When we accept the null hypothesis but it is false. Type II errors are denoted by β . In Hypothesis testing, the normal curve that shows the acceptance region is called the β region.

9.6 Testing a hypothesis about the mean of a population

We have the following steps:

1. Data: determine variable, sample size(n), sample mean(\bar{x}), population standard deviation or sample standard deviation(s) if unknown.

2. Assumptions : We have two cases:-

Case1: Population is normally or approximately normally distributed with known or unknown variance (sample size n may be small or large).

Case 2: Population is not normal with known or unknown variance (n is large i.e. $n \geq 30$).

3.Hypothesis: We have three cases:-

Case I : $H_0: \mu = \mu_0$ and $H_A : \mu \neq \mu_0$

E.g. We want to test that the population mean is different than 50.

Case II : $H_0: \mu = \mu_0$ and $H_A : \mu > \mu_0$

E.g. We want to test that the population mean is greater than 50.

Case III : $H_0: \mu = \mu_0$ and $H_A : \mu < \mu_0$

E.g. We want to test that the population mean is less than 50.

9.6.1 Testing hypothesis for the mean μ

Condition: When the value of sample size (n):

Case A: Population is normal or not normal ($n \geq 30$)

Case I: σ is known

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (9.1)$$

Case II: σ is not known

$$Z = \frac{\bar{X} - \mu_0}{S\sqrt{n}} \quad (9.2)$$

Case B: Population is normal ($n < 30$)

Case I: σ is known

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (9.3)$$

Case II: σ is not known

$$Z = \frac{\bar{X} - \mu_0}{S\sqrt{n}} \quad (9.4)$$

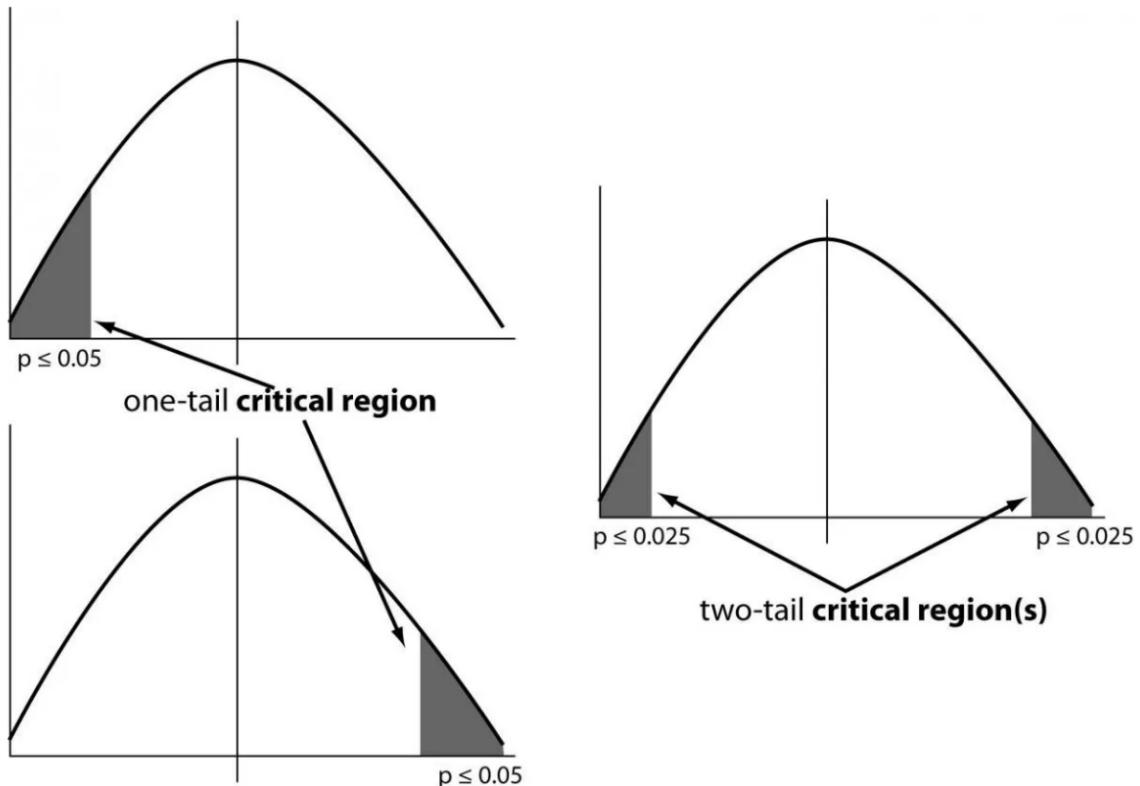
9.6.2 Test Types

One tailed test :- A test of a statistical hypothesis , where the region of rejection is on only one side of the sampling distribution, is called a one-tailed test.

Example :- a college has ≥ 4000 student or data science $\leq 80\%$ org adopted.

Two-tailed test :- A two-tailed test is a statistical test in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values. If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.

Example : a college has < 4000 student or data science $< 80\%$ org adopted.



Test Procedures:

Hypotheses	$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$	$H_0: \mu \leq \mu_0$ $H_A: \mu > \mu_0$	$H_0: \mu \geq \mu_0$ $H_A: \mu < \mu_0$
Test Statistic (T.S.)	Calculate the value of: $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$		
R.R. & A.R. of H_0			
Critical value (s)	$Z_{\alpha/2}$ and $-Z_{\alpha/2}$	$Z_{1-\alpha} = -Z_\alpha$	Z_α
Decision:	We reject H_0 (and accept H_A) at the significance level α if:		
	$Z < Z_{\alpha/2}$ or $Z > Z_{1-\alpha/2} = -Z_{\alpha/2}$ Two-Sided Test	$Z > Z_{1-\alpha} = -Z_\alpha$ One-Sided Test	$Z < Z_\alpha$ One-Sided Test

Test Procedures:

Hypotheses	$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$	$H_0: \mu \leq \mu_0$ $H_A: \mu > \mu_0$	$H_0: \mu \geq \mu_0$ $H_A: \mu < \mu_0$
Test Statistic (T.S.)	Calculate the value of: $t = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t(n-1)$ (df = v = n-1)		
R.R. & A.R. of H_0			
Critical value (s)	$t_{\alpha/2}$ and $-t_{\alpha/2}$	$t_{1-\alpha} = -t_\alpha$	t_α
Decision:	We reject H_0 (and accept H_A) at the significance level α if:		
	$t < t_{\alpha/2}$ or $t > t_{1-\alpha/2} = -t_{\alpha/2}$ Two-Sided Test	$t > t_{1-\alpha} = -t_\alpha$ One-Sided Test	$t < t_\alpha$ One-Sided Test

9.7 The Use of P-Values in Decision Definition Making

9.7.1 What is P-Values?

The P-value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H_0) of a study question is true — the definition of ‘extreme’ depends

on how the hypothesis is being tested.

P-value is the smallest value of α for which we can reject the null hypothesis H_0 .

Calculating P-value:

Calculating P-value depends on the alternative hypothesis H_A .

Suppose that

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (9.5)$$

is the computed value of the test statistic.

The following table illustrates how to compute P-value, and how to use P-value for testing the null hypothesis:

Alternative Hypothesis:	$H_A: \mu \neq \mu_0$	$H_A: \mu > \mu_0$	$H_A: \mu < \mu_0$
P-Value =	$2 \times P(Z > z_c)$	$P(Z > z_c)$	$P(Z < z_c)$
Significance Level =	α		
Decision:	Reject H_0 if P-value $< \alpha$.		

Decision:

If we reject H_0 , we can conclude that H_A is true.

If, however, we do not reject H_0 , we may conclude that H_0 is true.

9.7.2 An Alternative Decision Rule using the P-value Definition

The P-value is defined as the smallest value of α for which the null hypothesis can be rejected.

If the P-value is less than or equal to α , we reject the null hypothesis ($p \leq \alpha$).

If the P-value is greater than α , we do not reject the null hypothesis ($p > \alpha$).

If your P value is less than the chosen significance level then you reject the null hypothesis i.e. accept that your sample gives reasonable evidence to support the alternative hypothesis.

Example : You have a coin and you don't know whether that is fair or tricky so let's decide **null and alternate hypothesis**.

H_0 : a coin is a fair coin.

H_1 : a coin is a tricky coin and alpha = 5% or 0.05.

Now let's toss the coin and calculate P -value(probability value).

Toss a coin 1st time and result is tail, P -value = 50% (as head and tail have equal probability).

Toss a coin 2nd time and result is tail, now p -value = $50/2 = 25\%$.

and similarly we Toss 6 consecutive time and got result as P -value = 1.5 % but we set our significance level as 95% means 5% error rate we allow and here we see we are beyond that level i.e. our null- hypothesis does not hold good so we need to reject and propose that this coin is a tricky coin which is actually.

9.8 Hypothesis test for the population correlation coefficient ρ

This test is used to find if there is any linear relation between the two samples.

9.8.1 Steps for Hypothesis Testing for ρ

9.8.1.1 Hypotheses

Null Hypothesis $H_0: \rho = 0$

Alternate Hypothesis $H_A: \rho \neq 0$ or $H_A: \rho < 0$ or $H_A: \rho > 0$

9.8.1.2 Test Statistic

Test statistic is calculated using the following formula:

$$t^*: \frac{r\sqrt{n-2}}{\sqrt{1-R^2}}$$

Where r is the correlation coefficient and R^2 is its square.

For perfect Correlation both of these values will be 1.

9.8.1.3 P-Value

The P-value is determined by referring to a t-distribution with $n-2$ degrees of freedom.

P-value is the answer to the question "how likely is it that we'd get a test statistic t^* as extreme as we did if the null hypothesis were true.

9.8.1.4 Decision

If the P-value is smaller than the significance level α , we reject the null hypothesis in favor of the alternative. We conclude that there is sufficient evidence to prove that correlation exists between the two samples.

If the P-value is larger than the significance level α , we fail to reject the null hypothesis. We conclude that there is not enough evidence to prove that correlation exists between the two samples.

9.9 Problems based on rainfall data

You have to use Data of rainfall of last 42 years "Daily Data 42 Years 1981-2022.csv". Link

The Data consist the following information :

- The first three columns represent "Year", "Month" and "Day" respectively.
- The remaining columns represent the values of parameters of different locations.
- Each location has parameters Temperature at 2 meters (C) T2M, Temperature at 2 Meters Maximum (C) T2M MAX, Temperature at 2 Meters Minimum (C) T2M MIN, Specific Humidity at 2 Meters (g/kg), QV2M, Precipitation Prec, Wind Speed at 10 Meters (Degrees) WS10M and Wind Direction at 10 Meters (m/s) WD10M.
- The column indexes last part represent location id.
- Each row represent data at that given day.

```
1 # this is to make list of location coordinates to plot the graph
2 import geopandas as gpd
3 import matplotlib.pyplot as plt
4 import pandas as pd
5 import numpy as np
6 from scipy.stats import ttest_ind
7 from scipy.stats import ttest_rel
8 import matplotlib.pyplot as plt
9
```

```

10 import scipy.stats as stats
11 df=pd.DataFrame(pd.read_csv('Daily_Data_42_Years_1981-2022.csv'))
12
13 print(df.shape)
14
15 locations=[]
16 latitude = []
17 longitude = []
18 for i in range(4,3679,7):
19     # print(df.columns[i])
20     temp=df.columns[i].split("_")
21     if len(temp[2])==4:
22         lati=int(temp[2])/100
23     else:
24         lati=int(temp[2])
25
26     if len(temp[3])==4:
27         longi=int(temp[3])/100
28     else:
29         longi=int(temp[3])
30
31     locations.append([lati,longi])
32     latitude.append(lati)
33     longitude.append(longi)

```

9.9.1 Question 1

Find Sufficient evidence avg rainfall is significantly higher in summer than in winters on every location in 2022, and plot red color on the location where the hypothesis is rejected and green color where the hypothesis is accepted. ($\alpha = 0.05$)

9.9.1.1 Solution/explanation

Firstly we create Hypothesis i.e.

Null Hypothesis : $\mu_1 < \mu_2$

Alternate Hypothesis : $\mu_1 > \mu_2$

Where μ_1 and μ_2 are average rainfall in summer and winter respectively.

In this question we simply calculate the means for each and every location for both the seasons and compare them if the difference between them is more than 5% then we say that there is significant difference between the rainfall and the Alternate Hypothesis is accepted.

```

1 def rain_monthly_list(months):
2     df2=df[df.Year==2022]
3     df2=df2.reset_index()
4     x1=[]
5     for i in range(525):
6         x2=[]
7         temp=[]
8         for j in months:
9             df3=df2[df2.Month==j]
10            df3=df3.reset_index()
11            for k in df3.iloc[:,9+i*7]: # change x+i*7 for different variables
12                if k>0:
13                    temp.append(k)
14                else:
15                    temp.append(0)
16            x1.append(temp)
17     return x1
18 summer = rain_monthly_list([5,6,7])
19 winter = rain_monthly_list([10,11,12])
20 color = []
21 for i in range(525):
22     t_stat , p_value = ttest_ind(summer[i],winter[i])
23     # ttest_ind is used to calculate the p_value and t_statistic.
24     if(p_value < 0.05):
25         color.append('g')
26     else:
27         color.append('r')
28 shapefile=gpd.read_file("4-17-2018-899072.shp")
29 def Map_plot(longitude,latitude,value):
30     fig,ax=plt.subplots(figsize=(5,5))
31     plt.scatter(x=longitude , y = latitude,c =value)
32     plt.title('rainfall in summer vs winter')
33     plt.xlabel('Longitude')
34     plt.ylabel('Latitude')
35     shapefile.plot(ax =ax,color='black')
36     plt.legend(['winter','summer'])
37     plt.colorbar()
38     plt.show()
39 Map_plot(longitude,latitude,hypothesis_accept_reject)

```

Output :

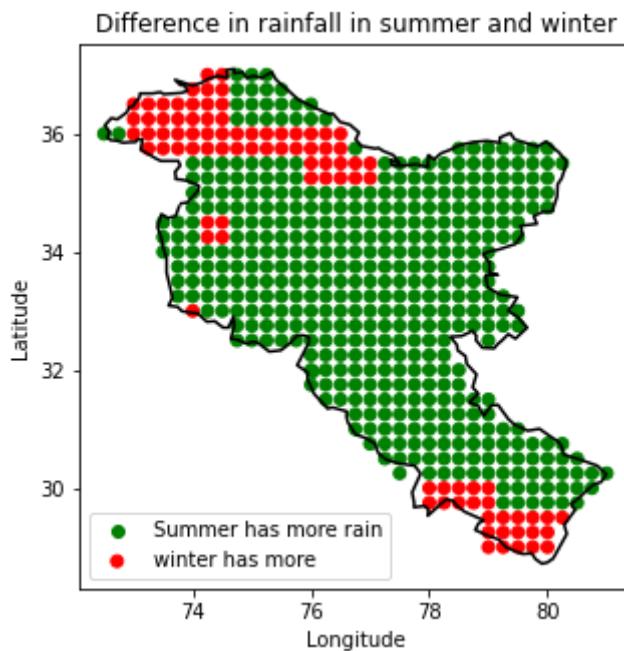


Fig. 9.12: Locations where rainfall is more in summer than in winter

9.9.2 Question 2

Perform the mean Comparison hypothesis test to show that rainfall is higher in recent years as compared to past. On location no. xxx where xxx is last three digits of your roll no. (B22xxx) (take $\alpha = 0.05$)

Recent years (2020-2022)

Past years (2013-2015)

9.9.2.1 Solution/explanation

Firstly we create Hypothesis i.e.

Null Hypothesis: $\mu_1 < \mu_2$

Alternate Hypothesis: $\mu_1 > \mu_2$

Where μ_1 and μ_2 are average rainfall in recent and past years respectively.

This question is about comparing means of one group at two different times, so we perform Two-tailed t-test.

To perform two tailed test we take monthly rainfall data for every every group of years i.e. (20,21,22) and (13,14,15) and then we make decision based on p-value.

Here is the code for all the locations Code:

```

1 def year_rainfall_monthly(year,i):
2     df3 = df[df.Year==year]
3     x2=[]
4     for j in range(1,13):
5         df4=df3[df3.Month==j]
6         df4=df4.reset_index()
7         temp=0
8         for j in df4.iloc[:,7+i*7]:
9             if j>0:
10                 temp+=j
11         temp=temp
12         x2.append(temp)
13     return x2
14 list_red_green = []
15 from scipy.stats import ttest_rel
16 for i in range(525):
17     y20 = year_rainfall_monthly(2020,i)
18     y21 = year_rainfall_monthly(2021,i)
19     y22 = year_rainfall_monthly(2022,i)
20     y13 = year_rainfall_monthly(2013,i)
21     y14 = year_rainfall_monthly(2014,i)
22     y15 = year_rainfall_monthly(2015,i)
23     sample1 = y20+y21+y22
24     sample2 = y13+y14+y15
25     t_stat, p_value = ttest_rel(sample1,sample2)
26     if p_value <= 0.05:
27         list_red_green.append('g')
28     else:
29         list_red_green.append('r')
30 shapefile=gpd.read_file("4-17-2018-899072.shp")
31 def Map_plot(longitude,latitude,value):
32     fig,ax=plt.subplots(figsize=(5,5))
33     plt.scatter(x=longitude , y = latitude,c =value)
34     plt.title('locations where rainfall in recent years has increased')
35     plt.xlabel('Longitude')
36     plt.ylabel('Latitude')
37     plt.legend(["increased in recent years","didn't increase"])
38     shapefile.plot(ax=ax,color='black')
39     plt.show()
40 Map_plot(longitude,latitude,list_red_green)

```

For instance if we look at location no. 50 we get the t-statistic = -0.93 and p-value = 0.3549.

locations where rainfall in recent years has increased

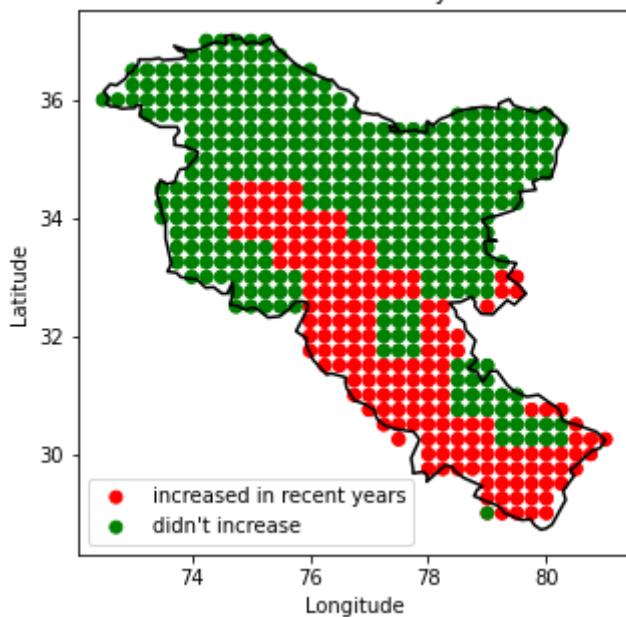


Fig. 9.13: locations where Rainfall in recent years is more

plot of rainfall in mm vs months for different years at location 50

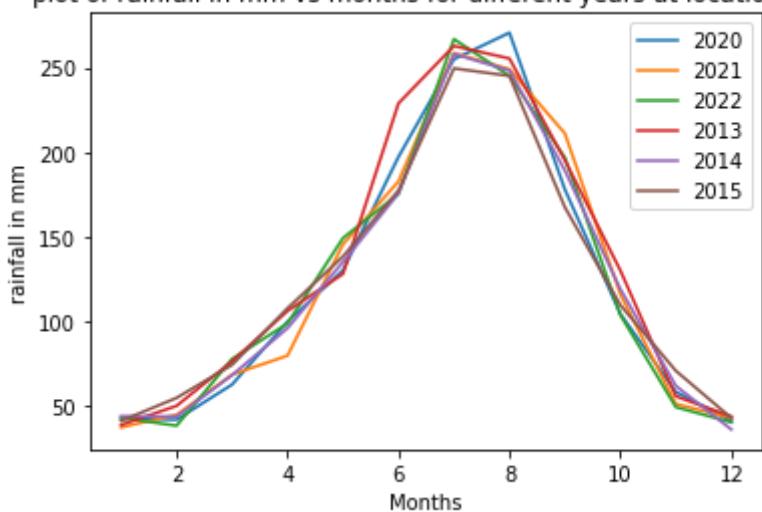


Fig. 9.14: monthly rainfall at location no. 50

While calculating for location no 150 we get the t -statistic = 2.24 and p -value = 0.031.

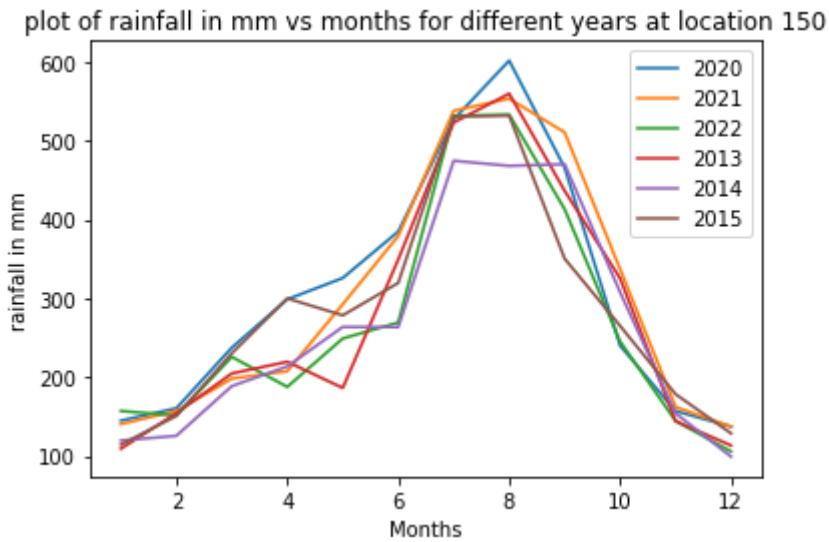


Fig. 9.15: monthly rainfall at location no. 150

At location no. 50 p-value is greater than $0.05(\alpha)$. which means that the null hypothesis holds.

9.9.3 Question 3

Perform the mean Comparison hypothesis test to show that average temperature is higher in recent years as compared to past. On location no. xxx where xxx is last three digits of your roll no. (B22xxx) (take $\alpha = 0.05$)

Recent years (2020-2022)

Past years (2013-2015)

9.9.3.1 Solution/explanation

This question is more or less the same as above question. But we will use t-statistic this time instead of p-value.

Null Hypothesis: Temperature in recent years did not increase as compared to past years.

Alternate Hypothesis: Temperature in recent years has increased as compared past years.

Code:

```

1 def year_temperature_monthly(year, i):
2     df3 = df[df.Year==year]
3     x2=[]
4     for j in range(1,13):
5         df4=df3[df3.Month==j]
6         df4=df4.reset_index()
7         temp=0

```

```

8     count = 0
9     for j in df4.iloc[:,4+i*7]:
10        if(temp < -30):
11            continue
12        else:
13            temp+=j
14            count +=1
15        temp=temp/count
16        x2.append(temp)
17    return x2
18 list_red_green = []
19 list_new = []
20 for i in range(525):
21     y20 = year_temperature_monthly(2020,i)
22     y21 = year_temperature_monthly(2021,i)
23     y22 = year_temperature_monthly(2022,i)
24     y13 = year_temperature_monthly(2013,i)
25     y14 = year_temperature_monthly(2014,i)
26     y15 = year_temperature_monthly(2015,i)
27     sample1 = y20+y21+y22
28     sample2 = y13+y14+y15
29     t_stat, p_value = ttest_rel(sample1,sample2)
30     critical = stats.t.ppf(q=0.05, df=35)
31     if(t_stat <= critical):
32         list_new.append('r')
33     else:
34         list_new.append('g')
35 shapefile=gpd.read_file("4-17-2018-899072.shp")
36 def Map_plot(longitude,latitude,value):
37     fig,ax=plt.subplots(figsize=(5,5))
38     plt.scatter(x=longitude , y = latitude,c =value)
39     plt.title('result increase in temperature or not hypothesis')
40     plt.xlabel('Longitude')
41     plt.ylabel('Latitude')
42     plt.legend(["increased in recent years","didn't increase"])
43     shapefile.plot(ax =ax,color='black')
44     plt.show()
45 Map_plot(longitude,latitude,list_new)

```

Output:

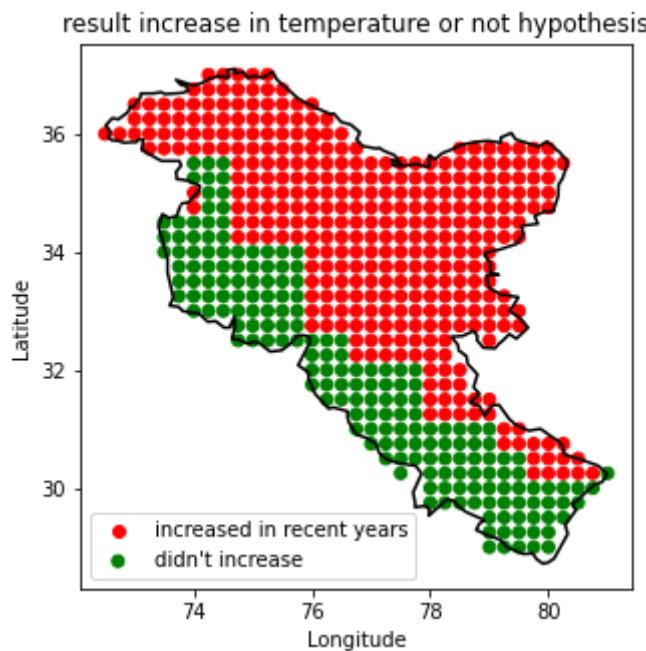


Fig. 9.16: locations where temperature has increased in recent years

We calculate t -statistic for location 150 which comes out to be -1.1538 while the critical value is -1.6895 , we can say that we reject the null hypothesis.

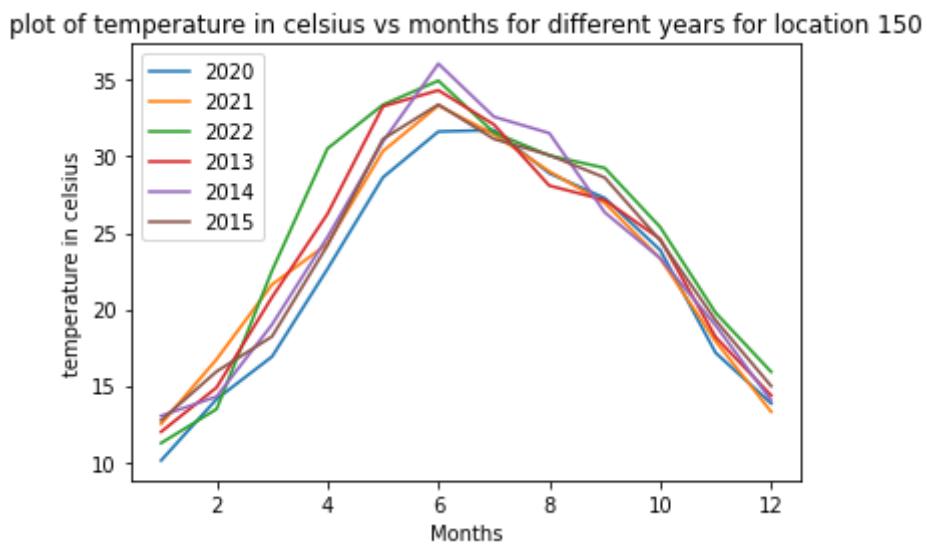


Fig. 9.17: monthly average of temperature at location 150

We calculate t -statistic for location 50 which comes out to be -1.8265 while the critical value is -1.6895 , we can say that we accept the null hypothesis.

plot of temperature in celsius vs months for different years for location 50

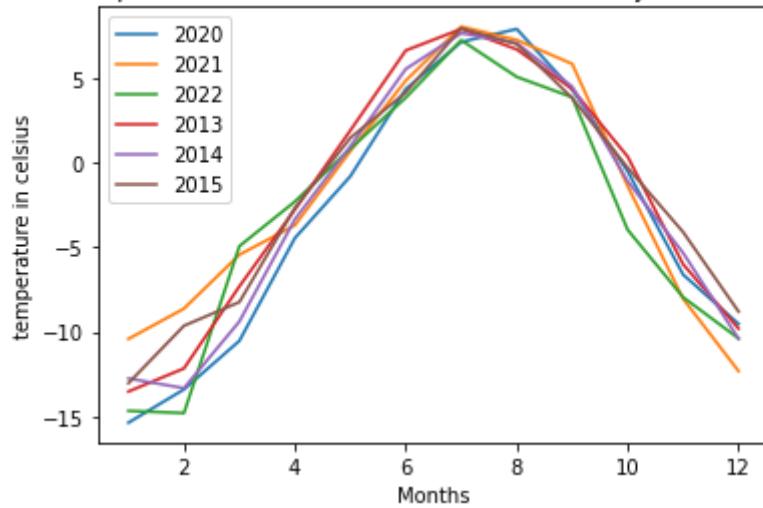


Fig. 9.18: monthly average of temperature at location 50

9.9.4 Question 4

Find Evidence to support the hypothesis that there is significant correlation between rainfall and Specific humidity in the year 2022 take significance level(α) as 0.01.

9.9.4.1 Solution/Explanation

Null Hypothesis *There is no correlation between rainfall and specific humidity*

Alternate Hypothesis *there is significant correlation between rainfall and humidity.*

code:

```

1 def month_avg(positi ,prm):
2     x_month=[]
3     for l in [2022]:
4         yr=l
5         df2=df[df.Year==yr]
6         df2=df2.reset_index()
7         x=[]
8         for i in range(1,13):
9             year=i
10            df3=df2[df2.Month==year]
11            df3=df3.reset_index()
12            #print(df3)
13            temp=0
14            for j in df3.iloc[:,7*positi+prm+4]:
```

```

16         #print(df3.columns)
17
18         if j>0:
19             temp+=j
20             temp=temp/len(df3.iloc[:,7*posi+prm+4])
21             x.append(temp)
22             # plt.plot(df3.index,df3.iloc[:,8])
23             # plt.show()
24             #print(sum(x)/len(x))
25             x_month.append(x)
26
27     return x_month
28
29 list_corr = []
30 corr = []
31
32 for i in range(525):
33     a = month_avg(i,4)
34     b = month_avg(i,2)
35     slope, intercept, r_value, p_value, std_err = stats.linregress(a[0],b[0])
36     corr.append(r_value)
37     if p_value < 0.01:
38         list_corr.append('g')
39     else:
40         list_corr.append('r')
41
42 shapefile=gpd.read_file("4-17-2018-899072.shp")
43 print(type(shapefile))
44
45 def Map_plot(longitude,latitude,value):
46     fig,ax=plt.subplots(figsize=(5,5))
47     plt.scatter(x=longitude , y = latitude,c = value)
48     plt.xlabel('Longitude')
49     plt.ylabel('Latitude')
50     plt.legend(['Correlation does not exists','correlation exist'])
51     shapefile.plot(ax=ax,color='black')
52     plt.show()
53
54 Map_plot(longitude,latitude,list_corr)
55 Map_plot(longitude , latitude , corr)

```

Output:

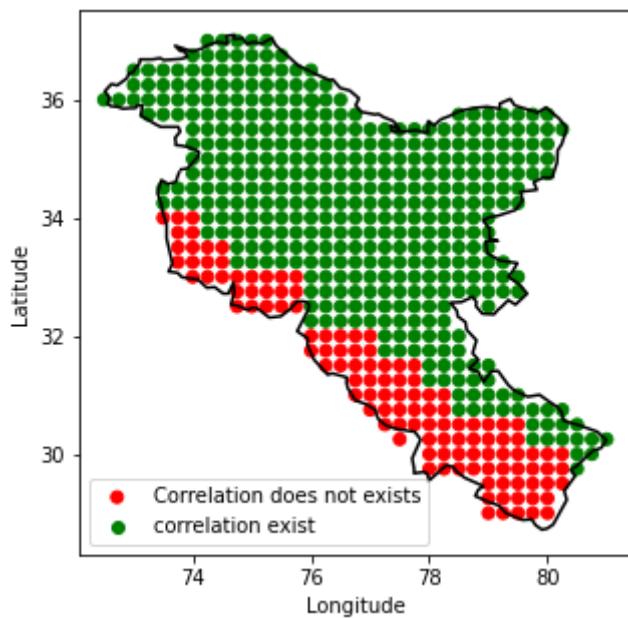


Fig. 9.19: locations where correlation between rainfall and Humidity exists

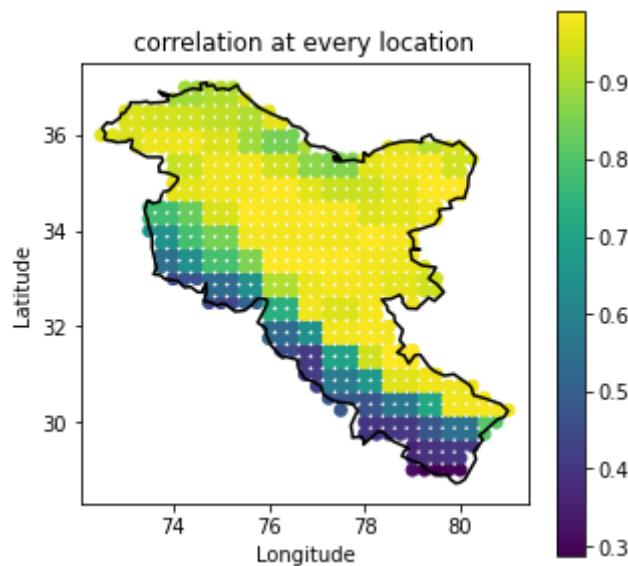


Fig. 9.20: correlation(r -value) at every location

Explanation: scipy.stats.linregress gives us r -value and the p -value. we make decision based on p -value as discussed above in theory. This Hypothesis tells us that correlation at green-locations is significantly different than "zero".

9.10 Problems

9.10.1 Task 1: Hypothesis Testing

A coffee company claims that their new coffee blend has a mean caffeine content of 120mg per cup. A consumer group tests 30 cups of coffee and finds a sample mean of 125mg with a sample standard deviation of 8mg. Test the claim made by the coffee company at a significance level of 0.01 using a one-tailed test.

9.10.1.1 Solution

```
1 import numpy as np
2 import scipy.stats as stats
3 import matplotlib.pyplot as plt
4
5
6 # Task 1
7 # Sample data
8 n = 30
9 sample_mean = 125
10 sample_std = 8
11
12 # Null hypothesis
13 null_mean = 120
14
15 # Calculate t-statistic and p-value
16 t_statistic, p_value = stats.ttest_1samp(np.random.normal(loc=null_mean, scale=
17     sample_std/n**0.5, size=n), sample_mean)
18
19 # Test hypothesis at significance level of 0.01
20 alpha = 0.01
21 print(t_statistic, p_value)
22 if p_value / 2 < alpha and t_statistic > 0:
23     print("The result is statistically significant and suggests that the caffeine
24         content of the coffee is higher than claimed by the company.")
25 else:
26     print("The result is not statistically significant and there is not enough evidence
27         to reject the company's claim.")
```

9.10.1.2 Output

$t\text{-statistic} = -22.0505$

$p\text{-value} = 1.1469e-19$

The result is not statistically significant and there is not enough evidence to reject the company's claim.

9.10.2 Task 2: Confidence Intervals

Using the same data as in Task 1, calculate a 99% confidence interval for the true mean caffeine content of the coffee. Interpret the results.

9.10.2.1 Solution

```
1 import numpy as np
2 from scipy.stats import t
3
4 # Sample data
5 n = 30
6 x_bar = 125
7 s = 8
8
9 # Degrees of freedom
10 df = n - 1
11
12 # t-value for 99% confidence interval
13 t_val = t.ppf(0.99, df)
14
15 # Margin of error
16 moe = t_val * (s / np.sqrt(n))
17
18 # Confidence interval
19 lower_ci = x_bar - moe
20 upper_ci = x_bar + moe
21
22 print("99% Confidence Interval: {:.2f}, {:.2f})".format(lower_ci, upper_ci))
```

9.10.2.2 Output

99% Confidence Interval: (121.40, 128.60)

9.10.3 Task 3: Sample Size Calculation

Suppose that the consumer group in Task 1 wants to test the claim made by the coffee company with a power of 0.90 at a significance level of 0.01. How many cups of coffee should they sample assuming that the true mean caffeine content is 122mg and the standard deviation is 8mg?

9.10.3.1 Solution

```
1 from scipy.stats import norm
2 import numpy as np
3 # Population parameters
4 mu = 122
5 sigma = 8
6
7 # Significance level and power
8 alpha = 0.01
9 power = 0.90
10
11 # Z-values for one-tailed test
12 z_alpha = norm.ppf(1 - alpha)
13 z_beta = norm.ppf(power)
14
15 # Calculate required sample size
16 n = ((z_alpha + z_beta) * sigma / (mu - 120)) ** 2
17
18 # Round up to nearest integer
19 n = int(np.ceil(n))
20
21 print("Required Sample Size:", n)
```

9.10.3.2 Output

Required Sample Size: 209

9.10.4 Task 4

Plot the sampling distribution of sample means and plot the null-mean and p-value

9.10.4.1 Solution

```
1 # Visualize the sampling distribution
2 import numpy as np
3 import matplotlib.pyplot as plt
4 null_distribution = np.random.normal(loc=null_mean, scale=sample_std/n**0.5, size
5 =100000)
6 fig, ax = plt.subplots(figsize=(10, 5))
7 ax.hist(null_distribution, bins=100, density=True, alpha=0.5)
8 ax.axvline(sample_mean, color='red', label='Sample Mean')
9 ax.axvline(null_mean, color='black', label='Null Mean')
10 ax.set_title('Sampling Distribution of the Sample Mean', fontsize=14)
11 ax.set_xlabel('Caffeine Content (mg)', fontsize=12)
```

```

11 ax.set_ylabel('Density', fontsize=12)
12 ax.legend(fontsize=12)
13 plt.show()
14
15 # Visualize the p-value
16 fig, ax = plt.subplots(figsize=(10, 5))
17 ax.hist(null_distribution, bins=100, density=True, alpha=0.5)
18 if t_statistic > 0:
19     ax.axvline(sample_mean + abs(null_mean - sample_mean) * 2, color='red', label='p-
        value')
20 else:
21     ax.axvline(sample_mean - abs(null_mean - sample_mean) * 2, color='red', label='p-
        value')
22 ax.axvline(sample_mean, color='red', label='Sample Mean')
23 ax.axvline(null_mean, color='black', label='Null Mean')
24 ax.set_title('Sampling Distribution of the Sample Mean with p-value', fontsize=14)
25 ax.set_xlabel('Caffeine Content (mg)', fontsize=12)
26 ax.set_ylabel('Density', fontsize=12)
27 ax.legend(fontsize=12)
28 plt.show()

```

9.10.4.2 output

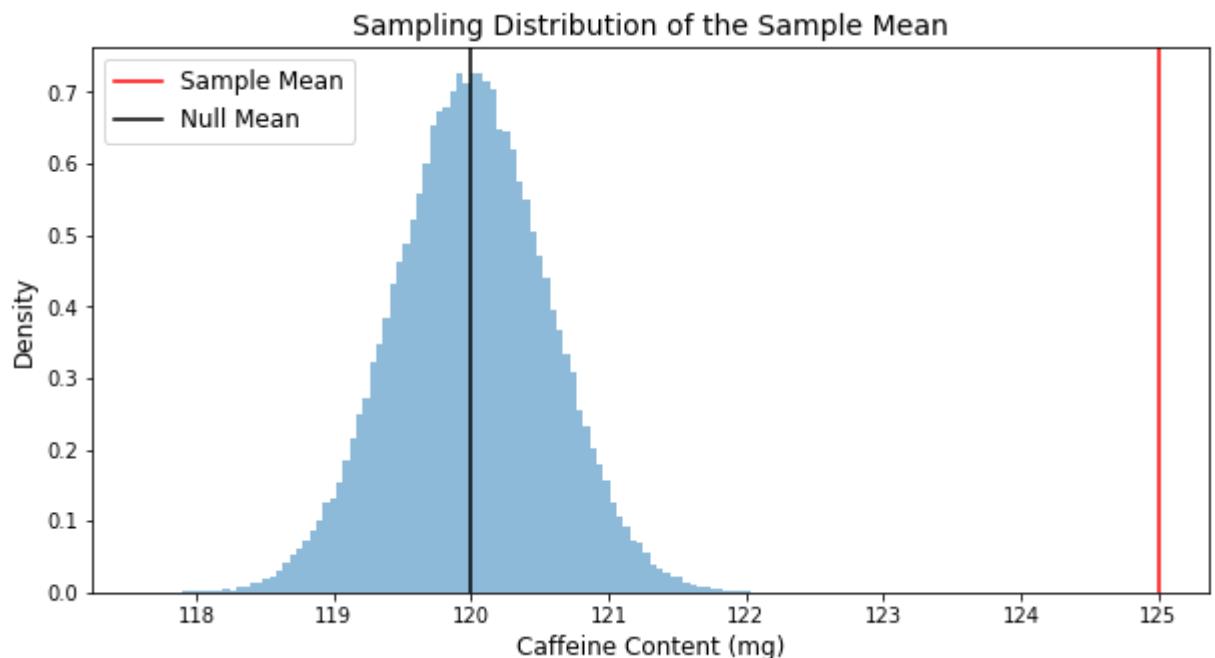


Fig. 9.21: Sampling distribution of sample mean

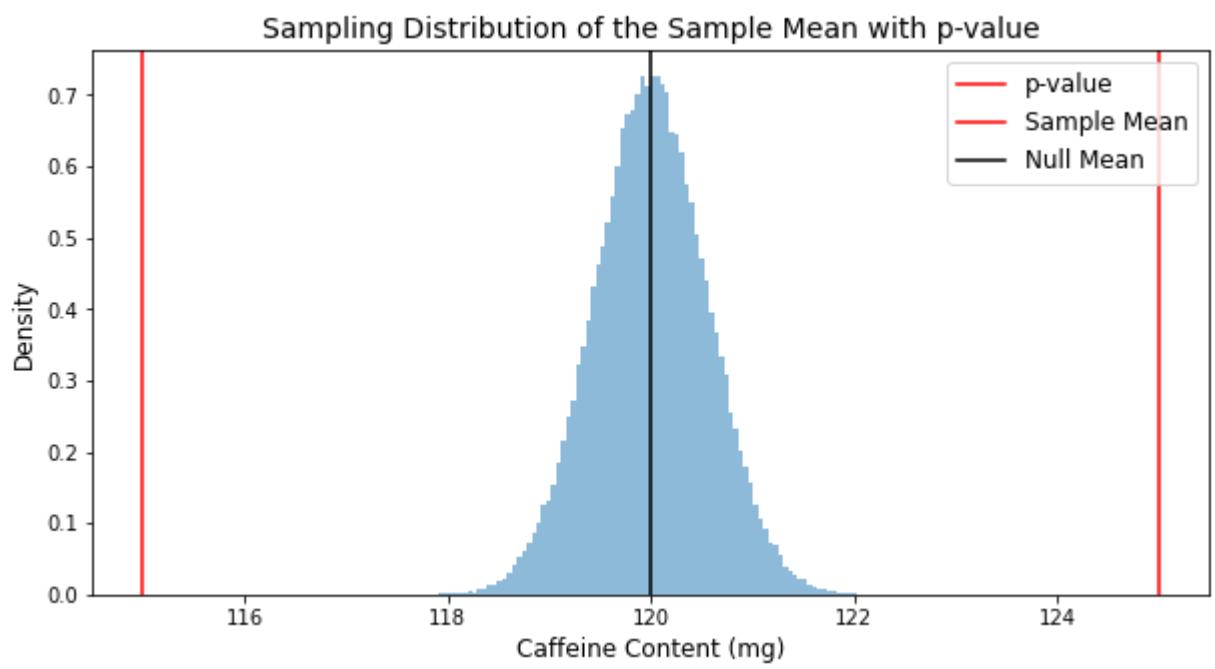


Fig. 9.22: Sampling distribution of sample mean with p-value

Chapter 10

Regression Analysis

Abstract: Regression is a statistical technique to establish a relationship between one dependent variable and one or more independent variables. This chapter explores using regression to model non-linear data, specifically the relationship between temperature and rainfall. Starting with simple linear regression and progressing to more complex multi-variate and polynomial regression, we examine how each model can capture the non-linear relationship between temperature and rainfall. We evaluate the performance of each model using metrics such as R-squared and mean-squared error, highlighting the importance of choosing the appropriate regression model to achieve accurate predictions.

10.1 Introduction

Regression is a statistical technique used to establish a relationship between one dependent variable and one or more independent variables. The aim of regression is to find a function that can accurately predict the value of the dependent variable given the independent variable(s).

When fitting a regression model, it is important to evaluate its performance in order to determine whether it is a good fit for the data. Two commonly used metrics to evaluate the performance of a regression model are mean squared error (MSE) and R-squared (R²).

10.1.1 Assumptions of Regression

10.1.1.1 Linearity

The relationship between the independent variables and the dependent variable is assumed to be linear. This means that the change in the dependent variable for a unit change in any independent variable is constant.

10.1.1.2 Independence

The observations used in the regression analysis should be independent of each other. This assumption implies that the values of the dependent variable for one observation should not be influenced by or related to the values of the dependent variable for other observations.

10.1.1.3 Homoscedasity

The variability of the errors, or the residuals, is assumed to be constant across all levels of the independent variables. In other words, the spread of the residuals should be consistent as the values of the independent variables change.

10.1.1.4 No endogeneity

There should be no correlation between the error term and the independent variables. This assumption ensures that the independent variables are not affected by any omitted variables or other sources of bias, which could lead to biased and inconsistent parameter estimates.

10.1.1.5 No perfect Multicollinearity

The independent variables should not be perfectly correlated with each other. Perfect multicollinearity occurs when one independent variable can be perfectly predicted from a linear combination of other independent variables, which makes it impossible to estimate the separate effects of the variables.

10.1.1.6 Normality of Residuals

The residuals, or the differences between the observed and predicted values of the dependent variable, should be normally distributed. This assumption is important for hypothesis testing and constructing confidence intervals.

10.1.1.7 No autocorrelation

The residuals should not exhibit any systematic patterns or correlations with each other. Autocorrelation, or serial correlation, occurs when the residuals from one observation are correlated with the residuals from previous or future observations. This assumption is particularly relevant in time series regression analysis.

Other than these, the accuracy of data, missing data, and outliers also affect the regression models and should be well-checked beforehand. By understanding and adhering to these assumptions, researchers can make reliable inferences and interpretations based on regression analysis results. It is crucial to assess these assumptions before drawing conclusions from a regression model.

10.1.2 Mean Squared Error

MSE measures the average squared difference between the predicted and actual values of the dependent variable. A lower MSE indicates a better fit of the model to the data, as it means that the predicted values are closer to the actual values.

10.1.3 R2 score

R2 measures the proportion of the variation in the dependent variable that is explained by the independent variable(s). R2 ranges from 0 to 1, with a higher R2 indicating a better fit of the model to the data, as it means that a larger proportion of the variation in the dependent variable can be explained by the independent variable(s).

When evaluating the performance of a regression model, both MSE and R2 should be considered together. A low MSE and a high R2 indicate that the model is accurately predicting the dependent variable and explaining a large proportion of the variation in the data. On the other hand, a high MSE and a low R2 indicate that the model is not accurately predicting the dependent variable or explaining much of the variation in the data, and may not be a good fit for the data.

10.2 Scatter Plots

A scatter plot is a type of graph used to display the relationship between two variables. It shows the values of one variable plotted against the values of the other variable, with each point representing an observation in the data.

Scatter plots are useful in doing correlation and regression analysis because they help to visually identify patterns and relationships between the two variables. In correlation analysis, we can use a scatter plot to examine the direction and strength of the relationship between two variables. A positive correlation between two variables is indicated by a general upward trend in the scatter plot, while a negative correlation is indicated by a general downward trend. The strength of the correlation is indicated by how tightly the points are clustered around the trend line.

In regression analysis, we use scatter plots to visually inspect the relationship between the dependent and independent variables. The scatter plot helps to identify the nature of the relationship, whether it is linear, non-linear or has outliers. It also helps to identify any potential issues with the data such as heteroscedasticity or non-constant variance.

After analyzing the scatter plot, we can then fit a regression line to the data to model the relationship between the variables. The regression line represents the best-fit line through the data, and allows us to estimate the value of the dependent variable for any given value of the independent variable. The regression line is chosen such that it minimizes the distance between the predicted values and the actual values of the dependent variable.

In summary, scatter plots are a useful tool in correlation and regression analysis as they help to visually identify patterns and relationships in the data. They provide a graphical representation of the data that can be used to identify potential issues and help guide the selection of the appropriate regression model.

10.3 Simple Linear Regression

Linear regression is a statistical technique used to establish a linear relationship between a dependent variable and one or more independent variables. It is a commonly used method in predictive modeling and is used to predict the value of the dependent variable based on the values of the independent variable(s).

10.4 Multivariate Regression

Multivariate regression is a statistical technique used to establish a linear relationship between a dependent variable and two or more independent variables. It is an extension of simple linear regression and is used to predict the value of the dependent variable based on the values of the independent variables.

10.5 Polynomial Regression

Polynomial regression is a form of regression analysis in which the relationship between the independent variable (x) and the dependent variable (y) is modeled as an n th degree polynomial. It is an extension of linear regression, which assumes a linear relationship between x and y . Polynomial regression can capture more complex relationships between the variables by allowing for curved or nonlinear relationships.

The equation for a polynomial regression model is:

$$y = b_0 + b_1x + b_2x^2 + \dots + b_n * x^n$$

where y is the dependent variable, x is the independent variable, and $b_0, b_1, b_2, \dots, b_n$ are the coefficients of the model. The degree of the polynomial is denoted by n .

10.6 Applying the regression on a Real-World Data

We are given with the monthly data of rainfall, temperature, specific humidity, wind direction, and wind speed data of different regions of Himachal pradesh for 21 years.

We have to use regression linear as well as non-linear to check which model fits the best and can be used to predict the rainfall in the different parts of the Himachal region.

Using linear regression on the data, we get the results as:

```

1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 from sklearn.linear_model import LinearRegression
5 from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
6
7 # Load the data from CSV file
8 data=pd.read_csv('2001-2021_Monthly_Data_525_Grids_7_Parameters.csv')
9
10 # Extract temperature and rainfall columns as numpy arrays
11 dfrm=data[['T2M_32785']]
12 #dfrm=data[['T2M_192-33-76', 'RH2M_192-33-76', 'PS_192-33-76', 'WS10M_192-33-76',
13 #WD10M_192-33-76', 'Prec_192-33-76']]
14 #dfrm=data[['T2M_296-34-79', 'RH2M_296-34-79', 'PS_296-34-79', 'WS10M_296-34-79',
15 #WD10M_296-34-79', 'Prec_296-34-79']]
16 dfmy=data['Prec_32785']
17 df_np=dfrm.to_numpy()
18
19 x_train,y_train=dfrm, dfmy
20 sklearn_model=LinearRegression().fit(x_train,y_train)
21 x_test=dfrm
22 y_test=dfmy
23
24 y_pred=sklearn_model.predict(x_test)
25 print("R2 score= ", r2_score(y_test,y_pred))
26 print("Mean squared error= ", mean_squared_error(y_test,y_pred))
27 print("Mean absolute error= ", mean_absolute_error(y_test,y_pred))
28
29 plt.plot(y_test,label='Actual Precipitation')
30 plt.plot(y_pred,'--',label='Predicted Precipitation')
31 plt.title('Actual Precipitation v/s Predicted Precipitation')
32 plt.ylabel('Precipitation')
33 plt.xlabel('Time (months)')
34 plt.legend()
35 plt.show()
```

Output of the above code:

```

1 R2 score= -0.07585809721537484
2 Mean squared error= 73.22770881577307
3 Mean absolute error= 7.111452209104175
```

Next we use the multivariate model of regression to predict and check the compatibility

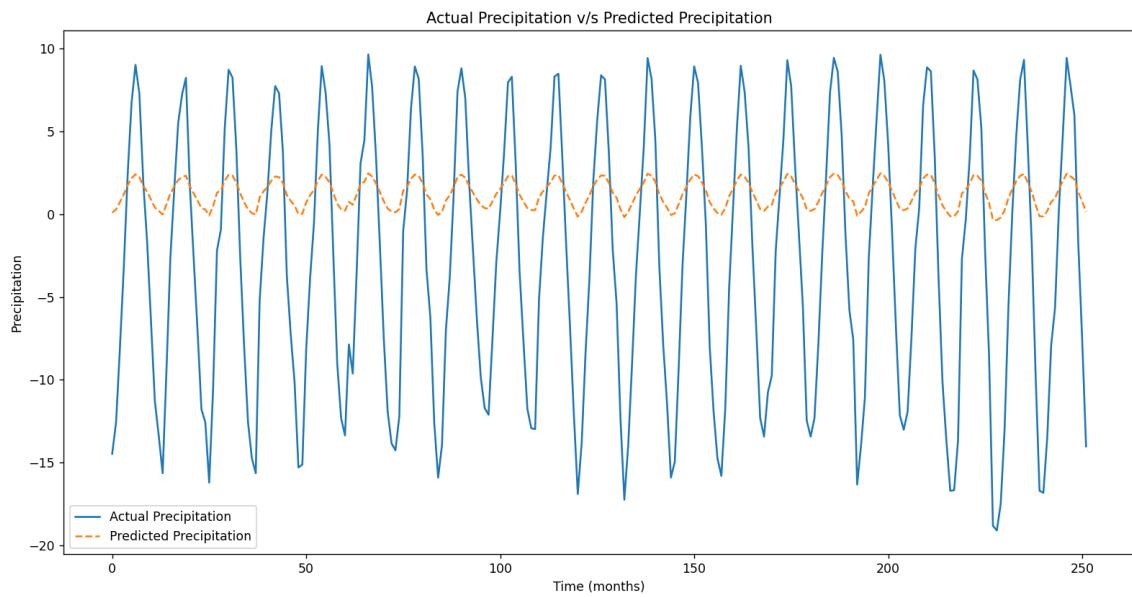


Fig. 10.1: Actual vs Predicted Data Comparison via Simple Linear Regression

of the model with the data. Using the multivariate model we get the results as below:

```

1 # Import necessary libraries
2 import pandas as pd
3 import numpy as np
4 import matplotlib.pyplot as plt
5 from sklearn.linear_model import LinearRegression
6 from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
7
8 # Load the data from CSV file
9 data=pd.read_csv('2001-2021_Monthly_Data_525_Grids_7_Parameters.csv')
10 #dfrm=data[['T2M_29_79','RH2M_29_79','PS_29_79','WS10M_29_79',
11 'WD10M_29_79','Prec_29_79']]
12 dfrm=data[['T2M_32785','RH2M_32785','PS_32785','WS10M_32785',
13 'WD10M_32785']]
14 #dfrm=data[['T2M_192-33-76','RH2M_192-33-76','PS_192-33-76','WS10M_192-33-76',
15
16 'WD10M_192-33-76','Prec_192-33-76']]
17 #dfrm=data[['T2M_296-34-79','RH2M_296-34-79','PS_296-34-79','WS10M_296-34-79',
18 'WD10M_296-34-79','Prec_296-34-79']]
19 dfrm_y=data['Prec_32785']
20 df_np=dfrm.to_numpy()
21 x_train,y_train=dfrm, dfrm_y
22 sklearn_model=LinearRegression().fit(x_train,y_train)
23 x_test=dfrm
24 y_test=dfrm_y
25

```

```

26 y_pred=sklearn_model.predict(x_test)
27 print("R2 score= ", r2_score(y_test,y_pred))
28 print("Mean squared error= ", mean_squared_error(y_test,y_pred))
29 print("Mean absolute error= ", mean_absolute_error(y_test,y_pred))
30 plt.plot(y_test,label='Actual Precipitation')
31 plt.plot(y_pred,'--',label='Predicted Precipitation')
32 plt.title('Actual Precipitation v/s Predicted Precipitation')
33 plt.ylabel('Precipitation')
34 plt.xlabel('Time (months)')
35 plt.legend()
36 plt.show()

```

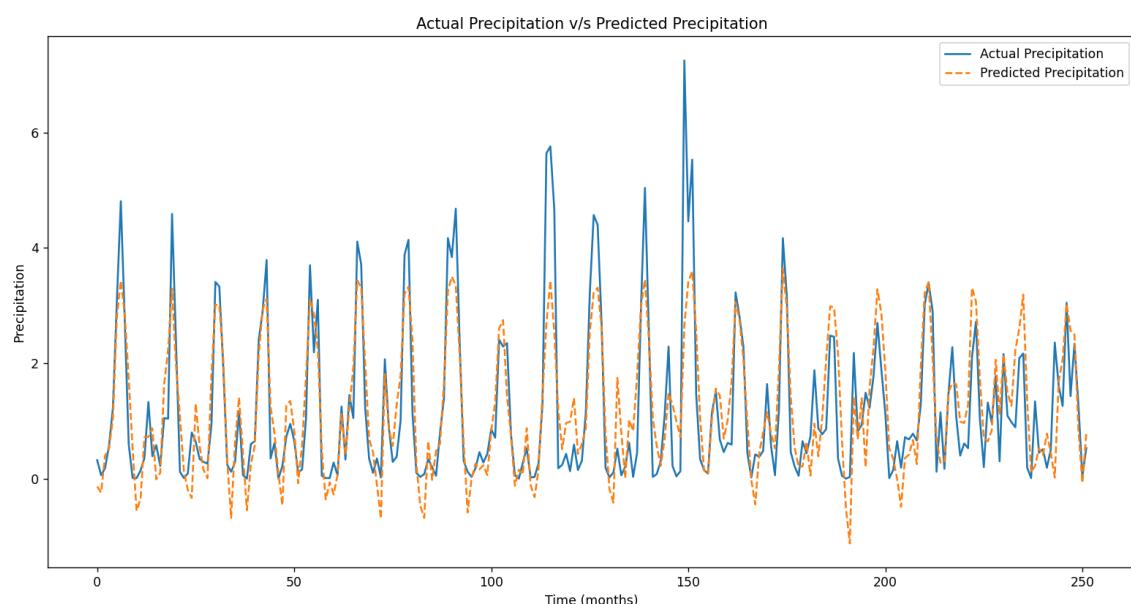


Fig. 10.2: Actual vs Predicted Data Comparison via Multivariate Linear Regression

Output of the above code:

```

1 R2 score=  0.7001131916301473
2 Mean squared error=  0.5614287626358583
3 Mean absolute error=  0.5456204111180978

```

Now we use polynomial regression model to predict the values and check if non-linear model is getting a better fit for the following data, so we got the results as below:

```

1 # Import necessary libraries
2 import pandas as pd
3 import numpy as np
4 import matplotlib.pyplot as plt
5 from sklearn.linear_model import LinearRegression
6 from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error

```

```

7
8 # Load the data from CSV file
9 data=pd.read_csv('2001-2021_Monthly_Data_525_Grids_7_Parameters.csv')
10 #dfrm=data[['T2M_29_79','RH2M_29_79','PS_29_79','WS10M_29_79',
11 'WD10M_29_79','Prec_29_79']]
12 dfrm=data[['T2M_32785','RH2M_32785','PS_32785','WS10M_32785',
13 'WD10M_32785']]
14 #dfrm=data[['T2M_192-33-76','RH2M_192-33-76','PS_192-33-76','WS10M_192-33-76',
15 'WD10M_192-33-76','Prec_192-33-76']]
16 #dfrm=data[['T2M_296-34-79','RH2M_296-34-79','PS_296-34-79','WS10M_296-34-79',
17 'WD10M_296-34-79','Prec_296-34-79']]
18 dfmyy=data['Prec_32785']
19 df_np=dfrm.to_numpy()
20 x_train,y_train=dfrm, dfmyy
21 x_test=dfrm
22 y_test=dfmyy
23
24 from sklearn.preprocessing import PolynomialFeatures
25 poly = PolynomialFeatures(degree=2)
26 X_poly_train = poly.fit_transform(x_train)
27 X_poly_test=poly.fit_transform(x_test)
28 # Fit the model using the transformed feature
29 model = LinearRegression()
30 model.fit(X_poly_train, y_train)
31 ypred = model.predict(X_poly_test)
32 print("R2 score= ", r2_score(y_test,ypred))
33 print("Mean squared error= ", mean_squared_error(y_test,ypred))
34 print("Mean absolute error= ", mean_absolute_error(y_test,ypred))
35 plt.plot(y_test,label='Actual Precipitation')
36 plt.plot(ypred,'--',label='Predicted Precipitation')
37 plt.title('Actual Precipitation v/s Predicted Precipitation')
38 plt.ylabel('Precipitation')
39 plt.xlabel('Time (months)')
40 plt.legend()
41 plt.show()

```

Output of the above code:

```

1 R2 score= 0.810621844376354
2 Mean squared error= 0.3545415824057068
3 Mean absolute error= 0.3750964470732207

```

10.6.0.1 Conclusion

From the above plots and parameters obtained from different models, it is very easy to guess non-linear model is more accurate on the data available to us for the above problem.

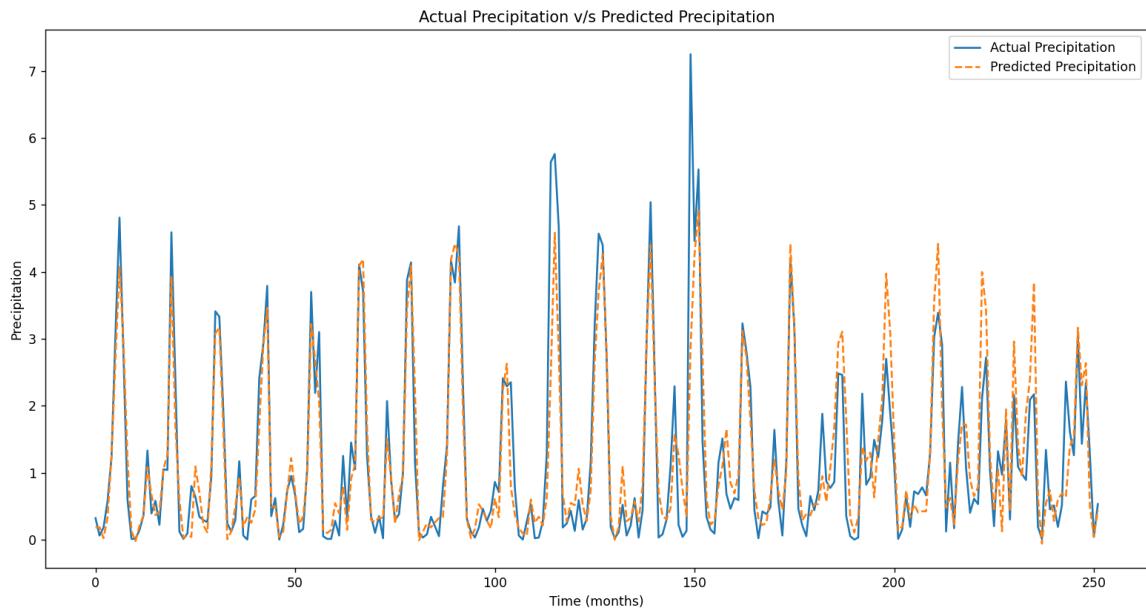


Fig. 10.3: Actual vs Predicted Data Comparison via Polynomial Regression

10.6.1 Covid Data

In some of the cases we don't have a proper model to fit our data, so we have to create a model of our own. Here is an example of the same in which there is no use of standard regression models, but we have used the standard models to create one of our own so that it is pretty easy to predict the variables(in this case mobility).

In this problem we were given different data sets, one is the vaccination data of different districts(our main aim is to model for Mumbai) and the mobility of the people. Different scenarios for achieving the desired immunization rates are evaluated using nonlinear regression models. The impact of recovery rates on mobility is also assessed to determine how the economy would have fared in various scenarios.

Here is the code for the problem

```

1 import pandas as pd
2 import statistics as st
3 import matplotlib.pyplot as plt
4 from sklearn.decomposition import PCA
5 import numpy as np
6 import math
7 from sklearn.model_selection import train_test_split
8 from sklearn.preprocessing import StandardScaler
9 from sklearn.neighbors import KNeighborsClassifier
10 from sklearn.metrics import confusion_matrix , accuracy_score
11 from sklearn import mixture

```

```

12 from scipy.stats import pearsonr
13 from sklearn.linear_model import LinearRegression
14 from sklearn.preprocessing import PolynomialFeatures
15 from csv import reader
16 import mobility # A python file
17 from mpl_toolkits import mplot3d
18 import pickle
19 from sklearn.metrics import r2_score
20 # import r0_calculation
21 from scipy.interpolate import make_interp_spline
22
23 df = pd.read_csv('cowin_vaccine_data_districtwise.csv', low_memory=False)
24 dfn = pd.read_csv('districts (3).csv', low_memory=False)
25
26 #####
27
28 death = list(dfn['Deceased'])
29 state = list(dfn['State'])
30 confirmed = list(dfn['Confirmed'])
31 recovered = list(dfn['Recovered'])
32 district = list(dfn['District'])
33 deathrate=[]
34 confirmed_per=[]
35 testing_per=[]
36 recovered_per=[]
37 for i in range(168736,358632):
38     if(district[i]== 'Mumbai'):
39         deathrate.append(death[i]/209614)
40         confirmed_per.append(confirmed[i]/209614)
41         testing_per.append(dfn.iloc[i,7]/209614)
42         recovered_per.append(recovered[i]/209614)
43 r_c=[]
44 for i in range(len(recovered_per)):
45     r_c.append(recovered_per[i]/confirmed_per[i])
46 d_c=[]
47 for i in range(len(deathrate)):
48     d_c.append(deathrate[i]/confirmed_per[i])
49 #####
50
51
52 fdm=[]
53 sessions1=[]
54 sites1=[]
55 f=393
56 l=394
57 for c in range(9, len(df.columns),10): # 9---> column for first dose

```

```

58     t=0
59     for r in range(f,l):
60         t=t+int(df.iloc[r,c])
61     fdm.append(t/20961400) # 20961400 is the population of Mumbai
62
63 for c in range(7,len(df.columns),10): # 7---> column for sessions
64     t=0
65     for r in range(f,l):
66         t=t+int(df.iloc[r,c])
67     sessions1.append(t/604)      # 604 sq.km is the area of Mumbai
68 for c in range(8,len(df.columns),10): # 8---> column for sites
69     t=0
70     for r in range(f,l):
71         t=t+int(df.iloc[r,c])
72     sites1.append(t/604)
73
74 ####
75
76 mob=list(mobility.avgm)[16:305]
77 c1,_=pearsonr(sites1,fdm)
78 c2,_=pearsonr(sessions1,fdm)
79 c3,_=pearsonr(deathrate,fdm)
80 c4,_=pearsonr(confirmed_per,fdm)
81 c5,_=pearsonr(recovered_per,fdm)
82 c6,_=pearsonr(mob,fdm)
83
84 print("Correlation between First dose and Sites : ",c1)
85 print("Correlation between First dose and Sessions : ",c2)
86 print("Correlation between First dose and deathrate : ",c3)
87 print("Correlation between First dose and Confirmed cases : ",c4)
88 print("Correlation between First dose and recovered : ",c5)
89 print("Correlation between First dose and Pharma Mobility : ",c6)
90 months = ['Jan 21','Feb 21','March 21','April 21','May 21','June 21','July 21',
91 'August 21','September 21','October 21']
92 xlen=[x for x in range(16,289,30)]
93 plt.plot([x for x in range(16,305)],fdm,color='red',linestyle='dashed')
94 plt.plot([x for x in range(16,305)],[((x*1.85) for x in fdm],color='purple',
95 linestyle='dashdot')
96 plt.plot([x for x in range(16,150)],[((x*6.76) for x in fdm[:134]],color='green')
97 plt.xticks(xlen,months,rotation = 90)
98 plt.ylim(0,100)
99 plt.legend(['Scenario 1','Scenario 2','Scenario 3'])
100 plt.ylabel('Vaccination coverage (%) ')
101 plt.xlabel('Time (in months)')
102 plt.grid()
103 plt.show()

```

```

104
105
106 ###
107 recovered_per=np.array(recovered_per)
108 confirmed_per=np.array(confirmed_per)
109 r=recovered_per/confirmed_per
110 Y = r
111 X=[]
112 for i in range(len(sites1)):
113     t=[]
114     t.append(fdm[i])
115     t.append(mob[i])
116     X.append(t)
117 X = np.array(X)
118 X_temp = X
119 Y = np.array(Y)
120 X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3,
121 random_state=42, shuffle=True)
122
123 ### Testing ----->>>>>>>>>
124 # Using the linear regression model from scikit learn
125 reg = LinearRegression().fit(X_train, Y_train)
126 y_pred_train=reg.predict(X_train)
127 y_pred_test=reg.predict(X_test)
128 print(reg.coef_)
129 corr_matrix = np.corrcoef(Y_test, y_pred_test)
130 corr = corr_matrix[0,1]
131 R_sq = corr**2
132 print("r2",R_sq)
133 for i in range(len(X_train)):
134     X_train[i][0]+=0.5
135 for i in range(len(X_test)):
136     X_test[i][0]+=0.5
137 y_pred_testv=reg.predict(X_test)
138 # print(max(y_pred_testv)-max(y_pred_test))
139 a=0
140
141 s=0
142 for i in range(len(Y_test)):
143     s=s + (Y_test[i]-y_pred_test[i])*(Y_test[i]-y_pred_test[i])
144 E_rmse=math.sqrt(s/len(Y_test))# Computing Prediction accuracy / RMSE on test data
145 print("ermse",E_rmse)
146
147 ###
148 poly_features = PolynomialFeatures(2) #p is the degree
149 x_poly = poly_features.fit_transform(X_train) # gives the polynomial matrix of X

```

```

150 regressor=LinearRegression()
151 regressor.fit(x_poly,Y_train)      # print(regressor.intercept_)
152 # k=regressor.coef_
153 # print(k)
154 # fitting data in our linear regression scikit model
155 x_poly_given=poly_features.fit_transform(X_temp)
156 y_pred_temp = regressor.predict(x_poly_given)
157 for i in range(len(X)):
158     # Vary the attribute accordingly to see the desired result
159     X_temp[i][0]=X[i][0]+0.365*X[i][0]
160 # fitting New data in our linear regression scikit model
161 x_poly_givenv=poly_features.fit_transform(X_temp)
162 y_predv_temp = regressor.predict(x_poly_givenv)
163 s=0
164 for i in range(len(y_predv_temp)):
165     s=s+(y_predv_temp[i]-y_pred_temp[i])
166 print(s/len(y_pred_temp))                                #Average difference
167 print(max(y_predv_temp)-max(y_pred_temp))               # Maximum difference
168 print(max(y_predv_temp))
169 print(max(y_pred_temp))
170 print(max(Y))
171 print(sum(Y)/len(Y))
172
173 print("#")
174
175 poly_features = PolynomialFeatures(1) #p is the degree
176 x_poly = poly_features.fit_transform(X_train) # gives the polynomial matrix of X
177 regressor=LinearRegression()
178 regressor.fit(x_poly,Y_train)      # print(regressor.intercept_)
179 # k=regressor.coef_
180 # print(k)
181 # fitting data in our linear regression scikit model
182 x_poly_given=poly_features.fit_transform(X)
183 y_pred = regressor.predict(x_poly_given)
184 for i in range(len(X)):
185     # Vary the attribute accordingly to see the desired result
186     X[i][0]=X[i][0]+5.8*X[i][0]
187 # fitting New data in our linear regression scikit model
188 x_poly_givenv=poly_features.fit_transform(X)
189 y_predv = regressor.predict(x_poly_givenv)
190 s=0
191 for i in range(len(y_predv)):
192     s=s+(y_predv[i]-y_pred[i])
193 print(s/len(y_pred))                                #Average difference
194 print(max(y_predv)-max(y_pred))                   # Maximum difference
195 print(max(y_predv))

```

```

196 print(max(y_pred))
197 print(max(Y))
198 print(sum(Y)/len(Y))
199
200
201
202 months = ['Jan 21', 'Feb 21', 'March 21', 'April 21', 'May 21', 'June 21',
203 'July 21', 'August 21', 'September 21', 'October 21']
204 x=np.array([i for i in range(16,305)])
205 xlen=[x for x in range(16,289,30)]
206 plt.plot(x,[y*100 for y in Y],color='red',linestyle='dashed')
207 plt.plot(x,[y*100 for y in y_predv_temp],color='purple',linestyle = 'dashdot')
208 plt.plot(x[:134],[y*100 for y in y_predv[:134]],color='green')
209 plt.xticks(xlen,months,rotation=90)
210 plt.ylabel('Recovery Rate (%)')
211 plt.legend(['Scenario 1','Scenario 2','Scenario 3'])
212 plt.xlabel('Time (in months)')
213 plt.ylim(70,100)
214 # plt.ylabel('Vaccination Coverage %')
215 plt.grid()
216 plt.show()
217
218
219 s=0
220 cc,_=pearsonr(r,fdm)
221 print(cc)
222 for i in range(len(y_pred)):
223     s=s+(y_predv[i]-y_pred[i])*209614
224 print(s)
225 print((max(y_predv)-max(y_pred)))
226 s=0
227 for i in range(len(Y_test)):
228     s=s + (Y_test[i]-y_pred[i])*(Y_test[i]-y_pred[i])
229 E_rmse=math.sqrt(s/len(Y_test))
230 print(E_rmse)
231
232 ##
233 ## A function to calculate E_rmse for different p
234
235 def Multiple_NLR(p,x_given,y_given):
236     poly_features = PolynomialFeatures(p) #p is the degree
237     x_poly = poly_features.fit_transform(X_train) # gives the polynomial matrix of X
238     regressor=LinearRegression()
239     regressor.fit(x_poly,Y_train)
240     # print(regressor.intercept_)
241     # print(regressor.coef_)
```

```

242 k=regressor.coef_
243 # fitting data in our linear regression scikit model
244 x_poly_given=poly_features.fit_transform(x_given)
245 y_pred = regressor.predict(x_poly_given)
246 s=0
247 for i in range(len(y_given)):
248     s=s + (y_given[i]-y_pred[i])*(y_given[i]-y_pred[i])
249 E_rmse=math.sqrt(s/len(y_given))
250 return E_rmse,y_pred
251
252 print("Prediction accuracies of Test data")
253 p2,y2=Multiple_NLR(2,X_test,Y_test)
254 p3,y3=Multiple_NLR(3,X_test,Y_test)
255 p4,y4=Multiple_NLR(4,X_test,Y_test)
256 p5,y5=Multiple_NLR(5,X_test,Y_test)
257 p6,y6=Multiple_NLR(6,X_test,Y_test)
258 corr_matrix = np.corrcoef(Y_test, y3)      #put any yi to check its R2
259 corr = corr_matrix[0,1]
260 R_sq = corr**2      #Godness of fit
261 print(R_sq)
262 for i in range(5):
263     print("Prediction Accuracy for p = ",i+2," : ",Multiple_NLR(i+2,X_test,Y_test)[0])
264
265 ##### Testing ----->>>>>>>
266
267 ## Vary sites
268 for i in range(len(X_train)):
269     X_train[i][0]+=1
270 for i in range(len(X_test)):
271     X_test[i][0]+=1
272 p3v,y3v,k3v=Multiple_NLR(3,X_test,Y_test,X_train,Y_train)
273 a=0
274 for i in range(len(y3)):
275     a=a+y3v[i]-y3[i]
276 print(a/len(y3))
277 plt.scatter(Y_test,y2,color='red')
278 plt.scatter(Y_test,y3v,color='black')
279 plt.show()
280 print(max(y3))
281 print(max(y3v))
282
283 ### To find the Correlation Table for Mumbai
284
285 data = {
286     'Vaccination': fdm,
287     'sites':sites1,

```

```

288     'Confirmed':confirmed_per,
289     'Recovered':recovered_per,
290     'Death':deathrate,
291     'Mobility':mob,
292 }
293 dataframe = pd.DataFrame(data,columns=data.keys())
294 matrix = dataframe.corr()
295 print("Correlation Matrix : ")
296 print(matrix)

```

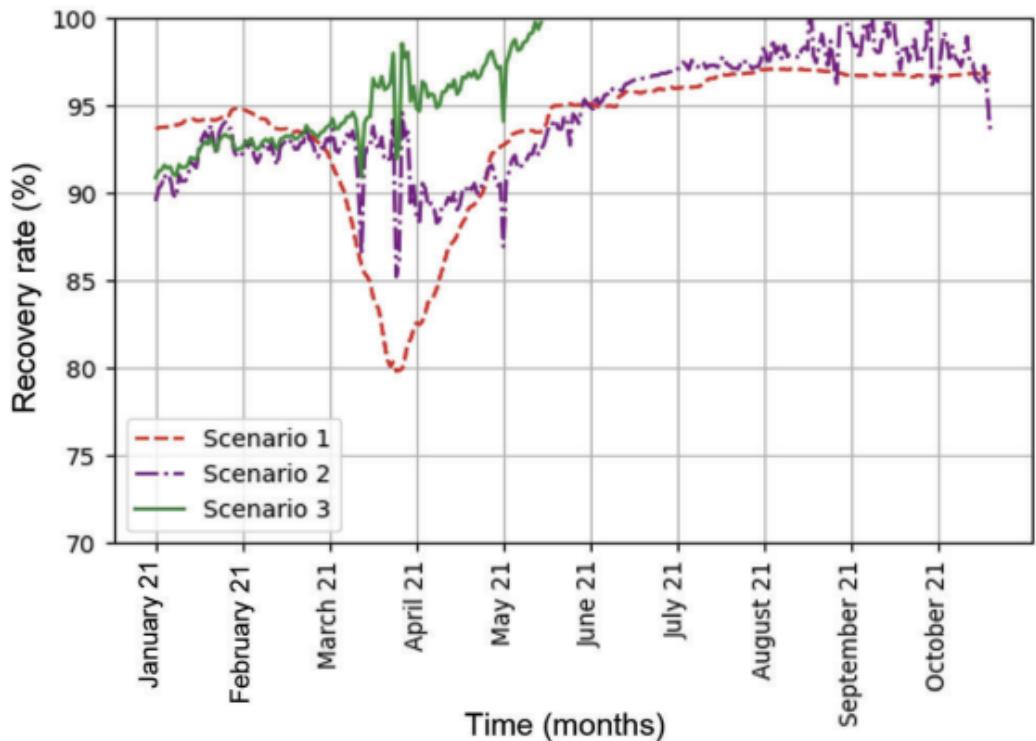


Fig. 10.4: Recovery Rate plot in all three scenarios

10.6.1.1 Conclusion

Lock-downs due to COVID-19, which restricted mobility, were the main cause of the decline in GDP. For the city of Mumbai in India, with an increase in recovery rate from 1% to 5%, it was observed that mobility and thus economic activity might have been restored to some extent. The findings presented here may aid the governing bodies in developing more effective emergency response plans.

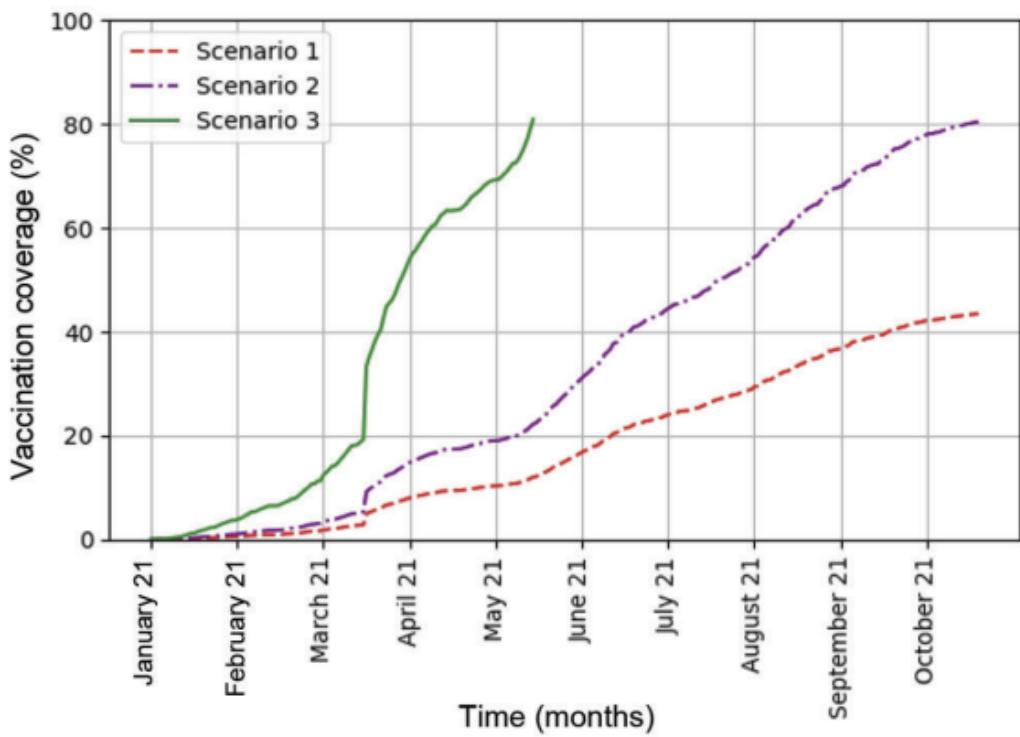


Fig. 10.5: Vaccination coverage plot in all three scenarios

10.6.2 Practice Problems

Import the given data file having You are given a data file of rainfall. The dataset has been prepared to make rainfall predictions for different parts of regions of Himachal. Rainfal.csv

Attribute information:-

Abbreviations of parameters are:

```

1 PS - Pressure (psi)
2 T2M - Temperature at 2 Meters (C)
3 RV2M - Specific Humidity at 2 Meters (g/kg)
4 T2M_MAX - Temperature at 2 Meters Maximum (C)
5 T2M_MIN - Temperature at 2 Meters Minimum (C)
6 WD10M - Wind Direction at 10 Meters (Degrees)
7 WS10M - Wind Speed at 10 Meters (m/s)
8 Prec - Precipitation

```

Perform the following task based on linear regression

Note:- use sklearn's train_test_split module to split the data into 70-30 train test split to tra

- 1) Find the correlation between precipitation and humidity/temperature at 2 meters/wind speed.

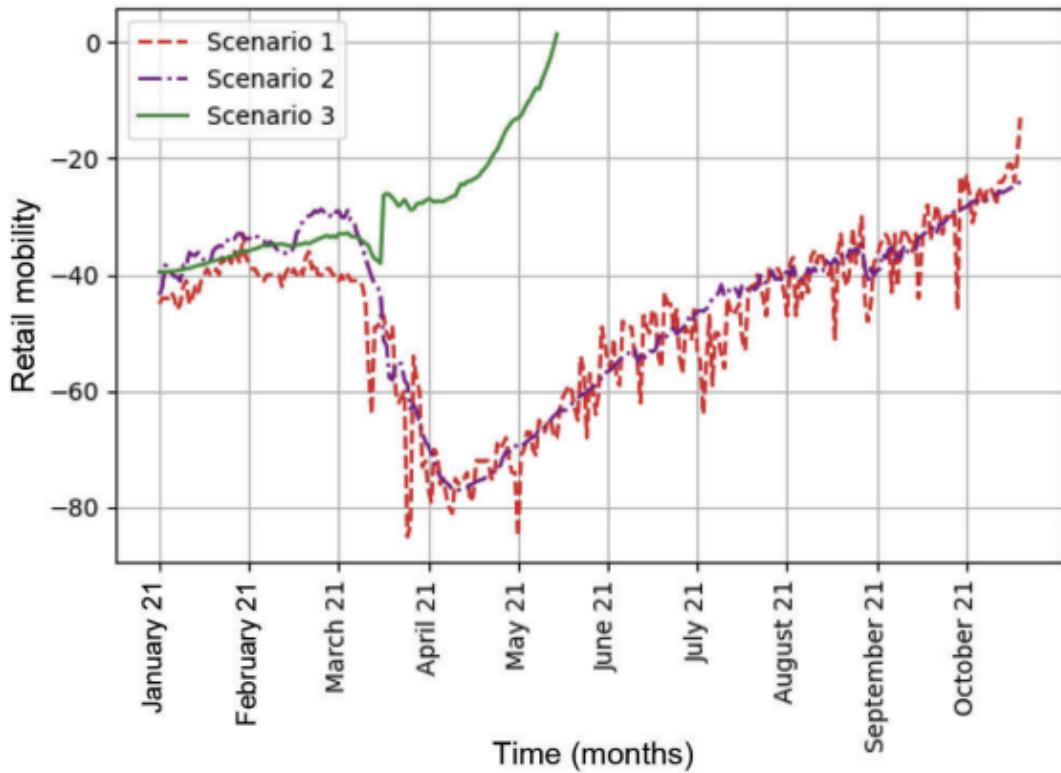


Fig. 10.6: Retail mobility plot in all three scenarios

Use the attribute with the highest correlation coefficient for the regression.

- 2) Build a simple linear regression model, to predict the rainfall using temperature data.
 - a) Plot the Actual precipitation vs Predicted precipitation plot and infer the results from the plot. Is the regression model accurate? (use scatter plot)
 - b) Plot Predicted precipitation vs attribute with the highest correlation coefficient
 - c) Find the prediction accuracy of the training data using mean squared error.
 - d) Find the prediction accuracy using r2_score and mean absolute error.(sklearn)

Some code snippets which will help you write the code:

For simple Linear regression:-

```

1 from sklearn.linear_model import LinearRegression
2 from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error

```

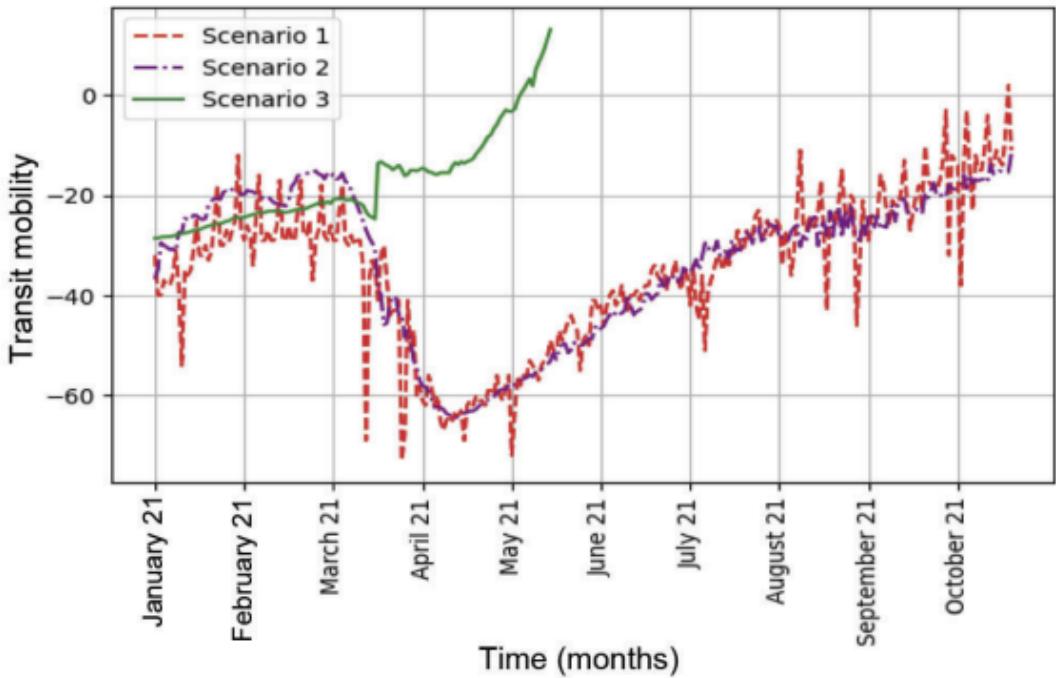


Fig. 10.7: Transit mobility plot in all three scenarios

```

3 x_train = data[['humidity']]
4 Y_train = data['precipitation']
5 sklearn_model=LinearRegression().fit(x_train,y_train)
6 y_pred=sklearn_model.predict(x_test)

```

Note: Use of standard libraries like sklearn is allowed and encouraged

10.7 Regression Questions

Q1. Students in a statistics class claimed that doing the homework had not helped prepare them for the midterm exam. The exam score y and homework score x (averaged up to the time of the midterm) for the 18 students in the class were as follows:

Find the prediction equation.

Q2.

(a) Find the linear relation between the number of widgets purchased and the cost per widget.

X : Number of widgets purchased – 1, 3, 6, 10, 15

Y : Cost per widget (in rupees) – 55, 52, 46, 32, 25

Table 10.1: Exam and Homework Scores

y	x
95	96
80	77
0	0
0	0
79	78
77	64
72	89
66	47
98	90
90	93
0	18
95	86
35	0
50	30
72	59
55	77
75	74
66	67

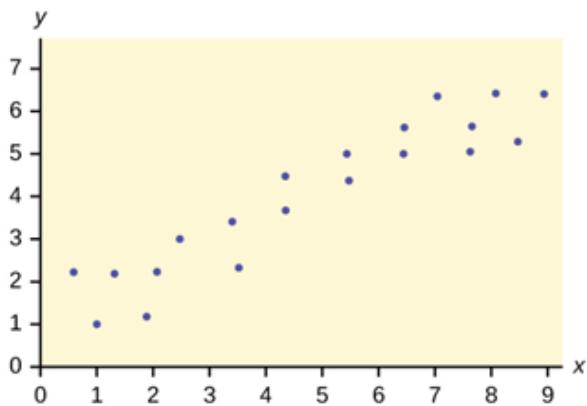
- (b) Suppose the regression line is $y = -3x + 60$. Compute the average price per widget if 25 are purchased. What inference do you get from it?
- (c) Under a "scatter diagram," there is a notation that the coefficient of correlation is 0.10. What does this mean?

Q3. The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

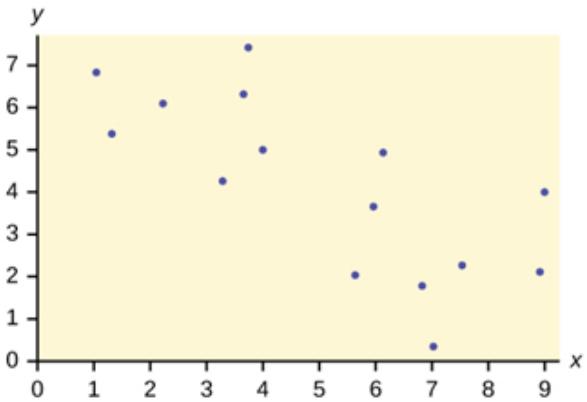
- (a) Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.
- (b) If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

Q4. Does the scatter plot appear linear? Strong or weak? Positive or negative?

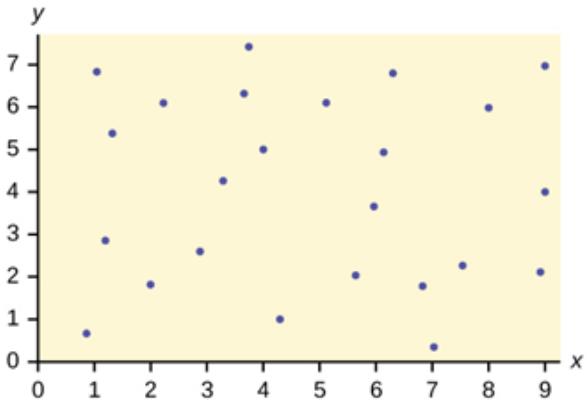
- (a)



(b)



(c)



Q5. A random sample of ten professional athletes produced the following data where x is the number of endorsements the player has and y is the amount of money made (in millions of dollars).

x	y
0	2
3	8
2	7
1	3
5	13
5	12
4	9
3	9
0	3
4	10

- (a) What is the slope of the line of best fit? What does it represent?
- (b) What does an r value of zero mean?
- (c) When $n = 100$ and $r = -0.89$, is there a significant correlation? Explain.

Q6. When testing the significance of the correlation coefficient, what is the alternative hypothesis?

solution 1.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{81195 - 18(58.056)(61.389)}{80199 - 18(58.056)^2} = 0.8726$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 61.389 - 0.8726(58.056) = 10.73$$

The predicted equation is thus given by $\hat{y} = 10.73 + 0.8726x$.

solution 2.

- (a) Using the equation of the solution from the first question, we get $\hat{y} = 57.4 - 2.2x$.
- (b) $y = -15$ rupees, which is obviously nonsense. This reminds us that predicting Y outside the range of X values in the data is a very poor practice.
- (c) On a scale from -1 to $+1$, the degree of linear relationship between the two variables is $+0.10$.

solution 3.

- (a) CI: $(7.9441, 8.4559)$
- (b) The level of confidence would decrease because decreasing n makes the confidence interval wider, so at the same error bound, the confidence level decreases.

solution 4.

- (a) The data appear to be linear with a strong, positive correlation.
- (b) The data appear to be non-linear with a strong, negative correlation.
- (c) The data appear to have no correlation.

solution 5.

- (a) The slope is 1.99 ($b = 1.99$). It means that for every endorsement deal a professional player gets, he gets an average of another 1.99 million in pay each year.
- (b) It means that there is no correlation between the data sets.
- (c) Yes, there are enough data points and the value of r is strong enough to show that there is a strong negative correlation between the data sets.

solution 6. $H_a: \rho \neq 0$

Chapter 11

Descriptive Statistics

11.1 What is Descriptive Statistics?

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).

11.1.1 Measures of Central Tendency

The mean, median, and mode are all measures of central tendency that describe where the center of a data set is located.

11.1.1.1 Mean

The mean is calculated by adding up all of the values in a data set and dividing by the total number of values. The formula for the mean is:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (11.1)$$

where \bar{x} is the mean, n is the total number of values, and x_i represents each individual value in the data set.

Example: Suppose we have the following rainfall data (in millimeters) for the last 10 years (2013-2022) in a city of Himachal Pradesh:

[1257, 730, 857, 716, 717, 852, 772, 849, 851, 556]

Find the mean rainfall of last 10 years at this place.

Solution: To calculate the mean rainfall over these 10 years, we use the formula:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (11.2)$$

where \bar{x} is the mean, n is the total number of values, and x_i represents each individual value in the data set.

Plugging in the values, we get:

$$\bar{x} = \frac{1257 + 730 + 857 + 716 + 717 + 852 + 772 + 849 + 851 + 556}{10} = \frac{8,157}{10} = 815.7 \quad (11.3)$$

Therefore, the mean rainfall over the last 10 years in a city of Himachal Pradesh is 815.7 millimeters.

11.1.1.2 Median

The median is the middle value in a data set when the values are arranged in order from smallest to largest. If there are an even number of values, the median is the average of the two middle values.

Example: Suppose we have the following minimum temperature data (in Celsius) of 10 winter days in a city of Himachal Pradesh:

$$[-16, -14, -15, -13, -17, -14, -19, -19, -16, -17]$$

Find the median minimum temperature of the given data.

Solution: Since we have an even number of values, the median is the average of the middle two values, which are $-16^\circ C$ and $-16^\circ C$. Therefore, the median minimum temperature is:

$$\text{median} = \frac{-16 + -16}{2} = -16 \quad (11.4)$$

Therefore, the median minimum temperature of 10 winter days of a city in Himachal Pradesh is $-16^\circ C$.

11.1.1.3 Mode

The mode is the value that occurs most frequently in a data set. It is possible for a data set to have more than one mode.

Figure 11.1 is an example of histogram of a data. As mode is most occurring data you can see that (red line) the data with highest peak is considered as mode of the data taken.(42 in this case)

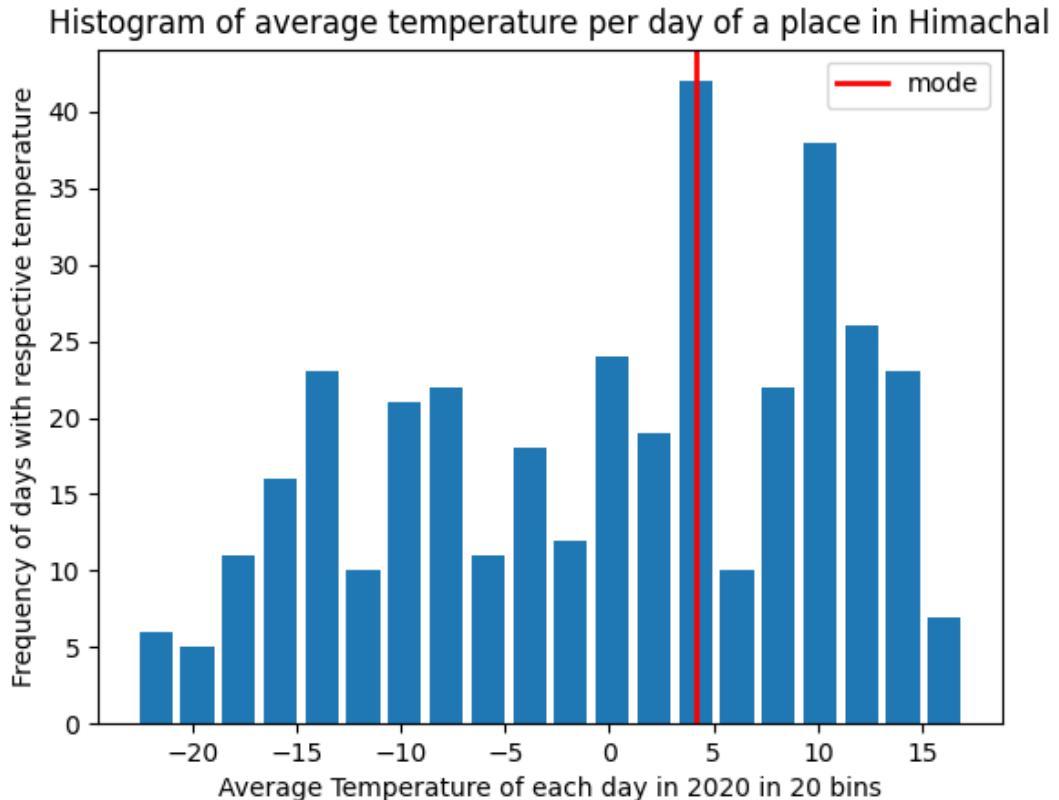


Fig. 11.1: Mode of Average Temp of each day in 2020

It is obtained using following code.

```

1 import pandas as pd
2 import matplotlib.pyplot as plt
3 df=pd.DataFrame(pd.read_csv('Daily_Data_42_Years_1981-2022.csv'))
4
5 df2=df[df.Year==2020]
6 df2=df2.reset_index()
7 import numpy as np
8 data=df2.iloc[:,4+7*400]
9 hist, bin_edges = np.histogram(data, bins=20)
10
11 max_freq_index = np.argmax(hist)
12 mode1=bin_edges[max_freq_index]+1
13 medn=data.median()
14 mn=data.mean()
15 plt.axvline(x=float(mode1),color='red', linewidth=2, label='mode')
16 # plt.axvline(x=float(medn),color='orange', linewidth=2, label='median')
17 # plt.axvline(x=float(mn),color='black', linewidth=2, label='mean')
```

```

18
19 plt.hist(data, bins=20, rwidth=0.8)
20
21 plt.xlabel('Average Temperature of each day in 2020 in 20 bins')
22 plt.ylabel('Frequency of days with respective temperature')
23 plt.legend()
24 plt.title('Histogram of average temperature per day of a place in Himachal')
25 plt.show()

```

11.1.2 Measures of Dispersion

Measures of dispersion describe how spread out the values in a data set are.

11.1.2.1 Range

The range is the difference between the largest and smallest values in a data set. The formula for the range is:

$$\text{range} = \text{maximum value} - \text{minimum value} \quad (11.5)$$

11.1.2.2 Variance

The variance is a measure of how much the values in a data set deviate from the mean. The formula for the variance is:

$$\text{variance} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (11.6)$$

where \bar{x} is the mean and x_i represents each individual value in the data set.

11.1.2.3 Standard Deviation

The standard deviation is the square root of the variance. It is a commonly used measure of dispersion that is often used in inferential statistics. The formula for the standard deviation is:

$$\text{standard deviation} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (11.7)$$

These measures of central tendency and dispersion are important tools for understanding and summarizing data.

11.1.3 Skewness and Kurtosis

Skewness and kurtosis are measures of the shape of a data set.

11.1.3.1 Skewness

Skewness is a measure of the asymmetry of a distribution. A distribution can be symmetrical, meaning that it is equally likely to be on one side of the mean as it is on the other side, or it can be skewed, meaning that it is more likely to be on one side of the mean than the other. Skewness can be positive, negative, or zero.

1. Positive Skewness :

A distribution is positively skewed if the tail on the right side of the distribution is longer or more spread out than the tail on the left side. In other words, the mean is greater than the median, and the mode is less than the median. Positive skewness is also known as right skewness.

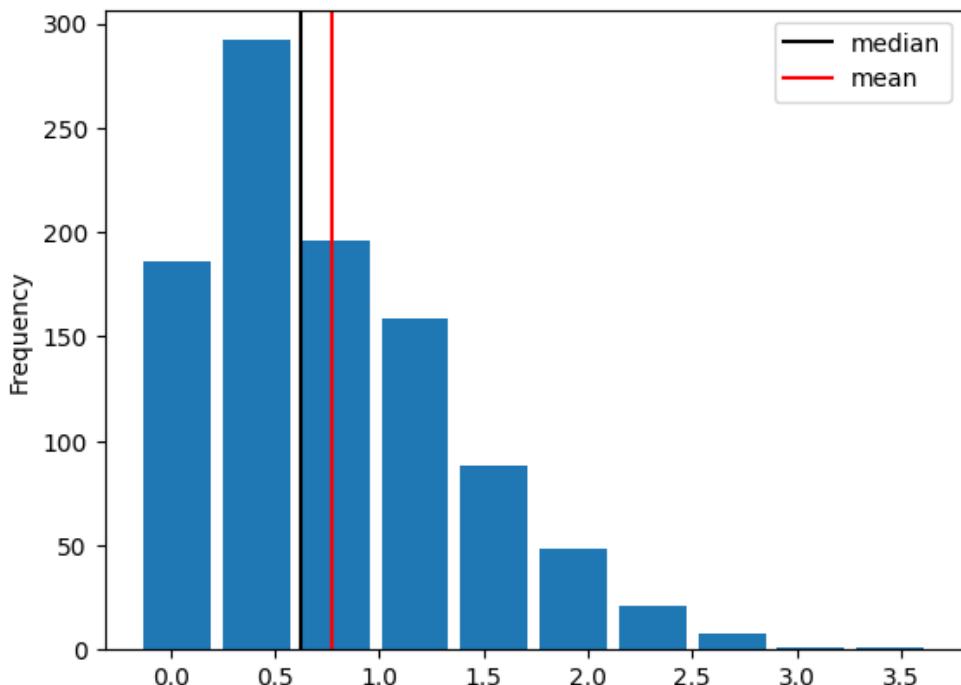


Fig. 11.2: Positively skewed data

2. Negative Skewness :

A distribution is negatively skewed if the tail on the left side of the distribution is longer or more spread out than the tail on the right side. In other words, the mean is less than the median, and the mode is greater than the median. Negative skewness is also known as left skewness.

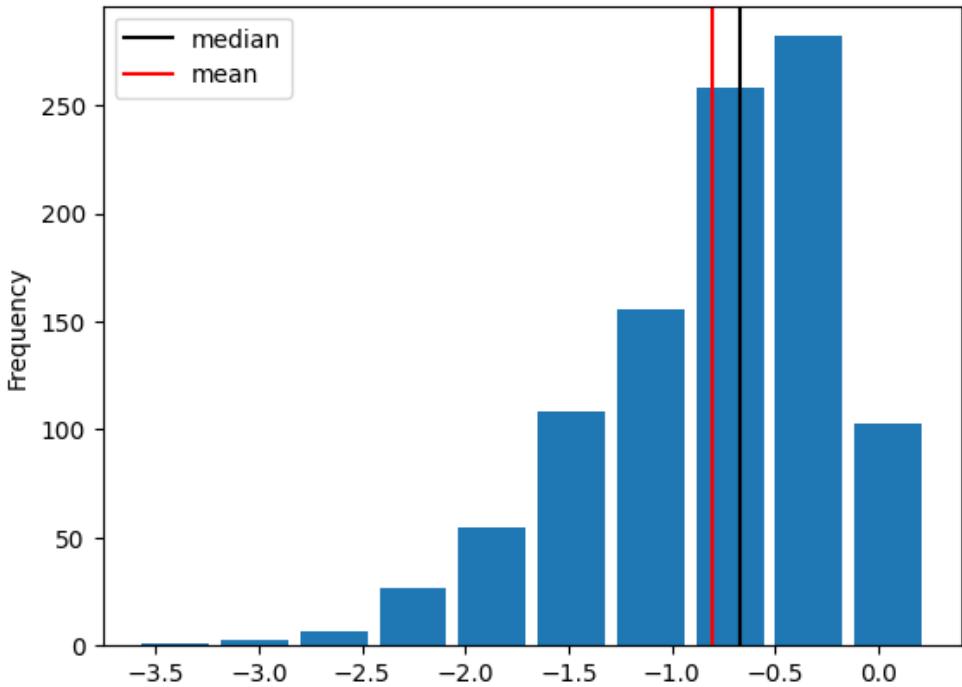


Fig. 11.3: Negatively skewed data

Figure 11.4 and Figure 11.3 were created using following code.

```

1  from scipy import stats
2  import numpy as np
3  import matplotlib.pyplot as plt
4  data= stats.skewnorm.rvs(a, size=1000)
5  plt.axvline(np.median(data), color='black', label='median')
6  plt.axvline(np.mean(data), color='r', label='mean')
7  plt.hist(data, rwidth=0.85)
8  plt.legend()
9  plt.ylabel('Frequency')
10 plt.show()
```

Changing the value of a in above code to positive give positively skewed data and with $a < 0$ we get negatively skewed data.

Skewness can be calculated using the formula:

$$\text{skewness} = \frac{\sum_{i=1}^n (x_i - \bar{x})^3 / n}{s^3} \quad (11.8)$$

where \bar{x} is the sample mean, s is the sample standard deviation, and n is the sample size.

Skewness is a useful measure of the shape of a distribution, as it can indicate whether the

distribution is normal or not. In a normal distribution, the skewness is zero, and the distribution is symmetrical. A positive skewness indicates that the distribution has a longer right tail, while a negative skewness indicates that the distribution has a longer left tail.

Let us now find the skewness of example taken in section 6.1.1.3. By finding Mean and Median of the data we can easily find the skewness of the data.

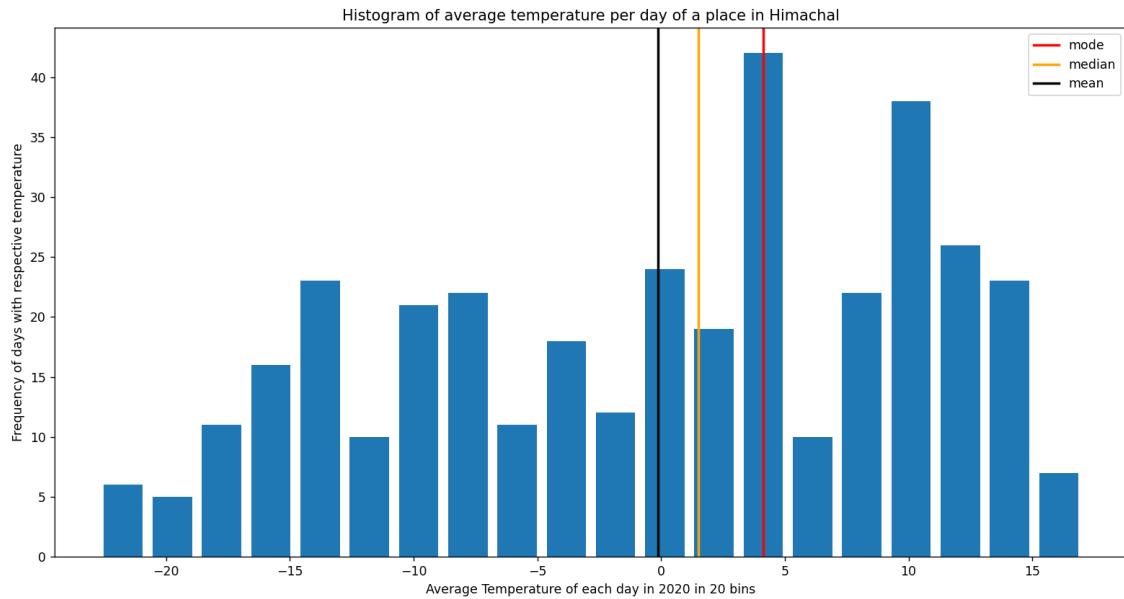


Fig. 11.4: Skewness of histogram of Average Temperature of each day in 2020

This plot was obtained using the following code:

```

1 import pandas as pd
2 import matplotlib.pyplot as plt
3 df=pd.DataFrame(pd.read_csv('Daily_Data_42_Years_1981-2022.csv'))
4
5 df2=df[df.Year==2020]
6 df2=df2.reset_index()
7 import numpy as np
8 data=df2.iloc[:,4+7*400]
9 hist, bin_edges = np.histogram(data, bins=20)
10
11 max_freq_index = np.argmax(hist)
12 mode1=bin_edges[max_freq_index]+1
13 medn=data.median()
14 mn=data.mean()
15 plt.axvline(x=float(mode1), color='red', linewidth=2, label='mode')
16 plt.axvline(x=float(medn), color='orange', linewidth=2, label='median')
```

```

17 plt.axvline(x=float(mn), color='black', linewidth=2, label='mean')
18
19 plt.hist(data, bins=20, rwidth=0.8)
20
21 plt.xlabel('Average Temperature of each day in 2020 in 20 bins')
22 plt.ylabel('Frequency of days with respective temperature')
23 plt.legend()
24 plt.title('Histogram of average temperature per day of a place in Himachal')
25 plt.show()

```

As we can notice that

$$\text{Mode} > \text{Median} > \text{Mean}$$

we can say that this data of Average temperature of each day in year 2020 is Negatively Skewed.

Also by using the formula in expression (6.8) we will get

$$\text{skewness} = -0.31$$

11.1.3.2 Kurtosis

Kurtosis is a measure of the “peakedness” or “flatness” of a probability distribution compared to the normal distribution. A distribution with high kurtosis has a sharp peak and heavy tails, while a distribution with low kurtosis has a flat peak and light tails. The most common measure of kurtosis is the fourth standardized moment, which is defined as:

$$\beta_2 = \frac{\mu_4}{\sigma^4} - 3, \quad (11.9)$$

where μ_4 is the fourth central moment and σ is the standard deviation.

A normal distribution has a kurtosis of 3, so a distribution with $\beta_2 > 3$ is said to have positive kurtosis (i.e., it is more peaked and has heavier tails than the normal distribution, also known as Leptokurtic), while a distribution with $\beta_2 < 3$ is said to have negative kurtosis (i.e., it is less peaked and has lighter tails than the normal distribution, also known as Platykurtic).

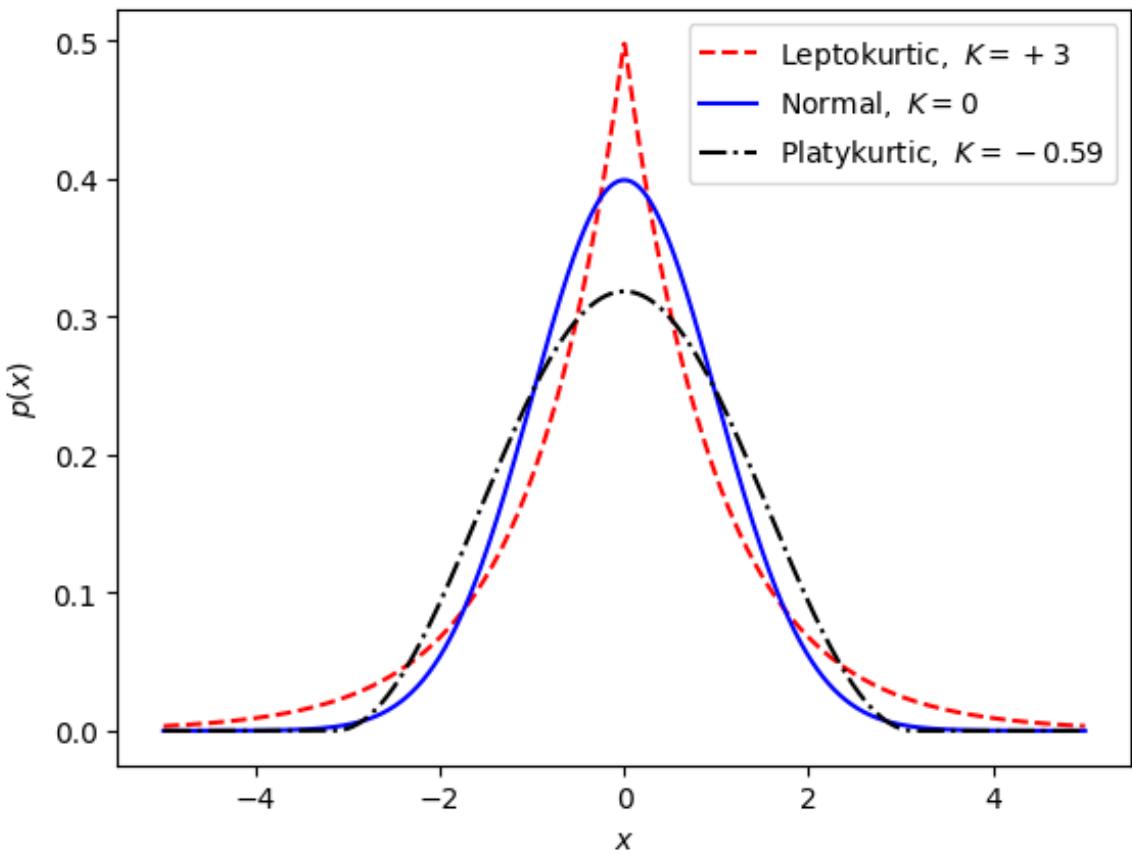


Fig. 11.5: Diffrent types of kurtosis

Figure 11.5 is obtained using following code.

```

1 import numpy as np
2 from scipy import stats
3 from matplotlib import pyplot as plt
4
5 x = np.linspace(-5, 5, 1000)
6 plt.plot(x, stats.laplace(0, 1).pdf(x), '--',
7           label=r'$\text{Leptokurtic, } K=+3$', color='r')
8 plt.plot(x, stats.norm(0, 1).pdf(x), '-',
9           label=r'$\text{Normal, } K=0$', color='b')
10 plt.plot(x, stats.cosine(0, 1).pdf(x), '-.',
11            label=r'$\text{Platykurtic, } K=-0.59$', color='black')
12 plt.xlabel('$x$')
13 plt.ylabel('$p(x)$')
14 plt.legend()
15
16 plt.show()

```

It's worth noting that there are different ways to define and interpret kurtosis, and some statisticians prefer to use alternative measures such as the excess kurtosis, which subtracts 3 from β_2 to make the normal distribution have a kurtosis of 0.

In practice, kurtosis can be used to identify outliers and assess the normality of a dataset. A high kurtosis value indicates that the dataset has more extreme values than a normal distribution, while a low kurtosis value indicates that the dataset is more spread out than a normal distribution.

However, it's important to be cautious when interpreting kurtosis, as it can be affected by the sample size and may not provide a complete picture of the distribution's shape. It's often useful to complement kurtosis with other measures of central tendency and dispersion, such as the mean and standard deviation, and to visualize the data using graphs such as histograms and box plots.

11.1.4 Boxplots and Interquartile Range

11.1.4.1 Boxplots

A boxplot is a way to visualize the distribution of a dataset. It is particularly useful when the dataset contains outliers or when the distribution is skewed. A boxplot consists of a box that represents the middle 50% of the data, with a line inside the box representing the median. The whiskers extend to the smallest and largest values within 1.5 times the interquartile range (IQR) from the box. Any data points beyond the whiskers are considered outliers.

Figure 11.7 shows an example of a boxplot. The bottom whisker extends to the minimum value that is not an outlier, which is 1. The top whisker extends to the maximum value that is not an outlier, which is 9. The box represents the middle 50% of the data, with the median line inside the box at 5. The outlier data point at 10 is plotted as a dot outside the whisker.

11.1.4.2 Interquartile Range

The interquartile range (IQR) is a measure of the spread of a dataset. It is defined as the difference between the third quartile (Q_3) and the first quartile (Q_1) of the dataset:

$$IQR = Q_3 - Q_1$$

The IQR is often used in conjunction with boxplots to determine the whisker length. The whiskers extend to the smallest and largest values within 1.5 times the IQR from the box. Any data points beyond the whiskers are considered outliers.

Boxplots and the interquartile range are useful tools for visualizing and analyzing datasets. Boxplots provide a quick summary of the distribution of the data, while the interquartile range provides a measure of the spread of the data. Together, they can help identify outliers and skewness in the data, and can provide insights into the underlying distribution.

11.1.4.3 Identifying Outliers by help of boxplot

To identify outliers in a boxplot, one needs to calculate the interquartile range (IQR) and determine the range beyond which data points are considered outliers. This range is calculated by multiplying the IQR by a constant of 1.5 and adding or subtracting it to the 25th or 75th percentiles. Any data points that fall beyond this range are considered outliers and can be marked on the boxplot. Outliers can indicate errors in data collection, measurement, or data entry, or may represent truly extreme values in the population. It is important to investigate outliers and determine whether they should be included or excluded in subsequent analyses, as their presence or absence may significantly affect the results.

11.1.5 Descriptive analysis on Rainfall data using Python

Now let's see how we can utilise the tools and theory we saw above on actual data with the help of Python.

11.1.5.1 Understanding how outliers affects the data

To see if our data has outliers the simplest way is to make a boxplot for our data. For that we will use the following code to make a boxplot of average temperature of each day in year 2022 at location 29 °N, 79 °E.

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import scipy.stats as stats
5
6 #Making a dataframe from csv containing the data
7 df=pd.DataFrame(pd.read_csv('Daily_Data_42_Years_1981-2022.csv'))
8
9 #Dataframe consisting only 2022 data
10 df_2022=df[df.Year==2022]
11 df_2022=df_2022.reset_index()
12
13 print("Mean of data before removing outlier",np.mean(df_2022.iloc[:,4]))
```

```

14 #plotting boxplot for average temperature
15 plt.boxplot(df_2022.iloc[:,4])
16 plt.ylabel("Average temperature in Degree Celsius")
17 plt.show()

```

The output of the following code is

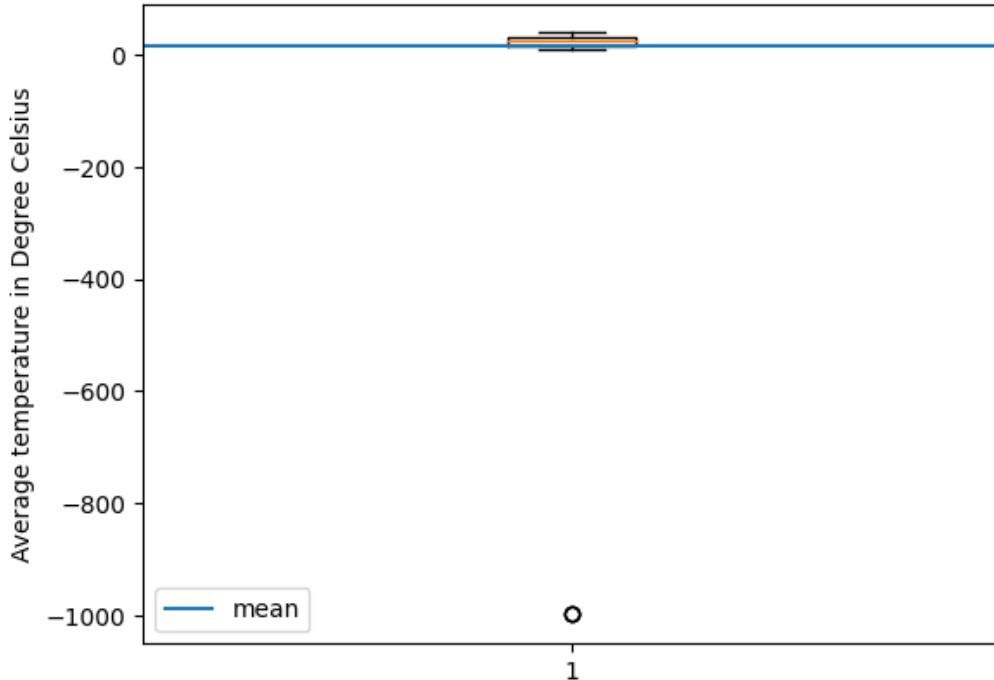


Fig. 11.6: Boxplot

As we can see in Figure 11.6 that there is one outlier which adversely affects the visualisation of Data we have and it must be removed to get better understanding of data. Also the mean we got the following data is 16.760301369863015.

To remove the outliers we will have to first identify them and replace them with some value. Its better to replace outliers with either median or mean, in our case we will be replacing outliers with median. Also here we have used the 3 standard deviation rule to find outliers but depending on data we can use IQR to find them. The code to do this is shown below.

```

1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import scipy.stats as stats
5
6 #Making a dataframe from csv containing the data
7 df=pd.DataFrame(pd.read_csv('Daily_Data_42_Years_1981-2022.csv'))
8

```

```

9 mdn=df.median() # calculating median of data
10 std=df.std() # calculating standard deviation of data
11 outliers=(df-mdn).abs()>3*std
12 df[outliers]=np.nan # replacing outliers with nan so that we can easily replace
    with median later using "fillna"
13 df.fillna(mdn,inplace=True) #replacing nan with median we calculated earlier

```

Now using the same code used to make boxplot above on the cleaned data we get following outputs.

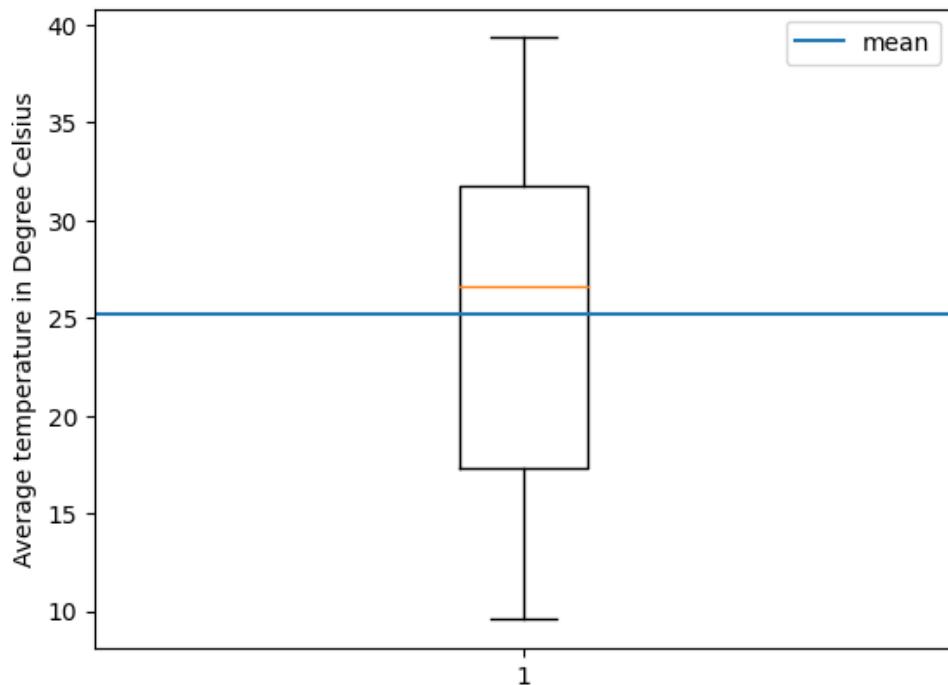


Fig. 11.7: Boxplot without outlier

The mean we got is also changed the new mean is 25.188986301369862. This shows the importance of data cleaning (removing outliers) as they affect the analysis and will prevent wrong results. The Figure 11.7 also shows how mean and median differs and also pointing out that there is skewness in the data. This also means when data is skewed it's better to take median over mean as it depicts where 50% of value lies whereas mean showing just average value which actually is not what majority of values depict and when data is not skewed we can take mean as factor to analyze.

11.1.5.2 Using different ways to analysis Data

To understand the data one way of plot might not work and one must use multiple ways to analyze the data visually and statistically. The data we have got 7 different parameters

including average temperature, maximum temperature, minimum temperature, specific humidity, precipitation, wind speed and wind direction of each day of 42 years from 1981 to 2022 for 525 different locations.

One way to visually analyze data when maximum and minimum data is given is to compare different year data with average range of few years.

Lets plot the average temperature data of whole NWH of 2017-2021 for every month and compare it with data of 2022 using following code (*Note: we will be using cleaned data in all codes used later in this chapter*).

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import scipy.stats as stats
5
6 #Making a dataframe from csv containing the data
7 df=pd.DataFrame(pd.read_csv('Daily_Data_42_Years_1981 -2022.csv'))
8
9 df_17to21=df.iloc[13145:14970] # data of 2017 to 2021
10 df_22=df[df.Year==2022] #data of 2022
11 tem=[] #monthly avg temperature of each location
12 temmax=[] #monthly max temp
13 temmin=[] #monthly min temp
14
15 for i in range(525):
16     temp=[]
17     for j in range(1,13):
18         df_temp=df_17to21[df_17to21.Month==j]
19         t=0
20         for k in df_temp.iloc[:,3+i*7]:
21             t+=k
22         t=t/len(df_temp.iloc[:,3])
23         temp.append(t)
24     tem.append(temp)
25
26 for i in range(525):
27     temp=[]
28     for j in range(1,13):
29         df_temp=df_17to21[df_17to21.Month==j]
30         t=0
31         for k in df_temp.iloc[:,4+i*7]:
32             t+=k
33         t=t/len(df_temp.iloc[:,3])
34         temp.append(t)
35     temmax.append(temp)
36
```

```

37 for i in range(525):
38     temp=[]
39     for j in range(1,13):
40         df_temp=df_17to21[df_17to21.Month==j]
41         t=0
42         for k in df_temp.iloc[:,5+i*7]:
43             t+=k
44         t=t/len(df_temp.iloc[:,3])
45         temp.append(t)
46     temmin.append(temp)
47 yaxis1=[]
48 for i in range(12):
49     avg=0
50     for j in tem:
51         avg+=j[i]
52     avg=avg/525
53     yaxis1.append(avg)
54
55 yaxis2=[]
56 for i in range(12):
57     avg=0
58     for j in temmax:
59         avg+=j[i]
60     avg=avg/525
61     yaxis2.append(avg)
62
63 yaxis3=[]
64 for i in range(12):
65     avg=0
66     for j in temmin:
67         avg+=j[i]
68     avg=avg/525
69     yaxis3.append(avg)
70 tem22=[] #monthly max for 2022
71
72 for i in range(525):
73     temp=[]
74     for j in range(1,13):
75         df_temp=df_22[df_22.Month==j]
76         t=0
77         for k in df_temp.iloc[:,4+i*7]:
78             t+=k
79         t=t/len(df_temp.iloc[:,3])
80         temp.append(t)
81     tem22.append(temp)
82

```

```

83 yaxs4=[]
84 for i in range(12):
85     avg=0
86     for j in tem22:
87         avg+=j[i]
88     avg=avg/525
89     yaxs4.append(avg)
90
91 months=['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August',
92          'September', 'October', 'November', 'December']
93 plt.plot(months,yaxs1,label='mean temp 2017-2021')
94 # plt.plot(months,yaxs2,label='max temp 2017-2021')
95 # plt.plot(months,yaxs3,label='min temp 2017-2021')
96 plt.fill_between(months, yaxs3, yaxs2, color='green', alpha=0.2)
97 plt.plot(months,yaxs4,label='max temp for 2022')
98 plt.legend()
99 plt.ylabel("Temperature")
100 plt.xticks(rotation=45)
101 plt.show()

```

The output of this code is :

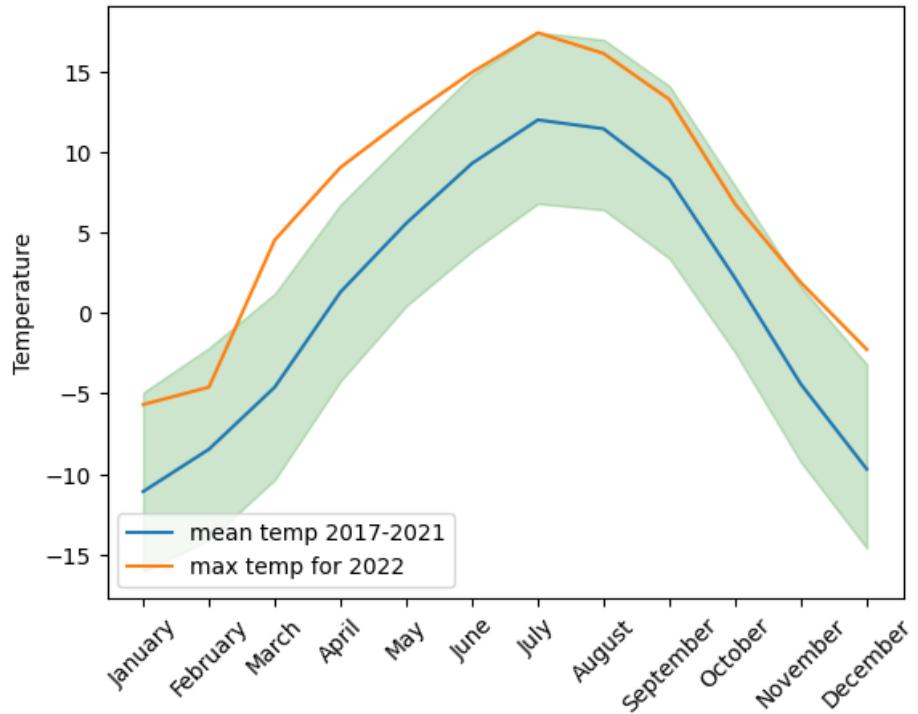


Fig. 11.8: Comparing 5 year data 2017-2021 with 2022

In figure 11.8 the green range shows min max temperature range. From this plot we can say

that maximum temperature of NWH in March-July is higher than last 5 year maximum means, or we can say that temperature in this months are increasing because of some factors.

11.1.5.3 Different parameters variance comparison

Standard deviation and variance also plays vital role when analysing data. Here is a little example of comparing variance in specific humidity and precipitation using below code.

```
1 df_22=df[df.Year==2022] #data of 2022
2 hum22=[]
3
4 for i in range(525):
5     temp=[]
6     for j in range(1,13):
7         df_temp=df_22[df_22.Month==j]
8         t=0
9         for k in df_temp.iloc[:,6+i*7]:
10            t+=k
11        t=t/len(df_temp.iloc[:,3])
12        temp.append(t)
13    hum22.append(temp)
14
15 prec22=[]
16
17 for i in range(525):
18     temp=[]
19     for j in range(1,13):
20         df_temp=df_22[df_22.Month==j]
21         t=0
22         for k in df_temp.iloc[:,7+i*7]:
23            t+=k
24        temp.append(t)
25    prec22.append(temp)
26
27 yaxis1=[]
28 for i in range(12):
29     avg=0
30     for j in hum22:
31         avg+=j[i]
32     avg=avg/525
33     yaxis1.append(avg)
34
35 yaxis2=[]
36 for i in range(12):
37     avg=0
38     for j in prec22:
```

```

39     avg+=j[i]
40     avg=avg/525
41     yaxis2.append(avg)
42
43 months=['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August',
44     'September', 'October', 'November', 'December']
45 plt.plot(months,yaxis1,label='humidity')
46 plt.plot(months,yaxis2,label='precipitation')
47 plt.axhline(np.mean(hum22),linestyle='--',color='red')
48 plt.axhline(np.mean(prec22),linestyle='--',color='green')
49 plt.legend()
50 plt.xticks(rotation=45)
51 print(np.var(prec22))
52 print(np.var(hum22))

```

The output of this code is :

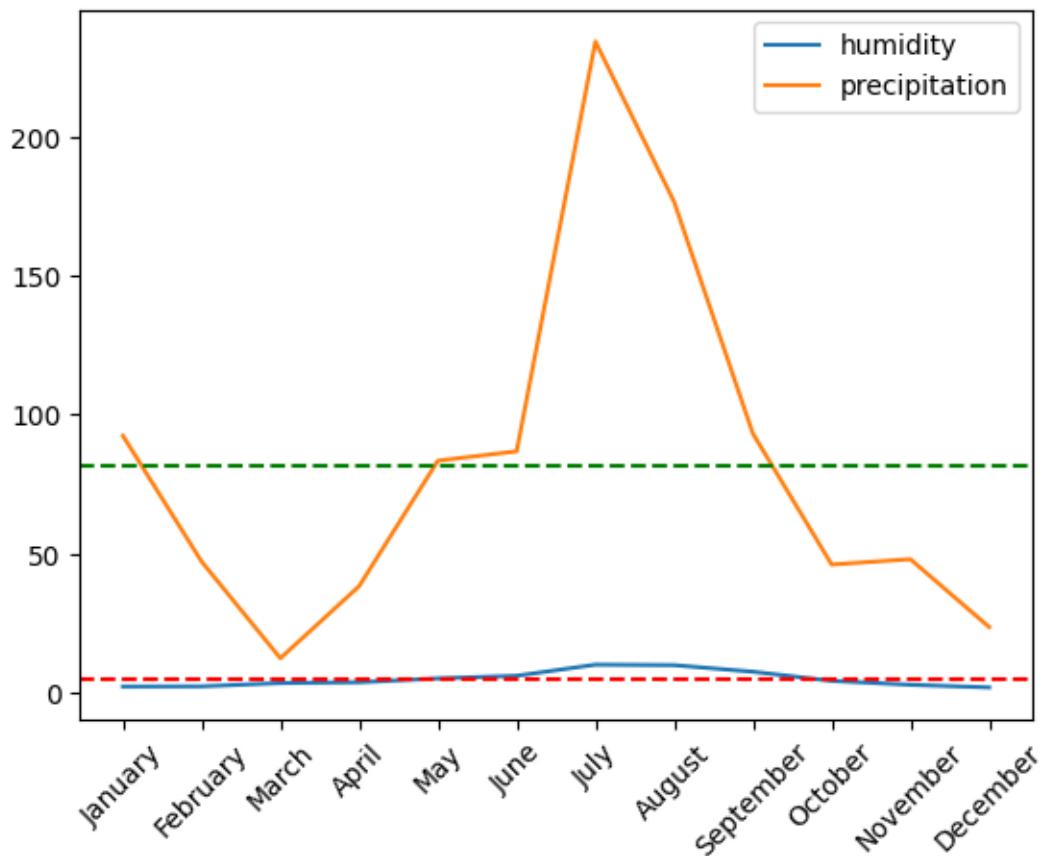


Fig. 11.9: Specific humidity and precipitation with there respective means

From the Figure 11.9 we can see that specific humidity is varying less than precipitation from

there respective mean. Same can be compared by values obtained for variance (13.785914232283062 for specific humidity and 7854.564563798689 for precipitation). Another observation we can see is that specific humidity tends to increase with precipitation.

11.1.5.4 Cumulative analysis

Some trends cannot be seen just by plotting and observing data at every point. Another important way we can use to analyze data is plotting cumulative data. For example we will be analysing the cumulative data of rainfall for last 5 years using the following code.

```

1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import scipy.stats as stats
5
6 #Making a dataframe from csv containing the data
7 df=pd.DataFrame(pd.read_csv('Daily_Data_42_Years_1981-2022.csv'))
8
9 lst=[]
10 for i in range(2017,2023):
11     dftemp=df[df.Year==i]
12     dftemp=dftemp.iloc[:,7::7]
13     dftemp["som"]=0
14     for j in range(525):
15         dftemp["som"]+=dftemp.iloc[:,j]/525
16     lst.append(dftemp["som"].cumsum())
17 for i in range(5):
18     plt.plot(range(len(lst[i])),lst[i],label=2017+i)
19 plt.legend()
20 plt.ylabel("cumulative rainfall")
21 months=['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August',
22          'September', 'October', 'November', 'December']
23 plt.xticks(np.linspace(0,365,12), months, rotation=45)
24 plt.show()
```

The output of this code is:

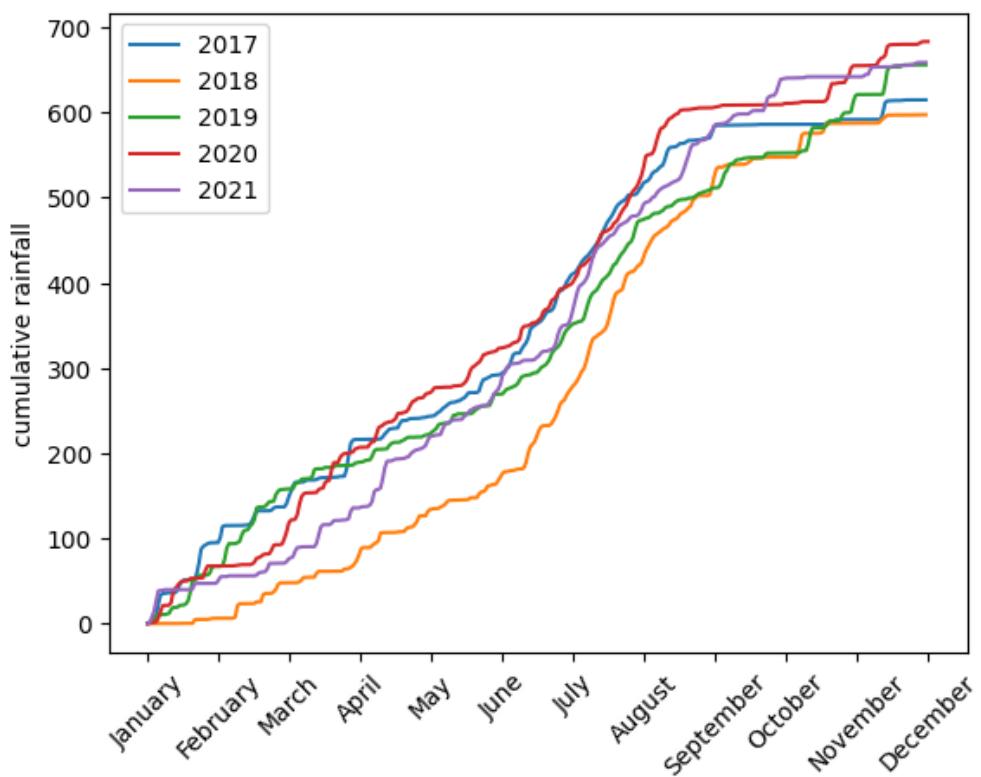


Fig. 11.10: Cumulative rainfall for last 5 years 2017-2021

Figure 11.10 shows the trend of rainfall of five years from 2017 to 2021 taken cumulatively. From this we can see how the total rainfall keep on adding with each month in a year. We can see that in 2020 and 2021 rainfall in first quarter is less comparative to 2017 and 2019 but still total rainfall at end of year are large compared to them. Also how rainfall was very less in 2018 in first half of year but sudden increase in later half.

11.1.5.5 Using spatial plots for location based data

As we have data for multiple locations in Himachal we can use geopandas to plot the map and show data using different colors. Here is example code to plot average yearly rainfall and average windspeed in 42 years in each location in NWH.

```

1 import geopandas as gpd
2 import pandas as pd
3 import numpy as np
4 import matplotlib.pyplot as plt
5 import scipy.stats as stats
6
7 #Making a dataframe from csv containing the data
8 df=pd.DataFrame(pd.read_csv('Daily_Data_42_Years_1981 -2022.csv'))

```

```

9  locations=[]
10 for i in range(4,3679,7):
11     # print(df.columns[i])
12     temp=df.columns[i].split("_")
13     if len(temp[2])==4:
14         lati=int(temp[2])/100
15     else:
16         lati=int(temp[2])
17
18     if len(temp[3])==4:
19         longi=int(temp[3])/100
20     else:
21         longi=int(temp[3])
22
23     locations.append([lati,longi])
24 shapefile=gpd.read_file("4-17-2018-899072.shp")
25 print(type(shapefile))
26 def Map_plot(longitude,latitude,value):
27     fig,ax=plt.subplots(figsize=(5,5))
28     plt.scatter(x=longitude , y = latitude,c = value)
29     plt.xlabel('Longitude')
30     plt.ylabel('Latitude')
31     shapefile.plot(ax =ax,color='black')
32     plt.colorbar()
33     plt.show()
34 Map_plot([x[1] for x in locations],[x[0] for x in locations],[sum(df.iloc[:,7+i]*7])/42 for i in range(525)])
35 Map_plot([x[1] for x in locations],[x[0] for x in locations],[np.mean(df.iloc[:,8+i*7]) for i in range(525)])

```

Listing 11.1: Code for spatial maps

The output of this code is :

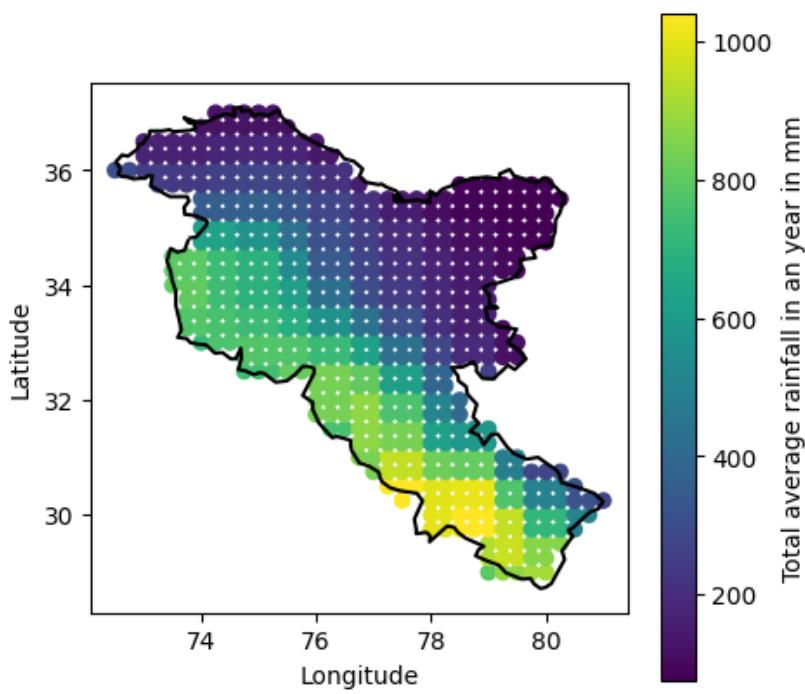


Fig. 11.11: Total average rainfall in an year in mm

Figure 11.11 was for rainfall per year average in mm and Figure 11.12 is for windspeed average of a day.

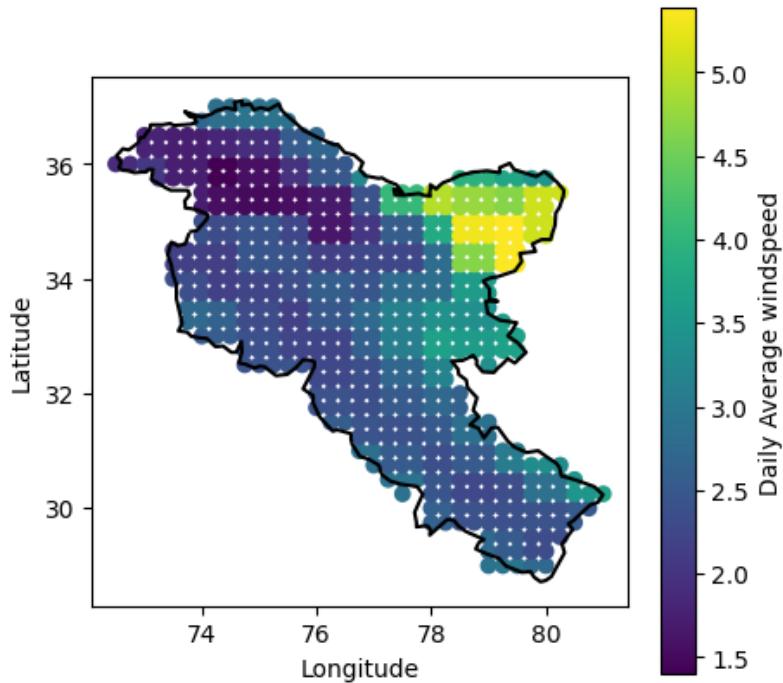


Fig. 11.12: Daily average windspeed

From both Figure 11.11 and 11.12 we can see that there is no clear relation between rainfall and windspeed. If we plot the daily average specific humidity map with same function used in previous code we will get output as presented in Figure 11.13.

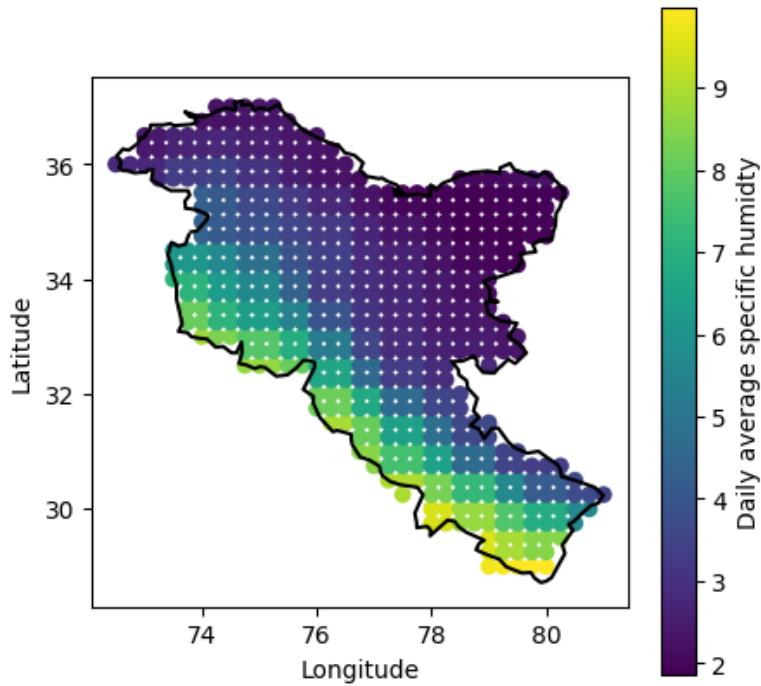


Fig. 11.13: Daily average specific humidity

Comparing Figure 11.11 and Figure 11.13 we can clearly see some similar pattern and hence confirming a correlation between them. To clearly see the correlation let's plot the correlation between rainfall and specific humidity by adding the following snippet in Code 11.1 .

```
1 Map_plot([x[1] for x in locations],[x[0] for x in locations],[np.cov(df.iloc[:,6+i*7],df.iloc[:,7+i*7])[0][0] for i in range(525)],"Correlation Between rainfall and specific humidity")
```

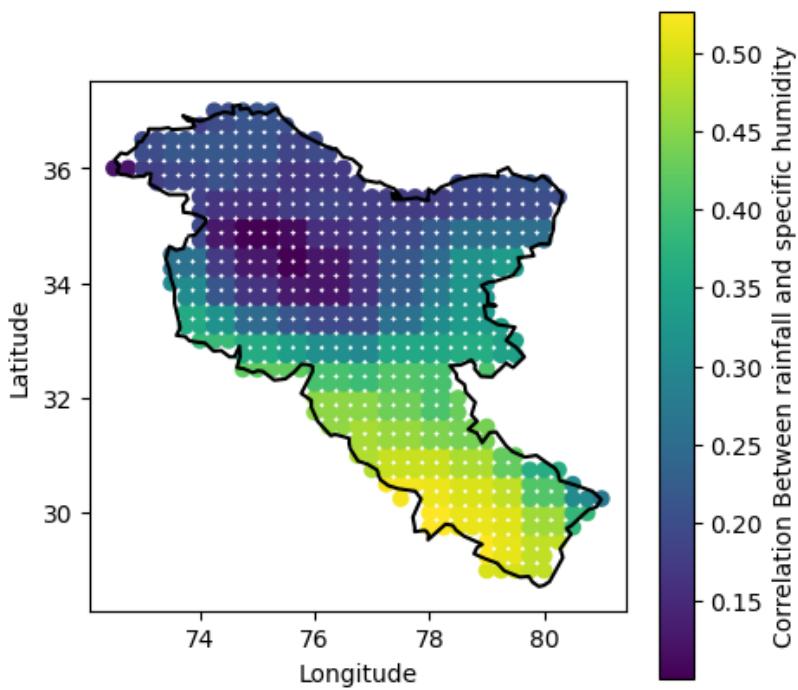


Fig. 11.14: Correlation Between rainfall and specific humidity

Figure 11.14 clearly shows that at lower latitudes the correlation between rainfall and specific humidity is higher.

The above techniques and examples shows how we can use Python's help to do Descriptive analysis on Data and how we can infer from the outputs.

11.2 Lab Assignment 10

Question 1. Create a python function that takes two input years, `year1` and `year2`, and returns a data frame containing statistical measures such as minimum, maximum, mean, variance, standard deviation, skewness, kurtosis, and interquartile range (IQR) for both temperature and precipitation datasets?

Question 2. Create a python function that accepts four parameters: `Latitude`, `Longitude`, `year1`, and `year2`. This function generates scatter plots, line plots, box plots, histogram plots, and violin plots based on the given input for both data sets. The goal is to derive meaningful inferences from the generated plots. Additionally, if any irregularities (e.g maybe min temp having a year) are detected in a particular year, further exploration will be conducted at the

monthly level?

Question 3. Generate spatial plots (e.g., fig-01) illustrating minimum, maximum, mean, variance, standard deviation, skewness, kurtosis, and interquartile range (IQR) for both temperature and precipitation datasets. Additionally, create a correlation plot to explore the relationship between the two datasets. Analyze the correlation plot to gain insights and identify any patterns or trends?

Question 4. Generate spatial plots illustrating clustering for both the dataset .Use K Means clustering and for choosing optimal number of clusters use elbow method. Analyze how the same cluster points are related by statistical measures such as minimum, maximum, mean, variance, standard deviation, skewness, kurtosis, and interquartile range (IQR)?

NOTE In the zip file two datasets are attached containing NWH (northern western himalayas) region temperature and precipitation(rainfall).

NWH Region - In the dataset, NWH region includes Jammu and Kashmir,Ladakh,Himachal Pradesh and Uttarakhand.

DATA - Temperature.csv contains temperature of NWH region in degree celsius. Precipitation .csv contains rainfall data in mm.

To generate spatial plots 4 files are given, You have to upload all four files on your system but have to call only dot shp file.

For spatial plots you can use GEOPANDAS.

To install geopandas:

```
1 pip install geopandas
```

To read NWH region shape file:

```
1 import geopandas as gpd
2 shapefile = gpd.read_file("/content/4-17-2018-899072.shp")
```

To plot shape file:

```
1 shapefile.plot (ax=ax, color='red'))
```

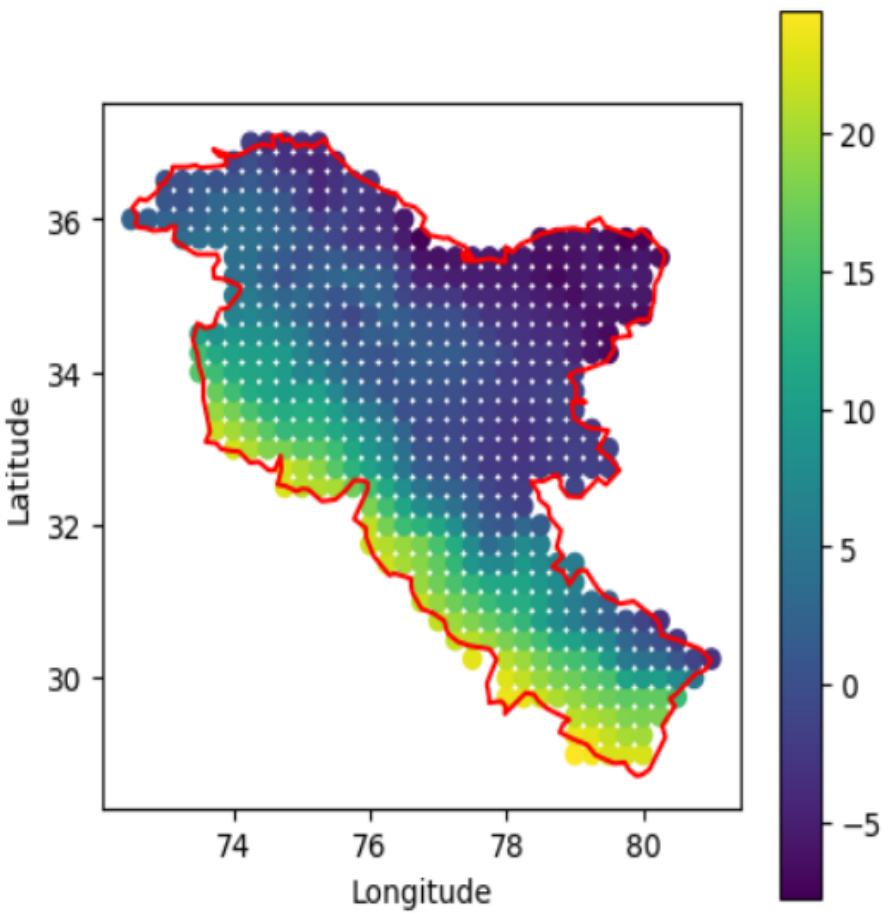


Fig. 11.15: (For mean of temperature (yearwise))

K means clustering -

```
1 from sklearn.cluster import KMeans
2 X = data
3 kmeans = KMeans(n_clusters=2).fit(X)
4 kmeans.labels_
5 kmeans.predict()
6 kmeans.cluster_centers_
```

Elbow Method to detect optimal numbers of clusters -

```
1 sse = {}
2 For k in range(1, 10):
3     kmeans = KMeans(n_clusters=k, max_iter=1000).fit(data)
4     data["clusters"] = kmeans.labels_
5     #print(data["clusters"])
6     # Inertia: Sum of distances of samples to their closest cluster center
7     sse[k] = kmeans.inertia_
8 plt.figure()
9 plt.plot(list(sse.keys()), list(sse.values()))
10 plt.xlabel("Number of cluster")
11 plt.ylabel("SSE")
12 plt.show()
```

