

School of Engineering and Applied Science (SEAS), Ahmedabad University

BTech(ICT) Semester IV: Probability and Random Process (MAT202)

Special Assignment-Final Report

- Group No. : SB - 4
- Group Members:
Harshil Mehta - AU1841010
Raj Mehta - AU1841018
Manav Patel - AU1841037
- Project Area : Biology
- Project Title : Disease Prediction and Analysis Of Schizophrenia

1 Introduction

1.1 General discussion

Schizophrenia is a chronic mental disorder. Many symptoms like Anger (Extreme), Mental confusion, Disorganised behaviour, unexpected memory loss leads to a disease named Schizophrenia. It's almost a non-curable disease. In some case it can improve conditions with help of happy environment, Psychotherapy and by some very expensive rehabilitation. On an average, only about 1% of the total population is affected by this disease. Mostly it occurs due to psychosocial factors, problems during birth, malnutrition before birth and tragic (obsessive) events caused by the environmental factors faced by people. It can also cause Mental Hallucinations.

1.2 General related work

How will we view schizophrenia in 2030? [1] Schizophrenia today is a chronic, frequently disabling mental disorder that affects about one per cent of the world's population. After a century of studying schizophrenia, the cause of the disorder still remains unknown. [2] Therefore there is an essential need for a probabilistic model which could quantify on the likeliness of a person having this disorder. [3] Also, there are theories and paperworks which emphasis on the complex modeling of schizophrenia but lack a basic inference - deriving model which could lead to better understanding. [4] The major modeling is done through Hardy - Weinberg Equilibrium. [5] Also, not much has been done in this field regarding probabilistic modeling and based on Inference deriving.

Other related articles:

- Bhattacharya, and Souvik. "Markov Chain Model to Explain the Dynamics of Human Depression." *Journal of Nonlinear Dynamics*, Hindawi, 18 Mar. 2014, www.hindawi.com/journals/jndy/2014/107164/.
- Vivian-Griffiths, Timothy, et al. "Predictive Modeling of Schizophrenia from Genomic Data: Comparison of Polygenic Risk Score with Kernel Support Vector Machines Approach." *Wiley Online Library*, John Wiley and Sons, Ltd, 4 Dec. 2018, www.onlinelibrary.wiley.com/doi/full/10.1002/ajmg.b.32705.
- "US20150224120A1 - Compositions and Methods for Treating Hyperprolinemia-Associated Mental Disorders." *Google Patents*, Google, www.patents.google.com/patent/US20150224120A1/en.
- Learning, Lumen. "Biology for Majors I." *Lumen*, www.courses.lumenlearning.com/wm-biology1/chapter/reading-penetrance-and-expressivi

1.3 Close related Work

Main articles:

- **Base article:**

Paek, Myung Jae, and Ung Gu Kang. "How Many Genes Are Involved in Schizophrenia? A Simple Simulation." *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, Elsevier, 24 Apr. 2012, www.sciencedirect.com/science/article/abs/pii/S0278584612000875.

- "Allele Frequency & the Gene Pool (Article)." *Khan Academy*, Khan Academy, www.khanacademy.org/science/biology/her/heredity-and-genetics/a/allele-frequency-th
- Insel, Thomas R. "Rethinking Schizophrenia." *Nature News*, Nature Publishing Group, 10 Nov. 2010, www.nature.com/articles/nature09552?page=7.

1.4 Motivation

The main motivation behind our current work is due to minimal quantitative work/analysis done in this subject. The Literature and Data analysis done by previous research has been based more upon Qualitative analysis of the disease. Higher algorithm like speech recognition, concepts of Deep - learning and Machine learning have been used to study cases and prepare data-sheets. So, our basic idea is to quantify schizophrenia and its chances simple concepts of Probability and Random Process and basic mathematical modeling.

1.5 Problem Case Study

Prediction of Schizophrenia is an uncertain problem.

In our probabilistic model, one of the category which affects Schizophrenia i.e. **Genes** is studied (and quantified).

Here some theories and concepts are applied to solve the uncertain problem.

There are different scenarios in which this is classified:

Consider Total population as X .

Now Let XP be the patient population, so the remaining population will be $X - XP$.

Now the cases that are to be calculated are quite obvious for prediction of Schizophrenia in offspring:

Mating Chances (Or Marriage Possibilities)

- 1 person from XP and 1 person from $(X - XP)$
- both the person from XP .
- both the person from $X - XP$. (For more details refer algorithm Section)

Theories used in this model:

Hardy Weinberg Equilibrium, Strachan Hierarchy, Total Probability and Relative frequency Approach.

In this work, the Hardy Weinberg Equilibrium and Joint Probability theory is applied along with the calculations of N and T using Total Probability Theorem. Where N is number of Genes and T is pathogenic genes involved in Schizophrenia. Strachan Hierarchy is used to plot bar chart and show comparison.

There are few assumptions of this theory:

Assumptions:

- Large population of Genes.
- No selection. (No biasing)
- No mutation. (No new genes can be produced from any combination of the two genes)
- Stable Allele Frequency.
- There are only two types of Genes involved. (For the sake of simplicity, just consider one of them as Dominant Gene and other as Recessive Gene)

2 DATA ACQUISITION

Yes, our special Assignment is Data Dependent.

So, as we don't have any particular data sheets. As we don't have available data/statistics online.

We have generated random data-sets according to realistic figures given in :

- Paek, Myung Jae, and Ung Gu Kang. "How Many Genes Are Involved in Schizophrenia? A Simple Simulation." *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, Elsevier, 24 Apr. 2012, www.sciencedirect.com/science/article/abs/pii/S0278584612000875.

The random data generated is of no. of genes in a particular person with the dominance of its alleles'.

After generating this data set, we calculate the likeliness according to the formulas.

3 Probabilistic Model Used/ PRP Concept Used

Probabilistic modelling

Assumptions based on data: (are shown above)

- 1) Total population of the world is approximately equal to 7 billion.
 - 2) Total schizophrenia patients in the world are approximately 1 % of the total population.
- ∴ Total patients = 7million,

Now,

Dividing the population into two part:

a) Schizophrenic patients which constitutes the **Patient population** denoted as **PP**.

b) Non - Schizophrenic patients which constitutes the **Non Patient population** denoted as **NP**.

Now, Two odds of the offspring having Schizophrenia is determined by **Hardy Weinberg equilibrium** denoted as **HWE**

According to HWE, total probability of NP and PP is equal = 1.

$$\therefore \boxed{PP + NP = 1}$$

Now, the genes of the offspring are the combinations of genes of both the parents.

∴ squaring both sides we get

$$\boxed{PP^2 + 2 \times PP \times NP + NP^2 = 1}$$

As the schizophrenic allele is dominant compared to the non-schizophrenic allele.

∴ Probability of the offspring having schizophrenia is $\text{Prob}(PP*PP) + \text{Prob}(2*NP*PP)$ and subtracting this from the total probability we get the probability of the offspring not having schizophrenia. which is $\text{Prob}(2*NP*NP)$.

Therefore,using the equations of **HWE** there are total 3 cases possible.

a) Offspring is the result of mating of two parents where both the parents belong to patient population (PP,PP).

Here allele frequency of the dominant allele = 0.80 and allele frequency of recessive allele is 0.20

b) Offspring is the result of mating of two parents where both the parents belong to non patient population (NP,NP).

Here allele frequency of the dominant allele = 0.99 and allele frequency of recessive allele is 0.01

c)Offspring is the result of mating of two parents where one of the parent belong to non patient population(PP) and other one belong to (NP).

Here allele frequency of the dominant allele = 0.99 and allele frequency of recessive allele is 0.01.

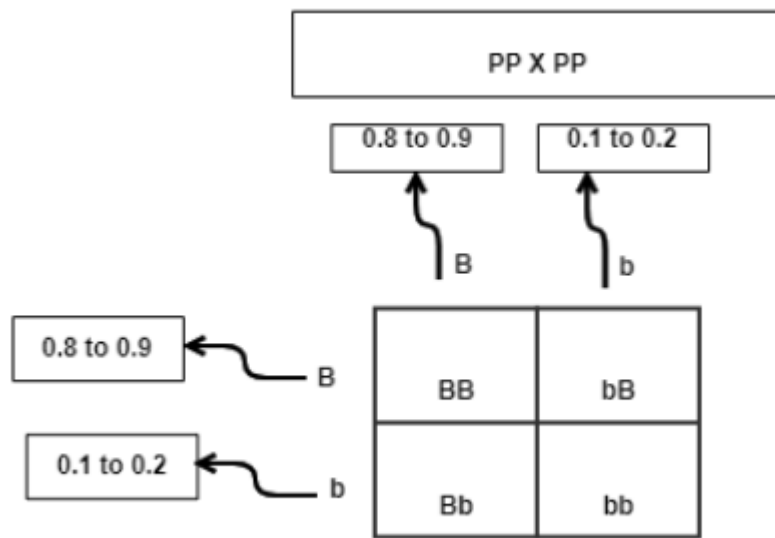


Figure 1: case a): PP X PP

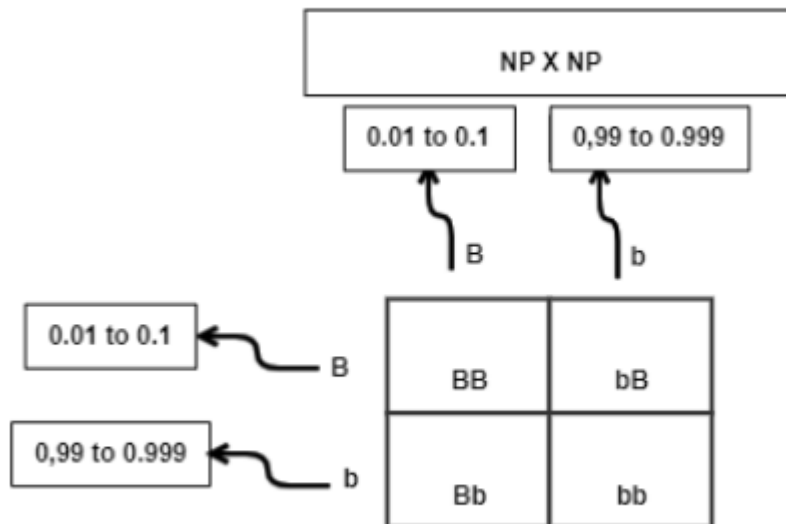


Figure 2: case b): NP X NP

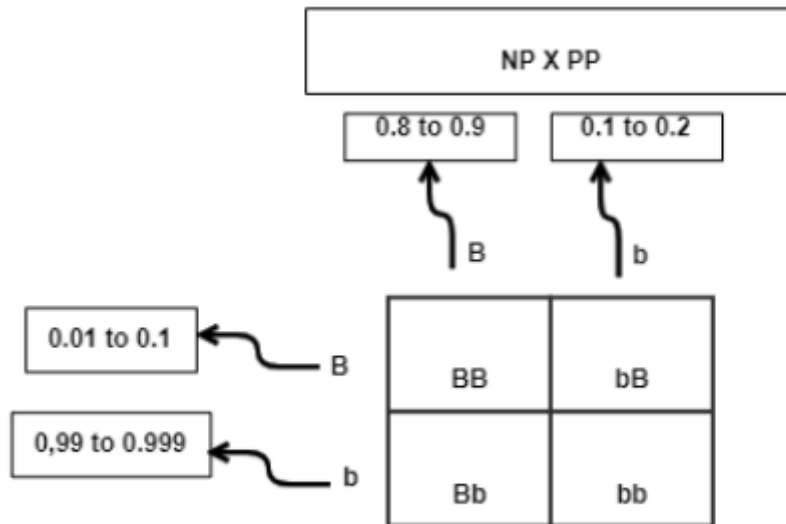


Figure 3: case c): NP X PP

∴ In all three cases,

$$\text{Chances of Schizophrenia} = \text{Allele_frequency} * (BB) + 2 * \text{Allele_frequency} * (Bb)$$

And

$$\text{Chances of the offspring of not having schizophrenia} = \text{Allele_frequency} * (bb).$$

∴ Both these values are different for all three cases.

4 Pseudo Code/ Algorithm

Pseudo Code

Algorithm 1 Function For Obtaining Probability according Hardy-Weinberg Equilibrium

```
1: function SOLVE_FORN_T(frequency,ratio)
2:   return ratio $\times$ (frequency * frequency + 2*(frequency)(1 - frequency))           ▶ Using HWE
3: end function
```

Algorithm 2 Function For Generating Random Data sets between input range

```
1: function GENERATERANDOMDATA(ratio,startPoint,endPoint)
2:   for  $i = 1, 2 \dots 100$  do
3:     Random_Value = Random(20, 160)                                           ▶ Uniform random data
4:     Threshold_Frequency.add(Random_Value)
5:     Probability_ans.add(Solve_forN_T(Random_Value,ratio))
6:   end for
7: end function
```

Algorithm 3 Function For Solving Case-wise with inter Population mating

```
1: function SOLVE_PROBABILITY_CASE(Allele_A_freq,Allele_B_freq)
2:   Offspring_Probability = [(Allele_A_freq * (1 - Allele_B_freq)) + (Allele_A_freq * Allele_B_freq) + ((1
   - Allele_A_freq)*(1 - Allele_B_freq))]
3:   return Offspring_Probability                                           ▶ Using Joint Probability Theorem
4: end function
```

We have also, used user defined functions and inbuilt functions for plotting different graphs and plots by programming.

Next is the algorithm to obtain bar graph deduced from strachan's hierarchy.

Algorithm 4 Function For Strachan Hierarchy

```
1: function SHOW_STRACHAN_HIERARCHY(Probability)
2:   performance = [Probability, Probability*0.92, Probability*0.74, Probability*0.53,
   Probability*0.48, Probability*0.31, Probability*0.28]
3:   plot performance[]
4:   return performance ► Hierarchy
5: end function
```

Algorithm 5 Procedure For Solving each Probabilistic cases with inter Population mating

```
1: procedure MAIN
2:   GenerateRandomData((0.01 * 0.01), 0.8, 0.99) ► For case of PP vs PP
3:   Rel_Probability1 ← solve_probability_case(0.8, 0.8)
4:   GenerateRandomData((0.99 * 0.99), 0.00001, 0.01) ► For case of NP vs NP
5:   Rel_Probability2 ← solve_probability_case(0.01, 0.01)
6:   GenerateRandomData((0.99 * 0.01), random.uniform(0.001, 0.01), random.uniform(0.8, 0.95)
7:   ► For case of PP vs NP
8:   Rel_Probability3 ← solve_probability_case(0.01, 0.8)
9: end procedure
```

5 Coding and Simulation

5.1 Simulation Framework

Assumptions:

- Total Population = 7 billion.
- Patient Population(Affected by Schizophrenia) = 1 % of Total Population
- Non Patient Population = Total - Patient Population
- 3 mating cases of case studies.

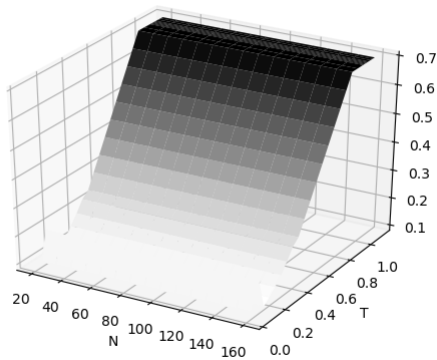
Now, The Number of genes(N) are generated randomly between 20 and 160 according to the base article.

The allele frequency(T) for the dominant gene is randomly generated between 0.8 to 0.95 and the allele frequency for the recessive gene is generated between 0.00001 to 0.01 according to the base article.

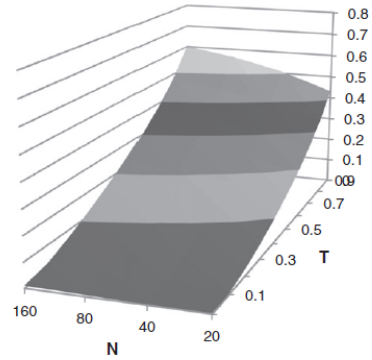
Both Relative and Absolute Probabilities are calculated.

5.2 Reproduced Figures

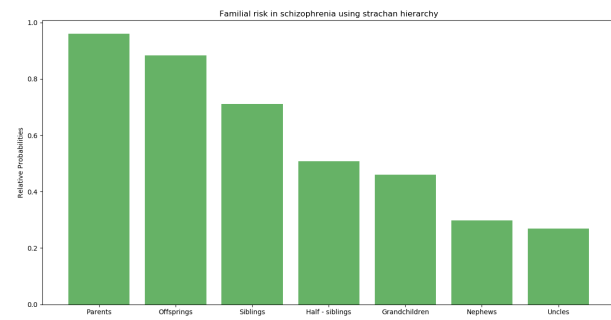
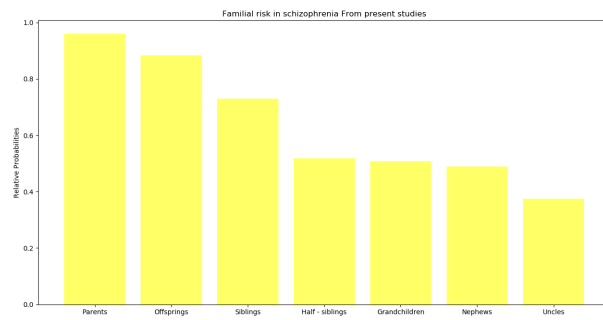
We have used Matplotlib in Python to produce desired plots.



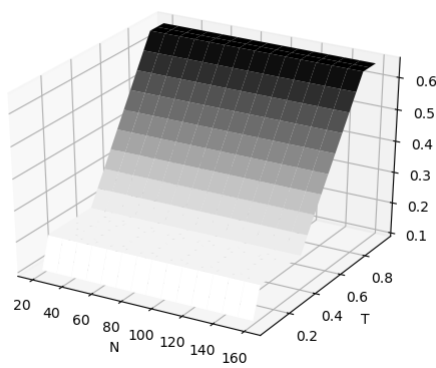
a) Allele frequency in NP according to N & T



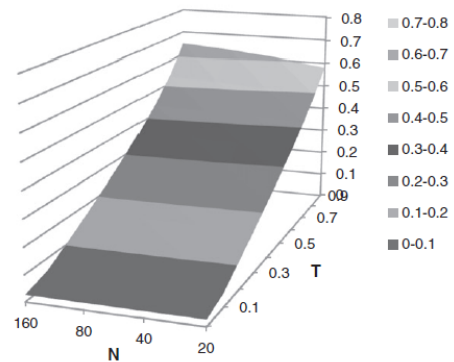
Allele frequency in NP according to N and T



Familial risk in Schizophrenia using present studies VS strachan hierarchy only for the case of PP vs PP



b) Allele frequency in PP according to N & T



Allele frequency in PP according to N and T

6 Inference Analysis/ Comparison

Inference 1:

- As per the graphs and outputs obtained from the code, we can infer that as the with Increase in Number of genes and Allele Frequency \rightarrow Increase in Probability of having Schizophrenia.
- But, as the Allele is of dominant nature, increase in Allele Frequency is responsible for more increase in Probability than increase in Number of genes.
- Dependency of Probability is governed by:
Number of genes(N) \ll Allele Frequency(T).

Inference 2:

- As, we can observe from the Probability of each case, it is clear that by taking into account the Person from the Patient Population, the Probability of having Schizophrenia increases exponentially.
- Probability(PP vs PP) \gg Probability(PP vs NP) \ggg Probability(NP - NP).
Here, PP denotes the person from patient Population and NP denotes the person from Non - Patient Population.

Inference 3:

- The Hierarchy which was prepared by Strachan in 1990 has shown relation between risk in families spread through Schizophrenia.
- Our studies has also shown very similar Hierarchy to Strachan's Hierarchy using bar graphs.

Thus, from this inferences all of our study can be expressed in a more user - understandable language.

These thesis can help to analyze and quantify Schizophrenia in a more scrutinized manner.

7 Contribution of team members

7.1 Technical contribution of all team members

Tasks	Harshil Mehta	Raj Mehta	Manav Patel
Task-1	Coding(Part A)	Mathematical analysis	Resource/Info gathering
Task-2	Simulation(bar charts)	Probabilistic Modeling	Coding(Part B)
Task-3	Data-Correction	Testing	Inferencing
Task-4	Inference-derivation	Simulation(graphs)	Case analysis

7.2 Non - Technical contribution of all team members

Tasks	Harshil Mehta	Raj Mehta	Manav Patel
Task-1	Introduction	Probabilistic Model Used	Problem/Case Study
Task-2	Pseudocode/Algorithm	Simulation Framework	Data Acquisition
Task-3	Inference analysis/Comparision	Reproduced Figures (Part1)	Reproduced Figures (Part2)

8 REFERENCES

- [1]. Marco Procopio(2005): *Does god play dice with schizophrenia? A probabilistic model for the understanding of causation in mental illness*
- [2]. Myung JaePaeka, Ung GuKangI(2012): *How many genes are involved in schizophrenia? A simple simulation*
- [3]. Ying Guo, DuBois Bowman and Clinton Kilts(2008):*Predicting the Brain Response to Treatment Using a Bayesian Hierarchical Model With Application to a Study of Schizophrenia*
- [4]. Elsevier Volume 127, Issues 1–3, April 2011, Pages 115-122 :*Probabilistic learning and inference in schizophrenia*
- [5].GENETIC EPIDEMIOLOGY , 05 June 2010: *Impact of Hardy–Weinberg equilibrium deviation on allele-based risk effect of genetic association studies and meta-analysis*