International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

STATISTICAL GENETICS

ARJUN GOVIND[1], Ms. AYESHA TARANNUM[2]

[1] 222PM40 St Joseph's University
[2] Under the Guidance of
Dept. Of Statistics
St Joseph's University

1.INTRODUCTION

Statistical genetics and genomics is a category of statistics that focuses on extensive knowledge of genetics. and the software used to assess the variability of genes and genomic expression in humans.It uses statistical approaches to study gene.It is a field found in biostatistics.

By using statistical techniques,it helps researchers to find the interplay between genetic variations ,environmental factors and phenotypic outcomes.It helps in genetic disorder risk prediction and finding other complexities based on genes.These methods involves collecting phenotypic and genotypic data from a data set to quantify the risk.Pearson's chi square test for goodness of fit is one of the most significant approaches used to find the deviation between the expected and observed values.This approach is used in examining Mendelian ratios.

The scientific discipline of statistical genetics is focused on the creation and use of statistical techniques for deriving conclusions from genetic data. The development of theory or methodology to enable research in one of three related areas is the typical focus of statistical genetics research:

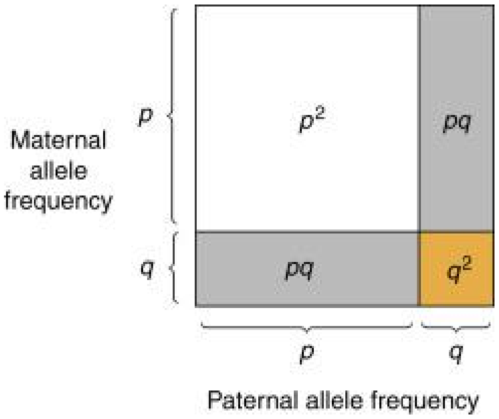
- Investigating how evolutionary processes affect genetic variation among species is known as population genetics.
- Investigating how genes affect disease is known as genetic epidemiology.
- Investigating how genes affect "normal" phenotypes is known as quantitative genetics.

Molecular biologists, physicians, geneticists, and bioinformaticians are frequently in close collaboration with statistical geneticists. Among the subfields of computational biology is statistical genetics.

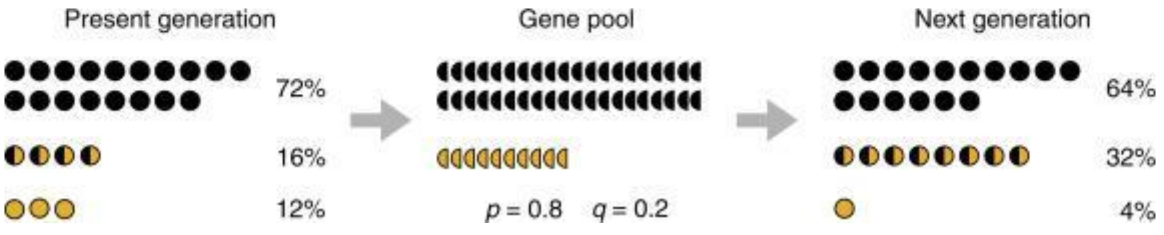
High-throughput techniques are developing quickly, making it possible for researchers to gather vast amounts of genetic data. This data can be used to answer more biological questions and generate more potent treatment plans for human diseases. Integrated genomic data analysis is becoming more and more common since different kinds of genomic data are frequently available both within and between research. It is possible to combine various genomic data types within the same set of samples or the same genomic data type across many studies.

2.HARDY-WEINBERG EQUILIBRIUM (HWE)

In a randomly mating population, the Hardy-Weinberg principle establishes a connection between genotype and allele frequencies. Assuming segregation of two alleles (*A* and *B*) at a single locus, where allele *A* frequency is denoted by *p* and allele *B* frequency by *q*, the principle predicts genotype frequencies of p^2, $2pq$, and q^2 after one generation of random mating. Hardy-Weinberg equilibrium occurs when genotype frequencies remain constant in the absence of evolutionary factors such as natural selection. Several assumptions underlie the Hardy-Weinberg principle: (1) random mating, without population structure or mating preferences based on genotype frequencies; (2) no natural selection; (3) large populations, minimizing genetic drift; (4) no gene flow or migration; (5) no mutation; and (6) autosomal locus. Deviations from Hardy-Weinberg proportions can occur when these assumptions are violated.



Looking at a real-world example can offer additional insight. Imagine a population with 18 individuals homozygous for the *AA* genotype, 4 heterozygotes with the *AB* genotype, and 3 individuals homozygous for the *BB* genotype. In this population, 20% of the alleles are *B* alleles, while 80% are *A* alleles. After one generation of random mating, the population exhibits Hardy-Weinberg proportions: 16 individuals with the *AA* genotype, 8 individuals with the *AB* genotype, and 1 individual with the *BB* genotype. It's important to note that allele frequencies remain constant.



Hardy–Weinberg illustration. Contributors to the gene pool are BB homozygotes (gold circles), AB heterozygotes (black and gold circles), and AA homozygotes (black circles). Alleles denoted as A and B are represented by black and gold half-circles, respectively. Following one generation of randomized mating, Hardy-Weinberg proportions are found.

The Hardy-Weinberg principle has several implications for evolution. Firstly, in large, randomly mating populations, genetic diversity is maintained. Secondly, the Hardy-Weinberg principle allows us to determine the frequency of individuals carrying a recessive gene. Thirdly, it's important to note that dominant alleles are not necessarily the most common in a population. Additionally, the principle suggests that heterozygous individuals are more likely to carry rare genes compared to homozygous individuals, due to the lower frequency of the recessive allele q^2 compared to the frequency of heterozygotes ($2pq$) when q is small.

The generalized Hardy-Weinberg principle extends to genes with more than two alleles and polyploid organisms. By using the multinomial expansion formula $(p_1 + ... + p_k)^n$, where *n* represents the number of sets of chromosomes in a cell and *k* is the number of segregating alleles, equilibrium genotype frequencies can be derived. For example, in tetraploid organisms (*n* = 4) with two alleles (*k* = 2), the expected genotype frequencies are: p_1^4 *(AAAA),* $4p_1^3p_2$ *(AAAB),* $6p_1^2p_2^2$ *(AABB),* $4p_1p_2^3$ *(ABBB),* and p^4 *(BBBB)*

. Similarly, in diploid organisms (*n* = 2) with three alleles (*k* = 3), the anticipated genotype frequencies are: p_1^2 *(AA),* p_2^2 *(BB),* p_3^2 *(CC),* $2p_1p_2$ *(AB),* $2p_1p_3$ *(AC),* and $2p_2p_3$ *(BC).* In each case, the sum of the genotype frequencies equals one.

It's important to note that while the Hardy-Weinberg principle can be applied to genes on sex chromosomes, such as the human X chromosome, it may take several generations for genotype frequencies at sex-linked loci to reach equilibrium levels.

3.GENOMIC WIDE ASSOCIATION STUDIES

Through genome-wide association studies (GWAS), scientists can identify genes associated with specific diseases or traits. By analyzing the entire set of DNA, or genome, of a large population, this method looks for small variations called single nucleotide polymorphisms (SNPs). In each study, thousands or even hundreds of SNPs can be analyzed simultaneously. Researchers then identify which SNPs are more prevalent in individuals with a particular disease compared to those without it. These SNPs are believed to be linked to the disease and can help researchers identify genes likely involved in the disease's development.

Genome-wide association studies are a potential approach to studying common, complicated illnesses where a person's risk is influenced by several genetic variants since they look at SNPs throughout the genome.By employing this approach, scientists have identified SNPs associated with several complex diseases, including diabetes, heart disease, Parkinson's disease, and Crohn's disease. SNPs have also been linked to an individual's response to certain medications and susceptibility to particular environmental factors, such as pollutants. Future genome-wide association studies are expected to uncover more SNPs linked to medication side effects and chronic illnesses.

Genome-wide association studies identify specific SNPs that individually contribute a small fraction to disease risk. However, when numerous SNPs distributed across the genome are combined, they can help estimate the overall risk of developing a disease or reacting to specific medications. Precision medicine has advanced to a point where researchers can more precisely anticipate which preventive and treatment approaches will be effective for specific populations by utilizing data from genome-wide association studies.

3.1.Methods of Study Design and Analysis:

Careful planning and exacting statistical analysis are necessary for the conception and implementation of GWAS:

- Sample Collection: To collect data for GWAS, large and diverse cohorts with precise phenotypic information about both affected and unaffected individuals (cases and controls) are needed.

- Genotyping: With high-throughput genotyping systems, single nucleotide polymorphisms (SNPs) ranging from hundreds of thousands to millions are examined across the genome.
- Quality control: Tight quality control procedures are used to guarantee the accuracy of the data. These procedures include removing low-quality SNPs and samples, determining population stratification, and correcting for relatedness among research participants.
- Statistical Analysis: To assess the relationship between each SNP and the desired characteristic, statistical tests like logistic regression or chi-square tests are performed, taking into consideration any confounding variables and correcting for multiple testing.

3.2.Applications in Statistical Genetics Research:

Understanding the genetic basis of numerous traits has advanced significantly thanks to GWAS:

- Disease Mapping: Hundreds of genetic variants have been connected to a range of diseases, including diabetes, cancer, cardiovascular disorders, and neurological issues, due to the application of GWAS.
- Phenotypic traits: The genetic foundation of quantitative traits such as height, blood pressure, cholesterol, and body mass index (BMI) has been shown by GWAS.
- Drug Response: Personalized medicine and treatment techniques have been made possible by the identification of genetic markers linked to medication responsiveness and adverse effects through the use of GWAS.
- Biological Insights: The results of the GWAS have yielded important information on the biological processes and mechanisms that underlie many illnesses and characteristics. This information has made it possible to identify novel targets for therapeutic treatments and drug development.

3.3.Challenges and Future Directions:

Even though GWAS has contributed significantly to the advancement of genetic research, there are still a number of issues and areas for development.

- Missing Heritability: Only a tiny portion of the heritability of complex characteristics is accounted for by several documented genetic variations, indicating the possibility of additional genetic components that have not yet been found.
- Rare variants: While the majority of GWAS research focuses on common genetic variations, more thorough sequencing technology should be used to investigate the effect of uncommon variants.
- Population diversity: To capture genetic variation across various ethnic groups and communities, it is imperative to ensure variety in research populations.
- Integration of Multi-Omics Data: A more thorough knowledge of the links between genotype and phenotype may be obtained by including data from other omics layers, such as transcriptomics, epigenomics, and metabolomics.

3.4.Recent Developments and Emerging Technologies:

The capabilities of GWAS have been extended, and recent developments have improved its usefulness in genetic research:

- Multi-Ethnic Studies: GWAS conducted in a variety of ethnicities guarantees the generalizability of results across many ethnic groups and aids in the discovery of genetic correlations specific to a certain community.
- Functional Genomics: By including functional genomic data—such as epigenetic changes and gene expression profiles—GWAS results may be prioritized and validated, offering mechanistic insights into the relationships between genes.
- Polygenic Risk Scores (PRS): By creating PRS from many genetic variations, it is possible to forecast a person's risk for a certain disease and enhance risk assessment for complicated illnesses.
- Rare Variant Analysis: The investigation of rare variants' involvement in complex features and disorders is made possible by the development of specific techniques for the identification and examination of these variations using whole-genome sequencing data.

With its significant insights into human biology and the genesis of illnesses, GWAS has completely changed our understanding of the genetic basis of complex characteristics and disorders. GWAS will continue to propel developments in statistical genetics research through sustained improvements in study design, analytic techniques, and the integration of multi-omics data. These developments might lead to the creation of precision therapies, tailored therapy, and better global healthcare outcomes for people.

4.LINKAGE ANALYSIS

A method used in genetics to map the locations of genes on chromosomes is called linkage analysis. It is particularly useful in identifying the genetic variants resulting in Mendelian traits, which are features that are strongly impacted by a single gene.GWAS sometimes require very large sample sizes in order to identify meaningful linkages throughout the genome that can be replicated. Software like CaTS14 or GPC15 may be used to calculate power and determine the optimal sample size. Research designs may contain cases and controls when the trait is dichotomous, or quantitative assessments on the whole study population when the characteristic of interest is quantitative. Additionally, there are population-based and family-based designs to choose from.The choice of data resource and research design for a GWAS is influenced by the intended sample size, the experimental question, and the ease of collecting new data or the accessibility of previously collected data.

4.1.Fields that use Linkage Analysis

1. Family Studies: Linkage analysis research often involves examining families in which many members are affected. By examining inheritance patterns within these families, researchers may identify regions of the genome most likely to have genes associated with the desired characteristic.
2. Genotyping: Family members are genotyped for single nucleotide polymorphisms (SNPs) and microsatellites, two types of genetic markers. These markers are distributed throughout the whole genome and are used to track inheritance patterns.
3. Statistical Analysis:Subsequently, statistical techniques are employed to examine the co-segregation between the genetic markers and the desired characteristic among families. Through a comparison of the inheritance of a characteristic, researchers can determine the genetic regions linked to that trait indicators indicating the presence or lack of the characteristic.
4. Linkage Maps: The end result of linkage analysis is a linkage map that shows the relative placements of genetic markers along the chromosomes and reveals regions most likely to house the gene or genes producing the trait.
5. Fine Mapping and Gene Identification: Further investigation is often conducted to determine the precise gene or genes responsible for the trait and to refine the region of interest if a linkage signal is detected. These techniques might involve functional studies, fine mapping with more markers, or sequencing.

5.CHI-SQUARE TEST FOR GENETIC ASSOCIATION

To determine if two category variables have a significant correlation with one another, a statistical method known as the chi-square test is employed.It is used to determine if there is a significant relationship between genetic variations (such single nucleotide polymorphisms, or SNPs) and phenotypic features (like illness status or other observable qualities). In genetic epidemiology, this test is frequently used to find genetic variables that can influence the onset of certain diseases or other characteristics.

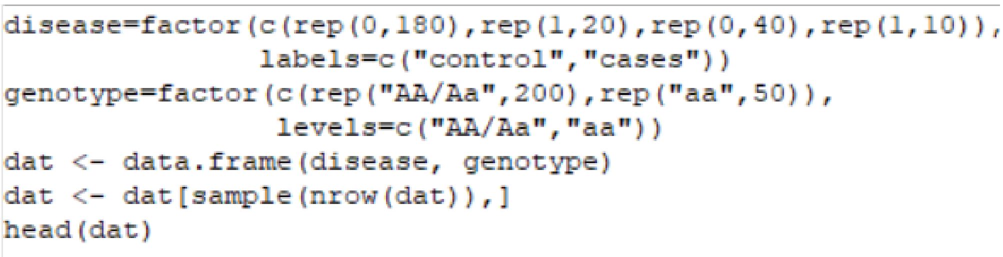
Based on the assumption that there is no correlation between the genetic variant and the trait, the chi-square test compares the observed frequencies of genotypes or alleles in cases—individuals who possess the trait of interest—and controls—individuals who do not possess the trait—to expected frequencies.

5.1.Establishing the Contingency Table:

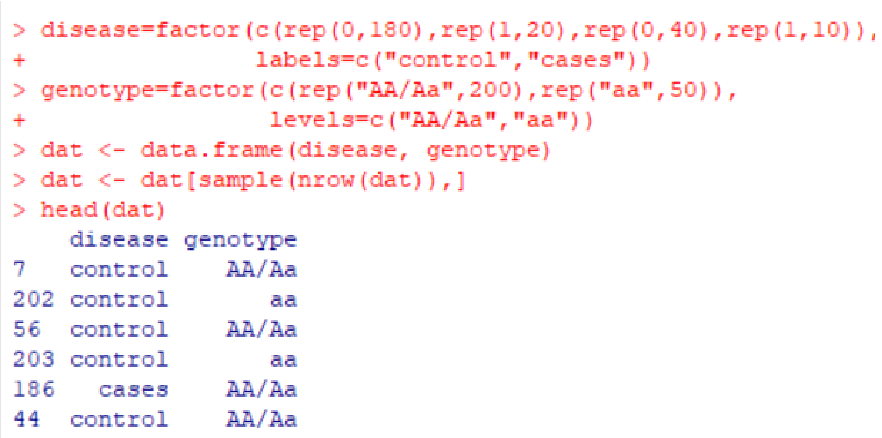
We usually put up a 2x2 contingency table that cross-tabulates the genotype frequencies between cases and controls in order to perform a chi-square test for genetic connection.

Assume that 250 persons total, of which some have a certain illness and the others do not. We discover that 10% of people who are homozygous for the minor allele (aa) do not have the disorder, whereas 20% of homozygous individuals do. Would we notice this once again if we selected 250 individuals?

Let's put these percentages in a dataset that we create:

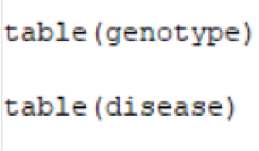


```
disease=factor(c(rep(0,180),rep(1,20),rep(0,40),rep(1,10)),
                   labels=c("control","cases"))
genotype=factor(c(rep("AA/Aa",200),rep("aa",50)),
                    levels=c("AA/Aa","aa"))
dat <- data.frame(disease, genotype)
dat <- dat[sample(nrow(dat)),]
head(dat)
```



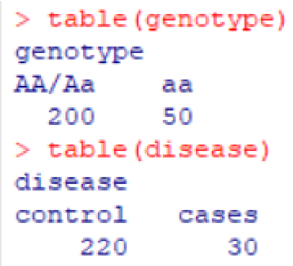
```
> disease=factor(c(rep(0,180),rep(1,20),rep(0,40),rep(1,10)),
+                    labels=c("control","cases"))
> genotype=factor(c(rep("AA/Aa",200),rep("aa",50)),
+                     levels=c("AA/Aa","aa"))
> dat <- data.frame(disease, genotype)
> dat <- dat[sample(nrow(dat)),]
> head(dat)
    disease genotype
7   control    AA/Aa
202 control       aa
56  control    AA/Aa
203 control       aa
186   cases    AA/Aa
44  control    AA/Aa
```

We will utilize the function table to generate the proper two-by-two table. The frequency of each level inside a factor is tabulated using this function. As an instance:



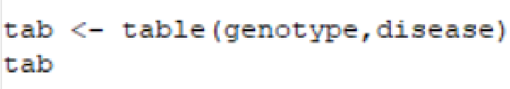
```
table(genotype)

table(disease)
```

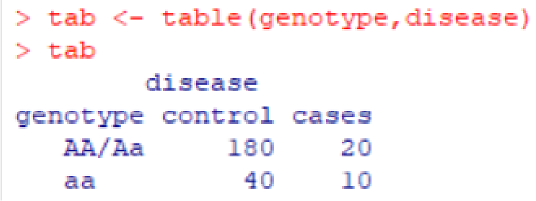


```
> table(genotype)
genotype
AA/Aa     aa
  200     50
> table(disease)
disease
control   cases
    220      30
```

If you give the function two factors, it will calculate every potential pair and generate the two by two table:



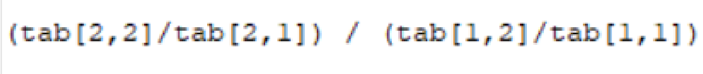
```
tab <- table(genotype,disease)
tab
```



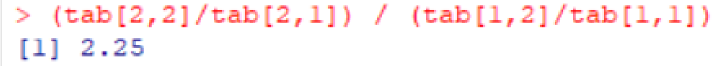
```
> tab <- table(genotype,disease)
> tab
        disease
genotype control cases
   AA/Aa     180    20
   aa         40    10
```

Keep in mind that supplying table n factors will result in all n-tables being tabulated.

In statistics, we frequently use the odds ratio (OR) to summarize these results. By calculating the probability of acquiring the disease if you are a "aa" (10/40) and the probability of acquiring the disease if you are a "AA/Aa" (20/180), we can get the ratio (10/40)/(20/180).

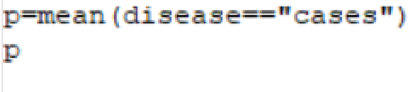


```
(tab[2,2]/tab[2,1]) / (tab[1,2]/tab[1,1])
```

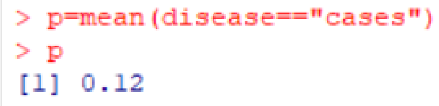


```
> (tab[2,2]/tab[2,1]) / (tab[1,2]/tab[1,1])
[1] 2.25
```

The OR is not used directly in the computation of a p-value. Instead, we make the assumption that there is no correlation between genotype and illness, after which we calculate our expectations for each table cell (note that the term "cell" in this context refers to pieces of a matrix or table, not real cells). The groups comprising 200 individuals and 50 persons were allocated the illness at random with equal probability under the null hypothesis. In such a scenario, the likelihood of illness is:

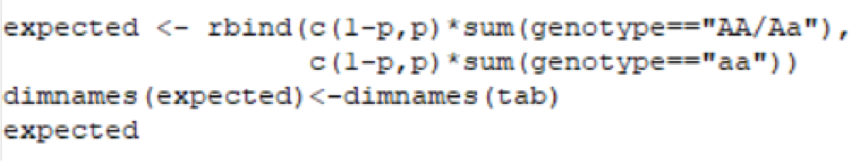


```
p=mean(disease=="cases")
p
```

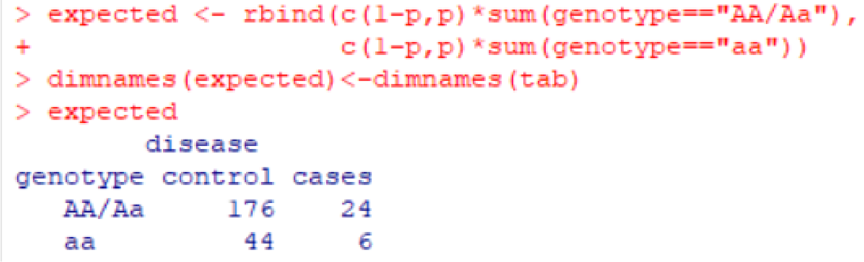


```
> p=mean(disease=="cases")
> p
[1] 0.12
```

The expected table is therefore:

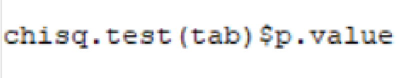


```
expected <- rbind(c(1-p,p)*sum(genotype=="AA/Aa"),
                  c(1-p,p)*sum(genotype=="aa"))
dimnames(expected)<-dimnames(tab)
expected
```

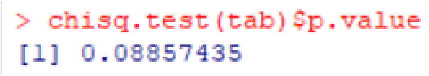


```
> expected <- rbind(c(1-p,p)*sum(genotype=="AA/Aa"),
+                   c(1-p,p)*sum(genotype=="aa"))
> dimnames(expected)<-dimnames(tab)
> expected
        disease
genotype control cases
   AA/Aa     176    24
   aa         44     6
```

An asymptotic result about the sums of the distinct binary outcomes is used in the Chi-square test. With this estimate, we can determine the probability of observing a deviation from the expected data as significant as that which we observed. The p-value for this table is as follows:



```
chisq.test(tab)$p.value
```



```
> chisq.test(tab)$p.value
[1] 0.08857435
```

6.MACHINE LEARNING IN GENETICS

The application of machine learning (ML) techniques in genetics and genomics has grown in popularity because of its capacity to forecast outcomes, recognize intricate patterns in large-scale genetic data, and analyze genetic data.

The creation and use of computer algorithms that get better with use falls under the umbrella of machine learning.

Machine learning approaches fall into three categories: unsupervised, semi-supervised, and supervised. Supervised methods use labels ('gene' or 'not gene') on labeled instances to learn and predict labels on new examples. Unsupervised methods, on the other hand, search data sets for patterns without categorizing them. Semi-supervised algorithms, which leverage patterns in unlabeled data to boost the strength of label prediction, combine these two approaches.

A number of machine learning approaches may be required, depending on whether an application is primarily interested in predicting capabilities or in comprehending the output model. Generally speaking, generative models—which assume a probabilistic distribution across input data—are better for interpretability, whereas discriminative models—which only seek to model labels—are most effective for predictive capability.

When a model is given more data, it may be trained more effectively by including prior knowledge. Additionally, it may be used to add data that the model does not directly utilize or simplify the model. Explicit or implicit prior information can be included into a probabilistic model by choosing features or similarity measures.

The choice of performance metric is heavily influenced by the application job. Machine learning approaches function best when they maximize a relevant performance metric.

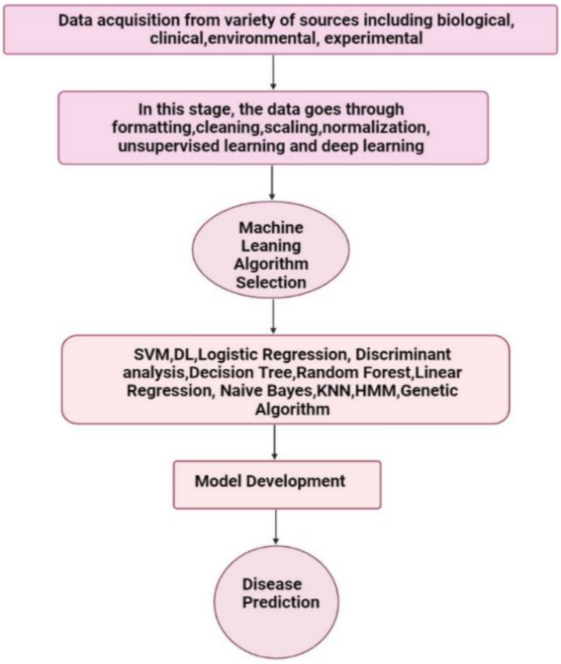
When there are complicated dependencies among samples in the data, network estimation techniques are acceptable. When indirect relationships are taken into consideration, these approaches perform optimally.

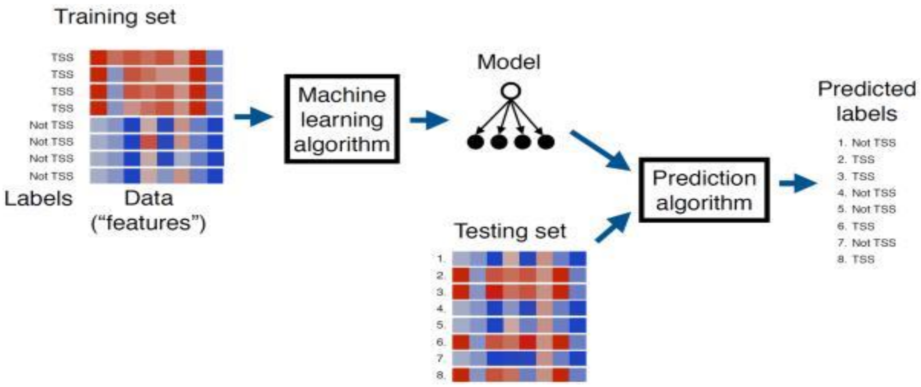
While the field of genomics is still in its infancy when it comes to the application of AI/ML techniques, researchers have already profited from creating customized programs.

6.1.Applications of Machine Learning:

- Applying facial analysis to individuals' faces allows AI systems to accurately identify genetic diseases.
- taking a liquid sample and using machine learning techniques to identify the primary cancer kind.

- determining how long a patient's particular cancer will last.
- To differentiate between genetic variants that cause disease and those that are benign, apply machine learning.
- deep learning to improve the performance of gene editing tools such as CRISPR.
- Genomic Variant Calling and Annotation: Single nucleotide variations (SNVs), insertions, deletions, and structural variants can all be reliably identified from sequencing data using methods from machine learning (ML) .Predicting the functional effects of genetic variants, such as how they may affect the structure and function of proteins, is known as variant annotation. Variants can be categorized by ML models according to their relation to particular diseases, regulatory effects, or pathogenicity.
- Genome-Wide Association Studies (GWAS): Information from GWAS studies are analyzed using machine learning techniques to find correlations between genetic variations and characteristics or illnesses.The generation of polygenic risk scores and the identification of new risk loci are made possible by the ability of machine learning algorithms to handle high-dimensional data and identify intricate interactions between genetic variations.
- Predictive Modeling for Disease Risk and Diagnosis: Using information from clinical trials, environmental variables, genetic profiles, and other sources, machine learning models are built to forecast a person's likelihood of contracting specific diseases.Predictive models can help identify people who are at high risk for cardiovascular disease or cancer, for example, and can also help with disease diagnosis and prognosis.
- Pharmacogenomics: Personalized medicine techniques are made possible by the application of machine learning algorithms to predict drug reactions based on genetic variation.Pharmacogenomic models have the ability to detect genetic markers linked to medication efficacy, side effects, and the best dosing schedules.
- Functional Genomics: Large-scale functional genomics data, including gene expression profiles, chromatin accessibility, and epigenetic alterations, are analyzed using machine learning approaches.Machine learning models have the ability to deduce gene regulatory networks, detect cis-regulatory regions, and rank potential genes for additional experimental verification.
- Drug Design and Protein Structure Prediction: ML algorithms are utilized in drug design to forecast drug binding affinities, protein-ligand interactions, and protein structures.By screening extensive compound libraries, creating unique drug candidates, and improving lead compounds, machine learning algorithms help hasten the process of finding new drugs.
- Population genetics and evolutionary genomics: Machine learning algorithms are used to examine genetic data related to populations, deduce past demographic patterns, and find evidence of natural selection.Machine learning algorithms are able to detect genetic adaptability to various environments, monitor population movements, and describe population composition.
- Functional Annotation of Non-Coding Regions: Long non-coding RNAs, promoters, enhancers, and other non-coding genetic variations are all predicted to have functional significance using machine learning algorithms.For functional validation, non-coding variations can be prioritized, and ML models can clarify their significance in gene regulation and illness.


Data acquisition from variety of sources including biological, clinical,environmental, experimental
In this stage, the data goes through formatting,cleaning,scaling,normalization, unsupervised learning and deep learning
Machine Leaning Algorithm Selection
SVM,DL,Logistic Regression, Discriminant analysis,Decision Tree,Random Forest,Linear Regression, Naive Bayes,KNN,HMM,Genetic Algorithm
Model Development
Disease Prediction



7.MENDELIAN RANDOMIZATION

Mendelian randomization, or MR for short, is a technique that looks at the direct relationship between an exposure and an outcome by measuring variations in genes. It evaluates the relationship of causation between a clinical result and a risk factor.Gray and Whitley initially proposed Mendelian randomization in 1986.

Mendelian randomization is one analytical method to determine the causal linkages of reported links between a potentially risk factor or modifiable exposure and a clinically relevant result. Randomized controlled trials (RCTs) are a technique used to establish causal relationships and lessen planning phase bias by generalizing observed findings from a sample selected through sampling.The main goal of statistics is to identify the characteristics of the populations of interest using the data gathered from the sample. The concept of randomization, which is comparable to random assignment, lends credence to this interpretation .In observational epidemiological research, however, this is challenging to accomplish.

An RCT is the "gold standard" for clinical research when it comes to experimentally evaluating scientific concepts. In this approach, people within a population are used as experimental units, and various treatments are assigned to them at random. In its most basic form, a "control group" is contrasted with one "active group" (such as an intervention on a risk factor).

Mendelian randomization is the process of drawing conclusions about causal effects from observational study data by treating instrumental variables (IVs) to be genetic variants. The reason Mendelian randomization is called "Mendelian deconfounding" since its goal is to produce a causality estimate free from confounding-related biases.

Recent studies have examined the causal association between lipid metabolism and insulin resistance, as well as between coronary heart disease and cardiovascular disease, using Mendelian randomization with pleiotropy and absence of genotype and ethnicity information. Mendelian randomization is an effective strategy for managing confounding and reverse causality, which are common problems in epidemiological research. To put it another way, it's a technique for estimating or testing the causal effect from confounded observation data.

It is challenging to determine whether a person's genetic variation inherited from their parents is entirely random, but it is reasonable to assume that a population's genetic variation is dispersed randomly in exposure and may act as an increased risk factor for a specific outcome (a disease of interest). To overcome these constraints, an attempt was made to apply the Mendelian random assignment approach.

The table below provides several examples of Mendelian randomization, broken down by type of exposure and outcome.

7.1.Evidence from epidemiology supporting causal links as determined by Mendelian randomization:

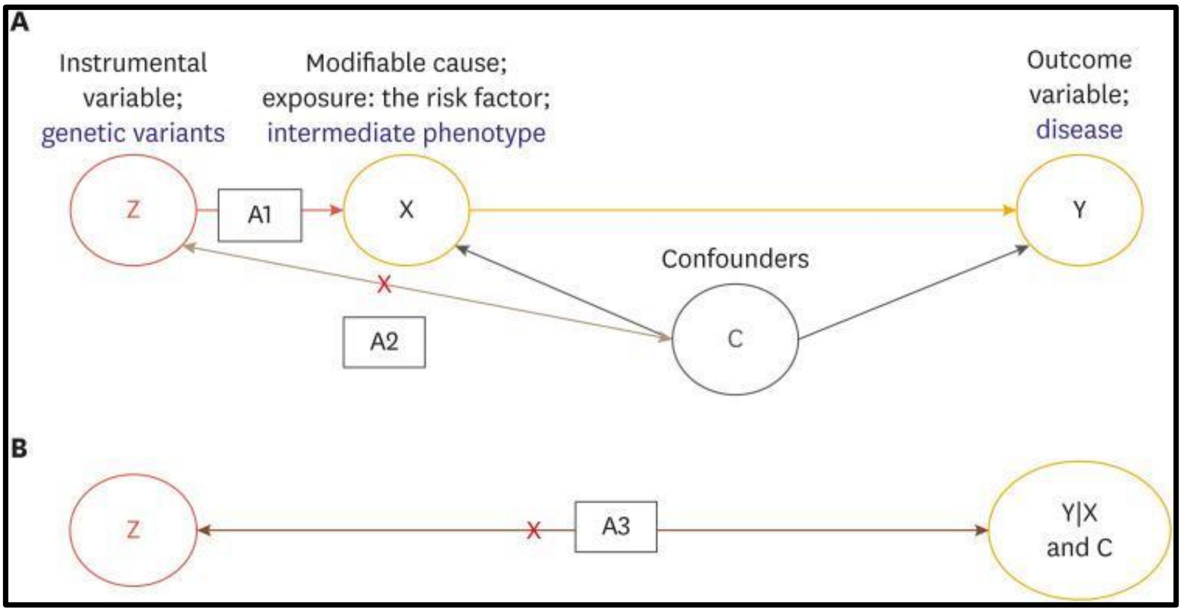
Exposure type	Exposure	Outcome
Dietary factors	Milk intake	Issue with metabolic syndrome
	Caffeine intake	The probability of stillbirth
	Alcohol intake	Elevated blood pressure
Physical characteristics	Body mass index	Early menarche
	Body fat mass	Academic achievement
Inter-generational effect	Intrauterine folate	Neural tube defects
Biomarkers	CRP	Increased body fat
		CIMT
		Cancer
		Elevated triglycerides, HDL cholesterol, and insulin resistance
	Apolipoprotein E	Cancer

	Homocysteine	A higher quantity of homocysteine causes stroke.
Pathological conditions	Mental health issues include obesity, depression, and ADHD.	Lower educational achievement

ADHD (Attention deficit hyperactivity disorder); CRP (C-reactive protein); CIMT (carotid intima-media thickness, cholesterol); HDL (high-density lipoprotein).

7.2.Explanation of variables:

Z is IV(Instrument variable), X is the variable that denotes the cause, and Y is the result, or illness. A variable Z that only affects X and has no effect on Y can be included to the model in order to verify that X has an impact on Y. So, X is a causative variable that has a significant impact on Y. This time, the variable Z is referred to as the IV. The theoretical concept of the Mendelian randomization with its three fundamental assumptions (A1, A2, and A3) is mirrored in the below figure.


A
Instrumental variable; genetic variants
Modifiable cause; exposure: the risk factor; intermediate phenotype
Outcome variable; disease
Z
A1
X
Y
Confounders
A2
C
B
Z
A3
Y|X and C

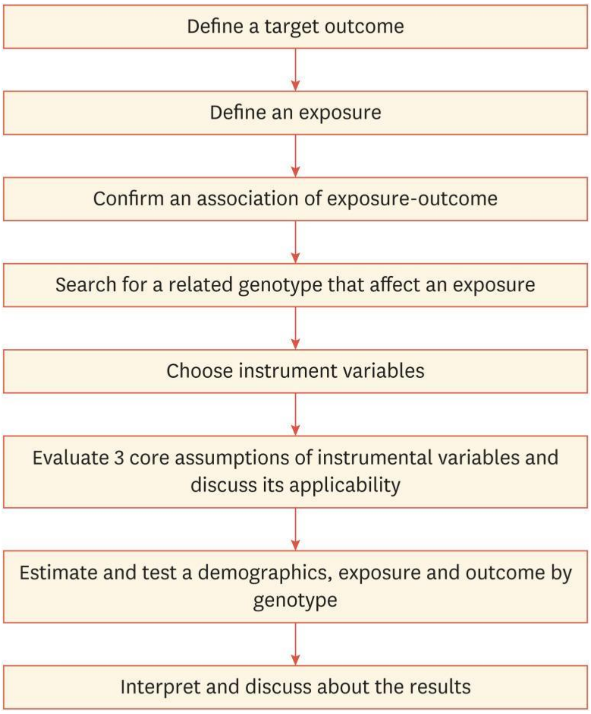
7.3.Statistical analysis in GWAS:

In GWAS, testing and estimating are the objectives of statistical analysis.Using tables, graphs, and statistics for prospective risk factors and demographics based on the genotypes of SNPs used as IVs, distribution and statistical tests are then carried out. When a significant number of SNP loci are examined for the contingency table, or when there is a genetic frequency difference of SNP loci between patients and normal, the odds ratio (OR) is calculated. If the environmental factors of the comparison target group differ, regression models are employed to explain features of the environment, such as age and gender, that are not related to the SNP locus. A machine learning model and the analysis of survival time are two further statistical analyses that are done, along with analysis of variance to demonstrate a correlation between possible risk variables and IVs.

7.4.R functions for Mendelian Randomization:

A R software package that utilizes summarized data for Mendelian randomization in R program version 3.0.1 or later was described by Yavorska and Staley. Many methods for applying Mendelian randomization utilizing condensed data from a consortium—which included a summary of the genetic relationships between exposure and outcome—were given in this report. This is one way that IV may be used to establish causality.

More trustworthy evidence is frequently obtained from a well-designed Mendelian randomization trial that meets the assumptions than from a traditional observational epidemiology study. However, while interpreting the data and contrasting them with prior research from other studies, care must be taken.


Define a target outcome
Define an exposure
Confirm an association of exposure-outcome
Search for a related genotype that affect an exposure
Choose instrument variables
Evaluate 3 core assumptions of instrumental variables and discuss its applicability
Estimate and test a demographics, exposure and outcome by genotype
Interpret and discuss about the results

8.RARE VARIANT ANALYSIS

Rare variants are single nucleotide polymorphisms (SNPs) having a minor allele frequency (MAF) of less than 0.01.They often have larger phenotypic consequences when compared to low-frequency (less common) (0.01< MAF <0.05) or common (MAF >0.05) disease-associated SNPs. Many investigations have looked at the underlying genetic etiology of complex illnesses and quantitative characteristics using genome wide association studies (GWASs).The experiment's design plan has a major impact on the outcome of rare variant research. The preprocessing and data selection establishes a feasible beginning point for the identification of important rare variants. Low-depth whole genome sequencing (WGS) is the suggested technique for investigating rare variants because extended WGS is sometimes expensive.Furthermore, low-depth sequencing can effectively identify illnesses and variations provided the sample is large enough. Although it only targets the coding region of the genome, exome sequencing is another technique that has been used to diagnose a number of Mendelian disorders, according to research.

The following gene-based tests are often used in rare variant association studies:
combination tests, variance component tests, adaptive burden tests, and burden tests .

By compressing the uncommon variant count, burden tests generate genetic scores.They aggregate the data from several genotypes of a single sample into a burden score that is easily used for association, and they are predicated on the fundamental idea that all rare changes that are causally and trait-associated have the same degree of impact. Cohort allelic sums test (CAST), available as a R package, is one such burden test. χ2 and Fisher exact tests can be used for burden testing, and they are similar to single-SNP analysis depending on the dataset under investigation.However, burden tests contain one flaw, they make the assumption that every variation affects the phenotypic in the same way.

When applied with set criteria, Adaptive burden tests are more reliable than burden testing. They permit the existence of null, trait-increasing, and trait-decreasing variations and do away with the restrictions of a burden test.Adaptive burden tests are computationally intensive since they depend on permutation to determine P values. They also employ the regression coefficients as weights for every choice.
Tests using variance components allow for a variety of effects, including both protective and risk variations, with varying impact sizes across unusual variation groups.

This approach is used in the sequence kernel association test (SKAT). It is applicable to binary and quantitative features alike. The foundation of SKAT-O is a combination test, sometimes referred to as a burden test and variance component test. It makes score statistics more flexible, which results in the best possible combination of effective calculations. The SKAT statistic may be optimized over a number of potentially significant SNV annotations using cSKAT. It works well with bigger sample sizes (N≥5,000) and SNV weights that are provided appropriately.

A new study suggests the Bayes Factor technique for the rare variant association test in sequencing data. Its sensitiveness to changes in the allelic distribution, variances in the uncommon variant count in binary studies, or both may be seen from its informative priors. While it hasn't been tested for different demographic patterns, imbalanced case-control study approaches may employ it. Another recently suggested technique that is more effective than the variance-component and burden tests for continuous phenotypes is adaptive hierarchically structured variable selection (HSVS-A). It may be used for analysis that is region-based or set-based. In addition to producing individual effect estimates for uncommon mutations, it can automatically regulate type I error rates.The association test for rare variants that utilizes algebraic statistics (ASRV) is a novel method for evaluating association when the causal variations have opposing effects. Single variant association tests resistant to genetic confounding, such Transmission Disequilibrium Tests (TDTs) or Family-based Association Tests (FBATs), can be employed in family-based association research. The Aggregated Cauchy Association Test (ACAT) is a set-based association test for sequencing research. The association test involving a trait and a variant-group is computationally efficient since it just requires P values. The R package RVFam provides methods to evaluate the association between rare variants and survival, continuous characteristics, or binary traits as evaluated in family datasets. It performs better than the Firth test and generalized linear model (GLM), which do not take sample-relatedness into account.A hybrid approach that use the GLM for gene-based tests and the Firth test with family data generally for rare variant association studies of binary outcomes proves to be highly efficient in the absence of variant filtering. The Integrated Nested Laplace Approximation (BATI) test, which is part of the rare-variant Genome Wide Association Study (rvGWAS) framework, incorporates both numerical and categorical variant features as variables for the Bayesian rare variant association test. Strong analysis in the case of loss-of-function variations is demonstrated by this test.

A pathway-based technique or multi-set testing for rare variant association test shows an improvement in statistical power and may also enhance putative disease-etiology elucidation when subsets of genes, such as exons, introns, or gene windows, comprise fewer variants overall. For causal variations with large effect sizes, a single-marker association test known as Copula-based Joint Analysis of Multiple Phenotypes (C-JAMP) works well. It uses a combined model of many phenotypes and variants or other factors and is implemented as a R program. Quantitative Phenotype Scan Statistic (QPSS) offers the advantage of localizing genomic areas with unusual quantitative-phenotype-associated variant groupings by using variant annotation to narrow a known region of interest.

Genotype imputation and population stratification are problems for investigations including uncommon variants. In these kinds of investigations, the former frequently serves as a confounding factor, therefore it should be corrected before continuing. For this reason, linear-mixed models and principal component analysis are frequently employed; however, their efficacy for uncommon variations is uncertain. Contrarily, genotype imputation has a detrimental impact on studies of rare variants because it reduces imputation accuracy with decreasing MAFs, which causes rare variants to be eliminated while in the quality-check phase. The implementation of a hybrid reference panel might potentially address this issue by improving the imputed accuracy of rare mutations.

For a genuine positive correlation between rare variants and a disease, the link must be replicated in a significant amount of samples, which often entails genotyping or sequencing.Once the first study has demonstrated its value, it may be beneficial to do follow-up research on high-priority changes across several samples. Once a definitive relationship between the uncommon variations and illness has been established, more research may be done to connect the variants to molecular and cellular activities.

9.GENETIC DRIFT AND EFFECTIVE POPULATION SIZE

9.1.Understanding the mathematics of drift:

Concerns regarding the effects of environmental change on the ecosphere are shared by environmentalists, conservationists, biologists, and knowledgeable individuals. Even though organisms are not able to plan for changes in their environment, populations with greater diversity are better equipped to deal with change when it does come. It should be noted that a population's degree of genetic variety is dynamic, reflecting the constantly shifting balance between random and nonrandom mechanisms that eliminate variation. Occasionally, the latter may become more dominant than the former, resulting in low levels of variation that are not recoverable across ecological time periods. Scientists are aware that natural selection has the power to eliminate variety from a population and that variation develops as a result of mutation and recombination.

Furthermore, scientists are fully aware that, contrary to what the Hardy-Weinberg model would have us believe, real-world populations are not limitless. When combined, these elements cause a constant loss of variation—a phenomenon known as genetic drift.

Our concerns about species with limited populations stem from genetic drift. Drift is more obvious in smaller populations because they have less diversity and, thus, a lower ability to adapt, or respond favorably, to varying circumstances.

For a gene with two alleles, *A* and *a*, imagine a population that is at Hardy-Weinberg equilibrium. Let $p = q = 0.5$, q = the relative frequency of the *a* allele, and p = the relative frequency of the *A* allele. The allele frequencies in succeeding generations need to stay at 0.5 in order for drift to not happen. The number of *A* alleles (designated k) is equal to $2pN$ if N is the size of the diploid organism population. How can we determine the precise likelihood that, following a generation of random sampling, k will stay equal to $2pN$ given this information? In order to accomplish this, we start by using the generic binomial distribution formula:

$$\Pr(k \mid p, n) = [n! / k!\ (n - k)!]p^{k}(1-p)^{n-k}$$

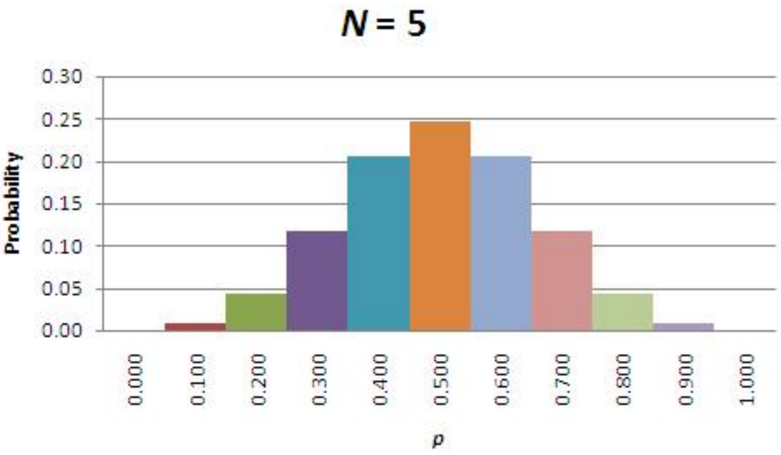
When there are two potential results in a trial, the likelihood of each outcome is constant throughout every trial, and each of the trials are independent of one another, the binomial distribution is utilized. In this case,we have $n = 2N$ trials (i.e., gametes sampled), and the *A* allele is obtained in k of these trials. The two possible results are *A* and *a*, which correspond to the allele sampled in a gamete. The probability of obtaining a gamete with the *A* allele is p.

[n! / (k! (n -k)!)] indicates the number of possible observations of exactly *k* "successes" (defined as *A* alleles in this case) and *n -k* "failures" (defined as *a* alleles in this case). The term *pk (1-p)n-k* is the actual likelihood of examining any particular sequence of *k* "successes" and *n - k* "failures" . The precise chance of seeing *k* "successes" and n - k "failures" is thus provided by the product of these variables, assuming that the order of the observations is unimportant.

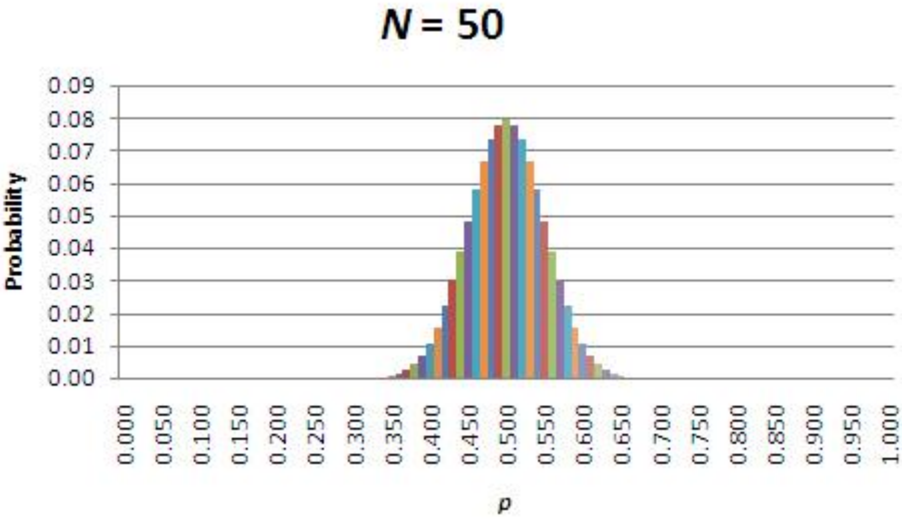
We can get the exact likelihood that *k = 2pN* for a range of *N*, using the binomial distribution formula. Following this leads to the following outcomes:

Population size (N)	2	5	10	50	100	500	1,000	10,000
Gene Copies (2N)	4	10	20	100	200	1,000	2,000	20,000
Pr(k | p, n = 2N)	.375	.246	.176	.080	.056	.025	.018	.006

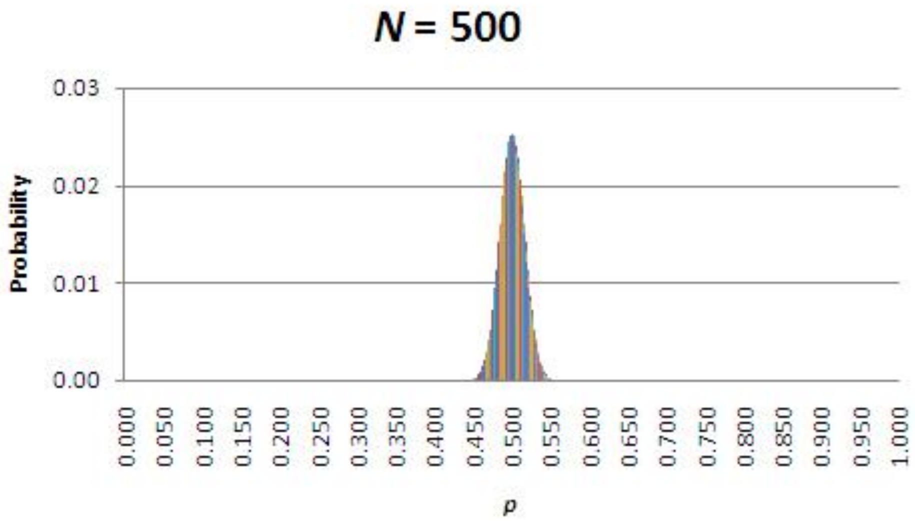
These results might appear reversed at first look. The data indicates that the smaller populations have a better likelihood of having unaltered allele frequencies. The likelihood of an alteration in allele frequencies is higher under these scenarios, and the significance of the alteration is what counts. Let's examine the population sizes where *N* = 5, *N* = 50, and *N* = 500 to understand what this implies. The probability of allele frequencies in each of these populations' subsequent generations are displayed in the below figures.



Allele frequency probabilities in the following generation within a population of five organisms (N=5).



Allele frequency probabilities in the following generation within a population of five organisms (N=50).



Allele frequency probabilities in the following generation within a population of five organisms (N=500).

It should be clear that as population numbers rise, the distribution's width gets smaller. This is because the sampling error has decreased.Hence, the amount of variance caused by sampling error will drop as the population develops, even though allele frequencies will almost certainly fluctuate with each generation.Perhaps most importantly, the change's directionality is unexpected; over time, allele frequencies will grow and decline in an unforeseen manner. In addition, sampling for the following generation will focus on the new *p* value when change does occur.Then with enough time, p will drift to either 0.0 or 1.0 in the absence of forces (such balancing selection) that retain both alleles; that is, one allele will eventually drift to fixation and the other towards extinction. The population size and the beginning frequencies of the alleles will determine how long it takes for this to happen.

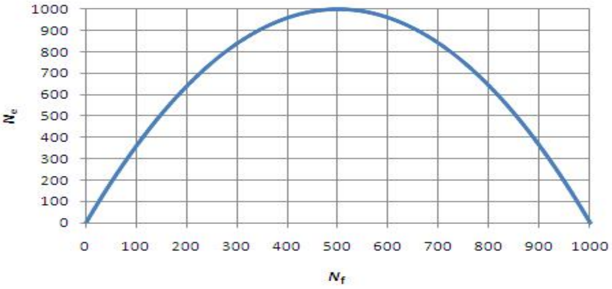
9.2.Effective population size:

It is reasonable to question whether genetic drift is actually all that relevant given the size of most populations. While most populations are enormous, that doesn't mean they behave as such. As a result, the rate of genetic drift and census population size (N_c) are not exactly proportionate. Instead, it is in line with a more abstract metric, which is the magnitude of the effective population (Ne). In a perfect population of sexually reproducing individuals, Ne equals Nc. The "ideal" population is characterized by the following characteristics, from which the bulk of deviations lower the effective population size:

The number of men and females is equal, and they can all reproduce.

- Every individual has the same probability of becoming a parent, and there is no variation in the number of children everyone produces beyond what chance would indicate.
- Mating occurs at random.
- The amount of individuals fit for breeding doesn't change from generation to generation.

Anything that increases the variation in reproductive success between individuals (beyond sampling variance) would effectively reduce *Ne* (the size of an ideal population that receives genetic drift at the rate of the population in question). Consider the impact of different numbers of men and females mating, for instance. Every man and every female in the population would have an equal chance of mating.Each sex has an equal chance of reproducing, but when one sex predominates over the other, a person's odds of reproducing are now determined by their sex. In this case, the formula $N_e = 4N_mN_f/(N_m + N_f)$, where the number of men is denoted by N_m and the number of females is denoted by N_f , may be used to determine the effective population size. The connection between N_e and N_f in an entire population of 1,000 mating individuals, is seen in the figure below. Every member of the population has an equal chance of passing on their genes in an ideal population. But this is rarely the case in real life, and Ne is especially sensitive to imbalances in the proportions of men and women in the population.



The relationship between Ne and Nf in a population of 1000 mating individuals.

REFERENCES

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4963036/
2. https://www.sciencedirect.com/topics/neuroscience/hardy-weinberg-principle
3. https://www.nature.com/articles/s43586-021-00056-9
4. https://genomicsclass.github.io/book/pages/association_tests.html
5. https://www.nature.com/articles/nrg3920
6. https://www.genome.gov/about-genomics/educational-resources/fact-sheets/artificial-intelligence-machine-learning-and-genomics#:~:text=Some%20examples%20include%3A,will%20progress%20in%20a%20patient.
7. https://www.frontiersin.org/articles/10.3389/fsysb.2022.877717/full
8. https://en.wikipedia.org/wiki/Mendelian_randomization
9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7379124/
10. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8267277/
11. https://www.nature.com/scitable/topicpage/genetic-drift-and-effective-population-size-772523/
12. https://www.nature.com/
13. https://atm.amegroups.org/
14. https://synapse.koreamed.org/
15. https://www.ncbi.nlm.nih.gov/
16. https://www.scribd.com/
17. https://www.ndsu.edu/pubweb/~mcclean/plsc431/mendel/mendel4.htm
18. https://en.wikipedia.org/wiki/Main_Page