Innovation in Science Pursuit for Inspired Research

Scholarship for Higher Education

Summer Research Project Report

Statistical analysis of some genotype-trait associations using R

Student

Tanikella Padma Ragaleena

3rd year Integrated M.Sc.

School of Mathematical Sciences

NISER Bhubaneswar (HBNI)

tp.ragaleena@niser.ac.in

Mentor

Prof. Madhuchhanda Bhattacharjee

Professor

School of Mathematics and Statistics

University of Hyderabad

mbsm@uohyd.ernet.in

July 13, 2019

Acknowledgements

I would like to express my sincere gratitude to my supervisor *Prof. Madhuchhanda Bhattacharjee* for her patience, motivation, and immense knowledge.

I would also like to thank INSPIRE for fellowship and the Mentorship Grant.

Contents

Introduction

The report is based on the book "Applied Statistical Genetics with R" by Andrea S. Foulkes [5]. In that book, the author introduces the reader to the techniques of handling genetic data arising from population based association studies.

The prime focus of the book is the population-based study of unrelated individuals. Family-based association studies are not discussed in detail.I would like to summarize some of the terms, their meanings and few facts which will be useful while understanding the report.

Gene : A length of the DNA that codes for the production of polypeptide molecules (i.e. proteins) [6, p.222].

Allele : Different variants of the same gene are called alleles [6, p.225].

Genotype : Alleles that an organism has form its genotype.[6, pp. 225-226].

Homozygous and Heterozygous : A genotype in which the two alleles of a gene are the same is called homozygous.If the two alleles are different for a genotype, then it is called heterozygous [6, p. 225].

Haplotype : Specific combination of alleles that are in alignment on one of the homologous chromosomes [5, §1.2.1].

Single Nucleotide Polymorphism (SNP) : An SNP represents a difference in a single DNA building block, called a nucleotide.Most Most SNPs in humans are bi-allelic.[1].

Functional SNP : It is the SNP most likely to be associated to the disease [5, §1.1.1].

Phenotype : The observable characteristics (corresponding to a genotype) of an individual are called their phenotypes [6, pp. 225-226].

Genetic Association Studies : The studies which relate genetic sequence information derived from unrelated individuals to a measure of disease progression or disease status. Different types of genetic association studies has been. explained in the textbook [5, p. 1].

Data Sets used in examples of report

There are two publicly available data sets being used in almost all examples of that book. They are :

1. **The FAMuSS data**: The Functional SNPs Associated with Muscle Size and Strength was to identify genetic determinants of skeletal muscle strength and size before and after exercise training using data from 1397 individuals [5, §1.3.3].
2. **The Virco data**: This data includes the protease sequence information of 1066 viral isolates along with fold resistance measures for protease inhibitors [5, §1.3.3].

.

[1]The details can be found at `https://ghr.nlm.nih.gov/primer/genomicresearch/snp`

Chapter 1

Some Statistical and Epidemiological Concepts

In this chapter, we will briefly discuss about topics like genetic association studies, some epidemiological concepts and statistical tests that allow us to handle traits during genotype-trait investigations. Firstly we will understand the genetic association studies in the section below which aims to characterize genetic contributers to disease, then in the next section, a few ways of handling genotype-trait data will be given. The discussions in this chapter will represent one class of investigations within the larger field of genomics. One of the most important reason for doing genetic association studies is that they help us uncover disease etiology.

1.1 Genetic Association Studies

In this section, the discussion is based on [5] and [4].A genetic association study aims to test whether a given sequence has involvement in controlling the phenotype of a specific trait, metabolic pathway, or disease.[1]. Genetic Association Studies are of two kinds:

- Population based Association Studies
- Family based Association Studies.

Population based studies aim to relate genetic sequence information from *unrelated* individuals to a measure of disease progression or disease status. On the other hand, family based study involves data collected from multiple individuals within the same family unit. The members of the same family are more likely to be similar than the members from different families. This phenomenon is called *clustering* and implies a within family correlation. This kind of within family correlation is not found in population based studies of unrelated individuals.

Different kinds of population based studies are:

- Candidate polymorphism studies
- Candidate gene studies
- Genome-wide Association Studies
- Fine mapping studies

In candidate gene (polymorphism) studies, we use the underlying biology of the disease known to us to pick candidate genes (SNPs) that we believe might be relevant to the disease phenotype under consideration. In these studies we aim to test the null hypothesis that the variable gene

[1] `https://www.nature.com/subjects/genetic-association-study`

(SNP) under investigation is indeed functional and hence an association exists between the gene (SNP) and the disease trait.

Considering the risk associated with just one gene, when we know that there might be a lot more genes associated with the disease phenotype is considered a drawback for the candidate gene approach. Also the biology known to us is limited. This also is a drawback since the choice of candidate gene depends on already known biology.

Hence statisticians moved to "Genome-Wide Association Studies" (GWAS). The idea here is that we scan the entire genome, without having to know anything about which genes are related to the disease phenotype, in order to identify associations between genes and disease trait. In fact this method, in theory, should let us identify a number of other genes associated with the disease.

The last type of studies called the Fine Mapping Studies aim to identify, with a high level of precision, the location of a disease causing variant. In other words, we aim to determine precisely where on the genome the mutation that causes the disease is positioned.

1.1.1 Correlation, Causation and Association

In general English, all the above words are used interchangeably but statistically, association is synonymous with dependence and is *different* from correlation. Association is *any* general relationship where one variable provides information about the other. Correlation on the other hand is a *type* of association and different correlation coefficients are used to quantify the increasing or decreasing trends of variable [1].

Causation $\Rightarrow$ Association $\not\Rightarrow$ Correlation

Causation $\not\Leftarrow$ Association $\Leftarrow$ Correlation

1.2 Confounding and Effect mediation

In this section some definitions and examples from [5, pp. 33-36], [10] and [12, §3.5] will be discussed. Since we have understood what genetic association studies are, we now try to understand some important concepts like confounding, effect mediation, interaction and conditional association which are relevant to any population based studies. We will see the importance of these concepts in the coming chapters, for e.g. the role of race and ethnicity as a confounder in genotype association studies has been discussed in the second chapter.

So understanding the concepts of this section will be useful to understand genotype trait associations better.

1.2.1 Exposure and Outcome

Our aim is to characterize the association between two variables, an *exposure* and an *outcome*. To understand the terms, consider the study where we want to determine if heavy alcohol consumption is associated with total cholesterol level. Here a person *exposed* to heavy alcohol consumption might have observed a change in cholesterol levels as an *outcome*. Hence, heavy alcohol consumption is the exposure variable and total cholesterol is the outcome variable.

1.2.2 Confounding

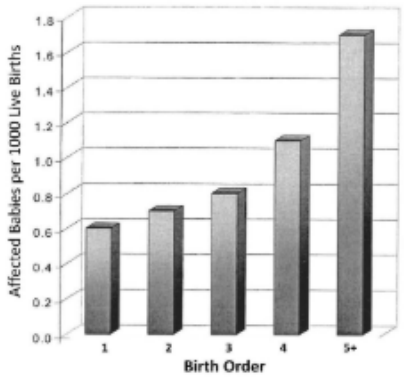
Confounding is a distortion in the estimated measure of association that occurs when the primary exposure of interest is mixed up with some other factor that is associated with the outcome.

Confounder is a variable which follows the following properties :

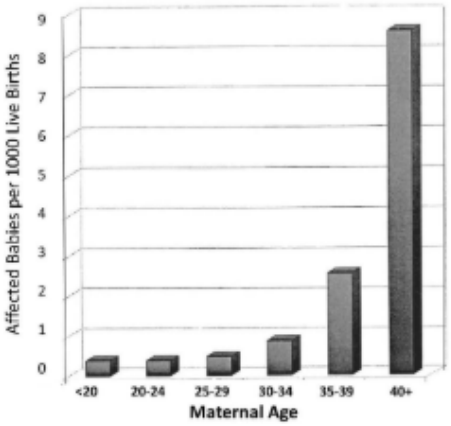
1. variable that is associated with exposure variable.
2. variable that is *independently* associated with outcome variable

3. variable that is *not* in causal pathway between exposure and disease.

Example 1.1. A study was conducted to investigate the association between *birth order* (exposure) and *risk of Down Syndrome*(outcome). The data obtained as the bar graph given in Figure 1.1 shows that birth order is associated to an increased risk of Down Syndrome.



(a) Birth Order vs No. of babies effected per 1000 live births.



(b) Birth Order and maternal age.

Figure 1.1: The graphs have been taken from the section "What is Confounding?" in [10].

But this analysis did not consider any risk factor other than birth order. Since birth order and maternal age during child birth are related, the same group had investigated the relationship between Down Syndrome and maternal age during child birth which turned out to have a more striking relationship as shown in graphs.

Conclusion: The effect of birth order on prevalence of down syndrome is exaggerated and hence incorrect due to the effect of age. The association between birth order and Down syndrome is exaggerated by the confounding effect of maternal age. So confounding variable here is maternal age when other two are exposure and outcome.

On the other hand if *risk of Down Syndrome* is outcome variable for the exposure variable *maternal age*, then birth order will be a counfounder **if** it is not independently associated with the risk of Down Syndrome. △

1.2.3 Effect Mediator

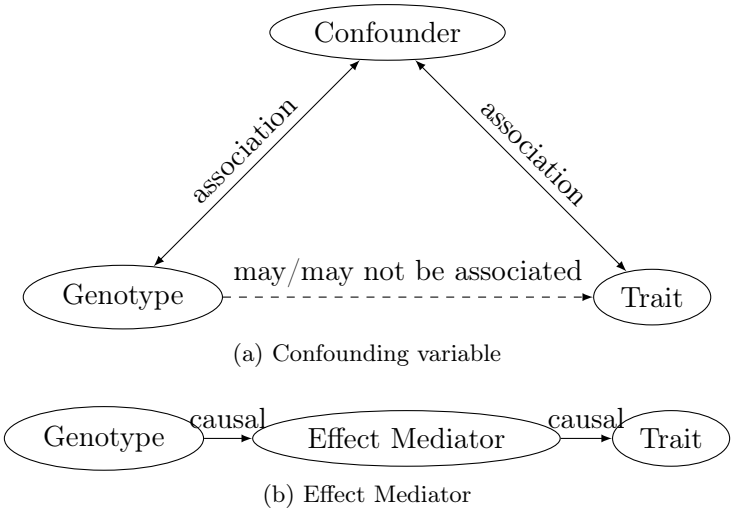
A variable that lies in the causal pathway between the exposure and outcome of interest is called an effect mediator or causal pathway variable.

Example 1.2. It is known from biology that *modest alcohol consumption* (exposure) reduces the *risk of coronary heart diseases* (outcome) among people. But this reduction in the risk of heart problems is due to increase in the blood levels of HDL (good cholesterol) as a result of modest alcohol consumption. Higher levels of HDL are associated with reduced risk of heart disease.

Conclusion: Amount of increase in levels of HDL is an effect mediator since it is part of the mechanism by which alcohol produces its beneficial effect. △

1.2.4 Effect Modification

Effect modification occurs when the effect of the primary exposure on the outcome differs depending on level of a third variable called **modifier**. When effect modification happens, we say



(a) Confounding variable

(b) Effect Mediator

Figure 1.2: Schematic diagram explaining the difference between Effect Mediator and Confounder

that the exposure variable and modifier *interact in a statistical sense* in their association with the outcome.

Interaction is defined statistically by the alternate hypothesis $H_1 : \beta_1 \neq \beta_2$. So rejection of the null hypothesis $H_0 : \beta_1 = \beta_2$. β_1 represents effect of exposure on outcome when effect modifier is 0 (i.e. not present) and β_2 represents effect of exposure on outcome when effect modifier is 1 (i.e. present)

Example 1.3. Scientists have observed association between taking aspirin during viral illness and development of Reye Syndrome. This happens more often for kids than adults i.e. kids 14 years and younger who have aspirin during the treatment of a viral illness are much more likely to develop Reye's syndrome than people older than 14 years.

Conclusion: Age of the patient here is the modifier whose level (i.e. younger or older than 14)affects the outcome (development of Reye's syndrome) for the given exposure (aspirin treatment) to the patient. △

1.2.5 Confounding versus Effect Modification

When confounding occurs, we observe an **incorrect association** because there is a third variable that is associated with both exposure and outcome but not a causal factor. On the other hand, no incorrect observation is obtained when effect modification happens. The effect observed is **real** but the magnitude of effect varies for different groups of individuals.

1.2.6 Conditional Association

Conditional association is said to exist when effect of variable x on y is statistically significant within any one or both levels of a third variable z. Statistically it tests the composite null $H_0 : \beta_1 = 0$ and $\beta_2 = 0$. If either one of β_i are not 0, then we can say that conditional association exists.

Example 1.4. Assume a hypothetical situation where we try to make a table showing number of men and women affected by cancer. The table changes completely once we put condition on type of cancer. If the table is for men and women having cancer conditional to the cancer being breast cancer, then the table will show more women to be affected than men. On the other hand cancer affected individuals conditional to it being prostate cancer, would have more men than women being affected. △

1.3 Few ways to handle a binary trait

The choice of test for association depends on the hypothesis under consideration and the structure of data given to us. A few types of tests and the setting under which it needs to be used is mentioned in this section [5, pp. 38-39] [14]. In particular, we will give a brief idea about how to deal with binary traits and methods to assess genotype-trait association.

1.3.1 Contingency table and Odds Ratio

Contingency tables or Two-way tables summarize the data in a table. As an example, consider the human genetic setting where genotype at a given SNP has three levels: homozygous wildetype (AA), heterozygous (Aa) and homozygous rare (aa). (Note that "A" is the major allele.) The Contingency table given below summarizes genetic data as shown below:

	aa	Aa	AA	total
$D+$	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
$D-$	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	N

Table 1.1: Contingency table for genotype-disease data

Here $D+$ and $D-$ represent "diseased" and "not diseased" respectively (trait). In a setting where we have a contingency table, a commonly used measure of association is the **Odds Ratio (OR)**.

In gene-disease association setting OR is defined as the ratio of the odds of the disease among the exposed to the odds of disease among the unexposed. "Exposure" for us is the genotype. Hence OR is ratio of odds of disease given a specific genotype to the odds of disease among individuals without the specified genotype. If $E+$ and $E-$ denote exposed and not exposed respectively, then

$$OR = \frac{Pr(D+|E+)/[1 - Pr(D+|E+)]}{Pr(D+|E-)/[1 - Pr(D+|E-)]}$$

In the above example about genotypes,

$$OR = \frac{(n_{11}/n_{\cdot 1})/(n_{21}/n_{\cdot 1})}{(n_{12}/n_{\cdot 2})/(n_{22}/n_{\cdot 2})} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

A test of no association between exposure and disease is equivalent to testing the null hypothesis $H_0 : OR = 1$.

The following function from "epitools" package is used to calculate Odds Ratio in R

```
oddsratio()
```

1.3.2 Pearson's chi-squared test

In general, the pearson's chi-squared test can be used for one of the following cases

- to test for **homogeneity** i.e. to test the null hypothesis that two or more multinomial distributions are equal.
- to test for **association** between two categorical variables. This is also called as the test for independence between variables. Here we test the null hypothesis that there is *no* association or that there is independence between variables.

- used as a **goodness of fit** test where we are given a theoretical distribution (eg. normal) and we aim to test for the null hypothesis that the data given to us is representative of the theoretical distribution.

In each of the above applications, we use the following statistic to arrive at a conclusion

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{{E^2}_{ij}}$$

O_{ij} represents the observed cell counts and E_{ij} represents the expected cell counts.

Here, I would like to focus on the Chi-squared test for association which is a non-parametric test that uses contingency table to analyse data. For a $m \times n$ table, number of degrees of freedom for χ^2 statistic is $(m-1)(n-1)$. The test assesses for associations between the categorical variables but does **not** provide any information about causation. It is used when less than 20 percent of the cells have *expected* (no restriction on observed count) frequencies below 5 .If the expected frequencies does not hold the above criterion, then we use the fisher's exact test which is explained in the next subsection.

To summarize, the following three requirements have to be met before we use the chi-squared test:

1. Sampling method involved is simple random sampling
2. Variables under study are categorical
3. When the data is displayed in a contingency table, then less than 20 percent of the cells should have expected count less than 5.
4. The observations are always assumed to be independent of each other. This means chi-squared cannot be used to test correlated data

Example 1.5. Chi-squared test for association: We use the FAMuSS study here to determine if there is an association between the SNPs within "ers1" gene and having a BMI more than 25 (BMI > 25).

Set-up: Here we have two categorical variables: BMI and genotype. BMI has two levels: > 25 and < 25. Genotype variable has three levels: AA, Aa, aa. First, we use R code to extract the variables corresponding to the SNPs within "ers1" gene as shown below. We construct a separate variable called fmsEsr1 to store the information of all SNPs within the ers1 gene.

```
fms <- read.delim(choose.files(), header=T, sep="\t")
attach(fms)
NamesEsr1Snps <- names(fms)[substr(names(fms),1,4)=="esr1"]

NamesEsr1Snps
[1] "esr1_rs1801132" "esr1_rs1042717" "esr1_rs2228480" "esr1_rs2077647"
[5] "esr1_rs9340799" "esr1_rs2234693"

fmsEsr1 <- fms[,is.element(names(fms),NamesEsr1Snps)]
```

Now we define our binary trait such that it is 1 when BMI is more than 25. We will see a few contingency tables of SNP vs trait to understand the data.

```
Trait <- as.numeric(pre.BMI>25)

table(esr1_rs1801132, Trait)
              Trait
```

```
esr1_rs1801132   0   1
            CC 349 189
            CG 252 136
            GG  50  19

table(esr1_rs1042717, Trait)
              Trait
esr1_rs1042717   0   1
            AA  30  30
            GA 246 130
            GG 380 184
```

Hence, we can have a contingency table of observed values w.r.t. each SNP of the esr1 gene. Similarly for each SNP, we can have a table of expected values which are calculated assuming independence of variables.

In terms of statistics, to find the association between a column (SNP) and the trait, we first get a contingency table of the SNP and trait under consideration. Now we apply chi-squared test for association to this table to test the null hypothesis that there is no association between given SNP and trait.

```
 newFunction <- function(Geno){
+  ObsTab <- table(Trait,Geno)
+  return(chisq.test(ObsTab)$p.value)
+  }
```

Above defined function called "newFunction" takes in the variable "Geno" (here, SNP corresponding to a column of fmsEsr1), constructs a contingency table w.r.t. "Geno" and "Trait" and then applies a chi-squared test to this table to finally return the p-value of the test.

```
apply(fmsEsr1,2,newFunction)
[1] esr1_rs1801132 esr1_rs1042717 esr1_rs2228480 esr1_rs2077647
     0.4440720      0.0264659      0.1849870      0.1802880
[5] esr1_rs9340799 esr1_rs2234693
     0.1606800      0.1675418
```

Above apply() function in R, applies the "newFunction" to the columns (2 = columns. 1 = rows) of fmsEsr1.

Conclusion: Hence at a level of significance of $\alpha = 0.05$, the null hypothesis corresponding to "esr1_rs1042717" can be rejected. In terms of biology, this test gives enough evidence to say that there is an association between the SNP "esr1_rs1042717" and BMI. △

Remark 1.1. Note that here we are doing multiple hypothesis testing. In such cases, corrections are required to control error which has been ignored in the above example.

1.3.3 Fisher's Exact Test

We use this test when there are two categorical variables and at least 20 percent of cells have the expected count less than 5. The data given to us is first presented in a contingency table. The overall idea here is to find the **exact** probability of observing the data at hand or a data that is more extreme than that.

In the case of Fisher's exact test, we test the null hypothesis that there is *no* association between the two categorical variables. Fisher's exact test is more accurate than the chi-squared test of independence when the expected numbers are small. In case of a chi-squared test, a large

sample size is required as the χ^2 statistic becomes approximately chi-squared as the sample gets very large. Hence Fisher's exact test is for a relatively small sample size and chi-squared test is for a relatively large sample size. Assumptions for conducting this test are:

1. both the variables involved are categorical.
2. the expected count is less than 5 for more than 20 percent of the cells when data is displayed as a contingency table.
3. Independence of individual observations.
4. Marginals are assumed fixed and given.

To understand how Fisher's exact test works, consider the special case where there are two categorical variables and each variable has two levels as shown below. I have used "success" and "failure" as two levels of one of the categorical variables to easily see the connection of the table with the hypergeometric distribution.

	Level 1	Level 2	Total
Success	x	$k-x$	k
Failure	$n-x$	$(N-k)-(n-x)$	$N-k$
Total	n	$N-n$	N

Table 1.2: Contingency table

From the table above, we can see that the probability of having the given values is given by

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

Note that $P(X=x)$ is the probability of obtaining exactly x successes from a sample of n elements drawn at random, without replacement, from a population of size N. There are a total of k successes in the population. Note that the number of degrees of freedom here is 1 since knowing any one of the values in the table is enough to find all other values, assuming we know the marginals. Hence the probability of observing a given data set is the same as $P(X=x)$.

Example 1.6. As an example, consider the following table :

	Level 1	Level 2	Total
Success	3	0	3
Failure	1	4	5
Total	4	4	8

Table 1.3: Contingency table

We wish to find all cases as extreme as Table 1.3 and more extreme than that assuming that the marginals are **given**. Now consider the following two extreme cases:

	Level 1	Level 2	Total
Success	3	0	3
Failure	0	5	5
Total	3	5	8

(a) Case-1

	Level 1	Level 2	Total
Success	0	3	3
Failure	4	1	5
Total	4	4	8

(b) Case-2

Table 1.4: Extreme cases

To find the required probability of getting data as extreme or more extreme than what we have (i.e. Table 1.3) , we do not consider Table 1.4a as the it does not keep marginals fixed i.e. the marginals of Table 1.3 and Table 1.4a are not equal. So while calculating total probability of extreme cases, we do not consider this case. So, in case of a **one-tailed** probability, we only consider total probability of getting as extreme or more extreme to be probability of obtaining data as in Table 1.3

On the other hand, if we wish to calculate a two-tailed probability, then we need to consider cases which are extreme in the opposite direction as well. Table 1.4b clearly keeps the marginals fixed and also is extreme in the direction opposite to that of Table 1.3. So for a **two-tailed** probability, we add probabilities of obtaining data as in Table 1.3 and Table 1.4b.

For the above example,

$$\text{one-tailed probability } = P(X=3) = \frac{\binom{3}{3}\binom{5}{1}}{\binom{8}{4}}$$

$$\text{two-tailed probability } = P(X=3) + P(X=0) = \frac{\binom{3}{3}\binom{5}{1}}{\binom{8}{4}} + \frac{\binom{3}{0}\binom{5}{0}}{\binom{8}{3}}$$

△

Hence, due to the above procedure of calculating exact probability, this method is called an "exact" test. In R, we use the following function for this test:

```
fisher.test()
```

Remark 1.2. When some of the expected values are small, Fisher's exact test is more accurate than the chi-square test of independence. If all of the expected values are very large, Fisher's exact test becomes computationally impractical; fortunately, the chi-square test will then give an accurate result.

Example 1.7. The second example is about using Fisher's exact test in R. We again use the FAMuSS study like in chi-square example. Since the sample size is large, both tests can be applied to get a result. The R program is very similar to the one we used for chi-squared test. Therefore, I will not explain the program again, please refer to Example 1.5.

```
fms <- read.delim(choose.files(), header=T, sep="\t")
attach(fms)
NamesEsr1Snps <- names(fms)[substr(names(fms),1,4)=="esr1"]

NamesEsr1Snps
[1] "esr1_rs1801132" "esr1_rs1042717" "esr1_rs2228480" "esr1_rs2077647"
[5] "esr1_rs9340799" "esr1_rs2234693"

fmsEsr1 <- fms[,is.element(names(fms),NamesEsr1Snps)]
Trait <- as.numeric(pre.BMI>25)

 newFunction2 <- function(Geno){
+  ObsTab <- table(Trait,Geno)
+  return(fisher.test(ObsTab)$p.value)
+  }

apply(fmsEsr1,2,newFunction2)
[1] esr1_rs1801132 esr1_rs1042717 esr1_rs2228480 esr1_rs2077647 esr1_rs9340799
```

```
     0.46053113      0.02940733      0.18684765      0.17622428      0.15896064
[6]  esr1_rs2234693
     0.16945636
```

Conclusion: From the p-values, we see that the null hypothesis corresponding to second SNP "esr1_rs1042717" can be rejected. Hence there is evidence that suggests an association between esr1_rs1042717 and BMI.

Observation: Observe that p-values corresponding to both fisher's test and chi-squared test are very close to each other. Both suggest that there is an association between the second SNP and BMI. This is because due to a large sample size, both tests could be used to produce results. △

Remark 1.3. Like in the chi-squared example, there is multiple hypothesis testing that has been done here. Such multiple testing requires corrections which haven't been applied here.

1.4 Few ways to handle a quantitative trait

The previous section gave us an idea about how to deal with a binary trait. Instead, if our trait is quantitative, our tests might be different. So here we try to look at a few tests that can help us understand association between a genotype and a quantitative trait.

1.4.1 Two-sample t-test

This test is used when we wish to compare the mean of two different populations i.e. when we want to test the null hypothesis, $H_0 : \mu_1 = \mu_2$. In order to test this null, we draw a smaller sample of individuals from the two populations and using these samples we try to infer about the populations as a whole.

Example 1.8. Assume that we want to test if the average number of individuals with homozygous wild type (AA) genotype is the same as that of the individuals with heterozygous genotype (Aa). Then in such a situation, we can use the two-sample t-test where μ_1 = population mean for people AA genotype and μ_2 = population mean for people with Aa genotype. △

The conditions under which we can use this test are :

1. The two samples which are used to test the null hypothesis should be independent of each other.
2. The two populations from which the data are sampled are each **normally distributed**.
3. Both populations should have **equal** variance (variance is unknown).

When above conditions are met, then we use the following t-statistic to arrive at a conclusion by either using the p-value approach or the critical value approach.

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_p^2[\frac{1}{n_1} + \frac{1}{n_2}]}} \sim T_{n_1+n_2-2}$$

$\bar{X}$ and $\bar{Y}$ indicate the means of the samples drawn from the two populations. s_p^2 represents the pooled variance. Let the two samples drawn have sample sizes n_1 and n_2. The above t-statistic follows a T distribution with $n_1 + n_2 - 2$ degrees of freedom when the null hypothesis is true.

On the other hand, consider the following conditions,

1. The two groups used to test the null hypothesis should be independent of each other.

2. The two populations from which the data are sampled are each **normally distributed**.

3. Both populations have **unequal** variance.

Statisticians haven't yet found a statistic whose distribution is known under null and can be used to derive conclusions when the above conditions are met. Instead, we use a statistic whose distribution is approximately known to us under the null, to arrive at a conclusion. That statistic is

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx T_\gamma$$

This statistic follows an approximate t-distribution with number of degrees of freedom equal to γ where γ is given by

$$\gamma = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

1.4.2 Mann Whitney U Test or Wilcoxon Rank Sum Test

To use the two-sample t-test, one of the assumptions given was that the outcome variable is normally distributed (we can even consider approximately normal distributions or the samples that are sufficiently large so that we can make use of the Central Limit Theorem). When comparing two independent samples, if the outcome is not normally distributed and the sample size is small, then the t-test mentioned above can't be used.

Since the sample is small in the above situation, we might have to look for a non-parametric test. A popular non-parametric choice which is used in the above mentioned scenario is the "Wilcoxon Rank Sum Test", also called as the "Mann Whitney U Test". Wilcoxon Rank Sum Test is used to test whether two samples have been derived from the same population or not (i.e., that the two populations have the same shape or not). Some investigators interpret this test as comparing the **medians** between the two populations. So the null hypothesis being tested is H_0 : Meadians of the two populations are equal.

The conditions under which we use this test is summarized below

1. The two samples which are used to test the null hypothesis should be independent of each other.

2. outcome is not normally distributed

3. Small sample size.

Hence this test a non-parametric analogue of t-test and is more appropriate when outcome is not normal and sample size is small. The test statistic involved in this test is denoted by U and assuming that the sample sizes of the two samples is n_1 and n_2 respectively, we have $U = \min\{U_1, U_2\}$ where

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Here, R_1 and R_2 are sum of *ranks* for first and second group respectively. Using the example below we see how to assign ranks to the data given.

Example 1.9. Assume that an experiment has been designed to investigate the effectiveness of a new drug for asthama. 10 participants are involved in this experiment where each one receives a placebo or the new drug. Participants are asked to record the number of episodes of shortness

Placebo	7	5	6	4	12
New Drug	3	6	4	2	1

Table 1.5: No. of episodes of shortness of breath

of breath over a 1 week period following receipt of the assigned treatment. Let the data obtained be as follows

Procedure to assign ranks to the outcome variable (no. of episodes of shortness of breath) is explained through an example. We first need to order the outcome variable from smallest to largest as shown in the table below. Finally, the procedure to assign ranks to this ordered data is easily observed from the same table:

Placebo	New Drug	Ordered placebo	Ordered new drug	Placebo ranks	New Drug ranks
7	3		1		1
5	6		2		2
6	4		3		3
4	2	4	4	4.5	4.5
12	1	5		6	
		6	6	7.5	7.5
		7		9	
		12		10	
				$R_1 = 37$	$R_2 = 18$

Table 1.6: Assigning Ranks

Hence using a table like Table 1.6, we can find out the value of the test statistic U.

$$U_1 = 5 \cdot 5 + \frac{5 \cdot 6}{2} - 37 = 3 \text{ and } U_2 = 5 \cdot 5 + \frac{5 \cdot 6}{2} - 18 = 22$$
$$U = \min\{U_1, U_2\} = 3$$

What do we observe in the table?
Let the placebo group be called group 1 and the new drug group be called group 2. The sum of the ranks of group 1 is 37 and sum of the ranks for group 2 is 18. As it can been seen in Table 1.6 , the higher rank values are clustered near the placebo group while the lower values are mostly given to new drug members. If the two populations were equal, then high and low scores are evenly distributed among both the groups, hence we can expect R_1 and R_2 to be close to each other.

At first glance, our data doesn't seem to support the null hypothesis. But is the observed difference in sum of ranks statistically significant or did it appear simply due to chance? To answe this , we need to understand the test statistic U even better.

There are two extreme cases possible in this set-up

1. Consider the situation where there is **complete separation** of the groups, supporting the research hypothesis that the two populations are not equal. If all of the higher numbers of episodes of shortness of breath (and thus all of the higher ranks) are in the placebo group, and all of the lower numbers of episodes (and ranks) are in the new drug group and that there are no ties, then:

$$R_1 = 6 + 7 + 8 + 9 + 10 = 40 \text{ and } R_2 = 1 + 2 + 3 + 4 + 5 = 15$$
$$U_1 = 5 \cdot 5 + \frac{5 \cdot 6}{2} - 40 = 0 \text{ and } U_2 = 5 \cdot 5 + \frac{5 \cdot 6}{2} - 15 = 25$$
$$U = \min\{U_1, U_2\} = 0$$

Hence a clear difference between the population gives $U = 0$

2. Consider a second situation where low and high scores are approximately evenly distributed in the two groups, **supporting the null** hypothesis that the groups are equal. If ranks of 2, 4, 6, 8 and 10 are assigned to the numbers of episodes of shortness of breath reported in the placebo group and ranks of 1, 3, 5, 7 and 9 are assigned to the numbers of episodes of shortness of breath reported in the new drug group, then

$$R_1 = 2+4+6+8+10 = 30 \text{ and } R_2 = 1+3+5+7+9 = 25$$
$$U_1 = 5\cdot 5 + \frac{5\cdot 6}{2} - 30 = 10 \text{ and } U_2 = 5\cdot 5 + \frac{5\cdot 6}{2} - 25 = 15$$
$$U = \min\{U_1, U_2\} = 10$$

Hence larger the value of U, the more evidence we have supporting the null hypothesis.

Now, we must determine whether the observed U in our example supports the null or not. To know that, we determine a critical value of U such that if the observed value of U is less than or equal to the critical value, we reject H_0 in favour of H_1 and if the observed value of U exceeds the critical value we do not reject H_0. There are tables which mention the critical values for U at different levels of significance. Using such a table, the critical value for us is 2 at $\alpha = 0.05$

Conclusion:Our decision rule is to reject $H0$ if $U < 2$. Since $U = 3 > 2$, we do not have enough evidence to reject the null at 5 percent level of significance. Hence the difference in ranks that we initially observed is not statistically significant. △

Since we have seen a computationally easy example, we will now see an example which involves the use of R. The following two examples use R to conduct the t-test and Wilcoxon rank sum test.

Example 1.10. Set-up: We use the FAMuSS data for this example.
Our aim is to determine whether having at least one copy of the variant allele for any of the SNPs within "resistin gene" is associated with a change in non-dominant muscle strength before and after exercise training. Change in non-dominant arm muscle strength before and after exercise training is given by "NDRM.CH"

In terms of statistics, we aim to test the null hypothesis that H_0 : Existence of a variant allele does not effect the change in non-dominant arm muscle strength before and after training i.e
if μ_1 = mean NDRM.CH value when a varient allele exists and let μ_2 = mean NDRM.CH value when there is no variant allele within the SNP, then we wish to test the null $H_0 : \mu_1 = \mu_2$.

To test the mentioned null, we use t-test.

```
%%extract all names of SNPs within resistin gene
NamesResistinSnps <- names(fms)[substr(names(fms),1,8)=="resistin"]

%%Now,all rows of data are included but columns are restricted to SNPS within
resistin gene
fmsResistin <- fms[,is.element(names(fms),NamesResistinSnps)]

library(genetics)

%%test to calculate p-value
TtestPval <- function(Geno){
alleleMajor <- allele.names(genotype(Geno, sep="", reorder="freq"))[1]
GenoWt <- paste(alleleMajor, alleleMajor, sep="")
GenoBin <- as.numeric(Geno!=GenoWt)[!is.na(Geno)]
Trait <- NDRM.CH[!is.na(Geno)]
return(t.test(Trait[GenoBin==1],Trait[GenoBin==0])$p.value)
}
```

In the above code, the variable "alleleMajor" stores the major allele in it. "GenoWt" stores the wildetype genotype in it. "GenoBin" is an indicator for at least one variant allele at the corresponding site. Finally a t-test is conducted between means of samples with GenoBin = 1 (i.e. with a copy of variant allele) and samples with GenoBin = 0 (i.e. no variant allele). Also the only output returned after the t-test is conducted is the p-value

Since we have a function which applies the t-test and returns to us the p-value, we can now use this function on the data called "fmsResistin" to obtain our results . To do that we use the "apply()" function in R as shown below.

```
apply(fmsResistin,2,TtestPval)

[1]resistin_c30t resistin_c398t resistin_g540a resistin_c980g resistin_c180g
   0.04401614     0.08098567     0.11578470     0.27828906     0.03969448
[2]resistin_a537c
   0.06573061
```

Seeing the above p-values, we can say that at $\alpha = 0.05$ we can reject the null hypothesis for the first and fifth SNPs i.e. for "resistin_c30t" and "resistin_c180g".
Conclusion: Hence the SNPs resistin_c30t and resistin_c180g within the resistin gene maybe associated with change in NDRM before and after exercise training.

△

Remark 1.4. We haven't done adjustments that are required for multiple testing. Hence our conclusions are not yet decisive.

To apply the Wilcoxon rank sum test under the same setting, we follow the same steps as above but instead of using "t.test()" , we write "wilcoxin.test" and run the programme.

1.4.3 Analysis of Variance (ANOVA)

ANOVA is the statistical method used to test differences between two or more means. Inferences about means are made by analyzing variance, hence it is called "Analysis of Variance". [11]. It is used to test general rather than specific differences among means as shown below.

Example 1.11. Assume we have population means of four different quantities and we want to test the null hypothesis that all four parameters are equal. That is, we want to test the null $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against the alternate $H_1 : \mu_i \neq \mu_j$ for at least one pair of $i \neq j$.

A non-specific null hypotheses like the one above are called omnibus null hypothesis. △

Note that rejecting the non-specific null as the one shown in example above, only implies that not all means are equal. It does not give information about which pair of means differ. ANOVA is an extension of the two sample t-test that was explained earlier.

The assumptions before we conduct an ANOVA are similar to that of a two sample t-test. They are:

1. Outcome variable is normally distributed or the samples are sufficiently large to apply the Central Limit Theorem.

2. All the groups involved have equal variances.

3. The groups involved are independent

1.4.4 Kruskal-Wallis Test

Like we found a non-parametric analogue for the t-test to use in cases when outcome is not normally distributed and sample size is small, similarly we have a non-parametric analogue for ANOVA which is the Kruskal-Wallis test (KW test).

Kruskal Wallis test is a non-parametric test used to compare outcomes among more than two independent groups. As mention earlier, this is more appropriate to use when sample size is small and the outcome is not normally distributed.

The null and alternate hypotheses for the Kruskal-Wallis non-parametric test are stated as follows:

$$H_0 : \text{The } k \text{ population medians are equal.}$$
$$H_1 : \text{The } k \text{ population medians are not all equal.}$$

Just like in the Wilcoxon rank sum test, we have to pool all the outcomes together and then order them from smallest to largest in order to assign ranks to them. Once the ranks are assigned, the test statistic to be used for this procedure is:

$$H = \left(\frac{12}{N(N+1)} \cdot \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1)$$

where $k =$ the number of comparison groups, $N =$ the total sample size, n_j is the sample size in the j^{th} group and R_j is the sum of the ranks in the j^{th} group. We must now determine whether the observed test statistic H supports the null or alternate hypothesis. Once again, this is done by establishing a critical value of H. If the observed value of H is greater than or equal to the critical value, we reject H_0 in favour of H_1. There exist tables describing critical values for H for $k = 3, 4, 5$.

If there are 3 or more comparison groups and 5 or more observations in each of the groups, it can be shown that the test statistic H approximates a chi-square distribution with $k-1$ degrees of freedom. Thus, in a Kruskal Wallis test with 3 or more comparison groups and 5 or more observations in each group, the critical value for the test can be found in the table of critical values of the chi-squared distribution.

Now we look at an R related example which involves ANOVA and KW test.

Example 1.12. Set-up: We will again use the FAMuSS data here. We want to determine whether there is an association between the "resistin_c180g" SNP and percentage change in non-dominant arm muscle strength before and after exercise training, as measured by NDRM.CH.

Statistically, we wish to check the null hypothesis that the average NDRM.CH value is the same irrespective of whether or not there is a polymorphism at that site. We can conclude this by performing ANOVA. The "lm()" function in R performs ANOVA as shown below

```
fms <- read.delim(choose.files(), header=T, sep="\t")
attach(fms)
Geno <- as.factor(resistin_c180g)
Trait <- NDRM.CH
AnovaMod <- lm(Trait~Geno, na.action=na.exclude)
```

In the above code, observe that we have considered genotype as a three level factor variable, unlike the way we dealt with it earlier as a binary trait. Also, the code "na.action=na.exclude" indicates that we want to exclude individuals with missing values for the trait, coded as NA. We now use "summary()" function in R to get details of this test.

```
summary(AnovaMod)

Call:
lm(formula = Trait ~ Geno, na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max
-56.054 -22.754  -6.054  15.346 193.946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   56.054      2.004  27.973   <2e-16 ***
GenoCG        -5.918      2.864  -2.067   0.0392 *
GenoGG        -4.553      4.356  -1.045   0.2964
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.05 on 603 degrees of freedom
  (791 observations deleted due to missingness)
Multiple R-squared:  0.007296,Adjusted R-squared:  0.004003
F-statistic: 2.216 on 2 and 603 DF,  p-value: 0.11
```

The summary above tells us that the over F statistic obtained is $F = 2.216$ and the p-value obtained is 0.11.

Conclusion for ANOVA: From the above observations, we conclude that we do not have enough evidence to reject the null hypothesis at 5 percent significance. Hence we do not have enough evidence for the association between resistin_c180g and NDRM.CH trait.

Now, to apply the KW test to the same data set, we use the "kruskal.test()" function as shown below:

```
kruskal.test(Trait, Geno, na.action=na.exclude)

Kruskal-Wallis rank sum test

data:  Trait and Geno
Kruskal-Wallis chi-squared = 4.9268, df = 2, p-value = 0.08515
```

Conclusion after KW test: The p-value above also suggests that we do not have enough evidence to reject the null at 5 percent significance level. Hence this test also does not provide evidence for association between resistin_c180g and percentage change in muscle strength of non-dominant arm as measured by NDRM.CH. △

Chapter 2

Some Genetic Data Concepts and Tests

The aim of this chapter is to introduce the reader to two of the most important concepts associated with population based studies, they are: Linkage Disequilibrium and Hardy-Weinberg Equilibrium. But before knowing what they are, we need to first know why these concepts are necessary.

Ignoring the concepts of Linkage Disequilibrium and Hardy-Weinberg Equilibrium during data analysis can lead to erroneous conclusions if proper statistical tools are not devised to deal with these phenomenon. For example, in candidate gene studies, our hypothesis of interest is whether the gene considered is in a causal pathway to the disease or not. In these kind of studies, multiple SNPs are considered within the gene which are chosen based on defined linkage disequilibrium blocks. Hence knowledge of linkage disequilibrium is of great help during candidate gene studies. Similarly, the test for Hardy-Weinberg Equilibrium is also useful because it helps population geneticists to think about what factors lead to evolution in a population. Here evolution means change of allele frequencies from one generation to another. Another application of HWE is in identifying genotyping errors. Note that both these concepts are based on the genetic component of our data. I would like to emphasize that what I mentioned above is one of the many reasons why its important to learn these concepts.

2.1 Linkage Disequilibrium and its measures

In this section the discussions will be based on [5, p. 65], [13], [9], [16], [15] and [8].

Linkage Disequilibrium is a measure of allelic association between two sites of the same chromosome. This concept does not involve the trait that we aim to associate with the available genetic information. To put in another way, linkage equilibrium exists when the law of independent assortment is followed. This statement has been explained below and a detailed explanation of what LD means is given below.

We know that the Mendel's law of **Independent Assortment** states that alleles of two or more *different* genes get sorted into gametes independently of one another. In other words, the allele a gamete receives for one gene does not influence the allele received for another gene. But this need not always be true. Various studies have later revealed that many genes were in fact **linked** i.e. due to less physical distance between two sites, some sets of genes have the tendency to get inherited together during a recombination event. Hence traits do not always assort independently. Mendel was lucky enough to choose traits that were not linked during his experiments with pea plants.

As we have seen, under the law of independent assortment, alleles need to sort independently of one another. So if "A" ,"a" are allels at locus 1 and "B" , "b" are alleles at locus 2 , then $P(AB) = P(A) \times P(B)$ (due to independence) where $P(AB)$ is the probability of the haplotype AB to occur on chromosome while $P(A)$ and $P(B)$ represent frequency/probability of the alleles

A and B respectively. Similarly we can find the probabilities of all possible haplotypes under independence as given in Table 2.1b.

Allele	SNP	Frequency
A	SNP-1	p_1
a	SNP-1	p_2
B	SNP-2	q_1
b	SNP-2	q_2

(a) Allele frequency

Haplotype	A	a	total
A	$p_1 \times q_1$	$p_1 \times q_2$	p_1
a	$p_2 \times q_1$	$p_2 \times q_2$	p_2
total	q_1	q_2	1

(b) Haplotype frequency under independence

Table 2.1

Assume we have n individuals for our study, which means we have $N = 2n$ homologs across n individuals. If our observed data has frequencies of haplotypes as in Table 2.1b, then we have evidence that the two loci are in **linkage equilibrium**. Consider now, a deviation from the equilibrium table as shown below :

Haplotype	B	b	total
A	$n_{11} = N(p_1 \cdot q_1 + D)$	$n_{12} = N(p_1 \cdot q_2 - D)$	$n_{1\cdot}$
a	$n_{21} = N(p_2 \cdot q_1 - D)$	$n_{22} = N(p_2 \cdot q_2 + D)$	$n_{2\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	N

Table 2.2: Expected number of haplotypes during deviation from equilibrium

Linkage Disequilibrium (LD) refers to the occurrence of specific allele combinations at the two loci with frequency greater than that expected by chance as shown in Table 2.2. D in the above table is a measure of magnitude of deviation from equilibrium. Note that linkage disequilibrium is a population level characteristic. In the above table, n_{ij} denotes the observed number of haplotypes corresponding to alleles of row i and column j. And n_{ij}/N gives the observed probability of the corresponding haplotype. This is denoted by p_{ij}.

Remark 2.1. LD measures allelic association in the **same** gamete.

Once we identify a departure from equilibrium, our next step should be to think about how to quntify the amount of departure. This has been discussed briefly in the following pages. The measures of LD which we will learn to quantify the amount of departure from equilibrium are D' and r^2.

2.1.1 Calculating D using given data

As mentioned in the previous section, we will discuss in this section about D' which is a measure of allelic association between SNPs in the same gamete. But before that we will see some results on D.

Lemma 2.1. *For $i, j \in \{1, 2\}$,*

$$D = p_{11} \cdot p_{22} - p_{12} \cdot p_{21}$$

where $p_{ij} = \frac{n_{ij}}{N}$.

Proof.

$$p_{11} \cdot p_{22} = (p_1 q_1 + D)(p_2 q_2 + D) = p_1 q_1 p_2 q_2 + p_1 q_1 D + p_2 q_2 D + D^2 \tag{2.1}$$

$$p_{12} \cdot p_{21} = (p_1 q_2 - D)(p_2 q_1 - D) = p_1 q_1 p_2 q_2 - p_1 q_2 D - p_2 q_1 + D^2 \tag{2.2}$$

if we subtract (2.2) from (2.1), we get:

$$p_{11} \cdot p_{22} - p_{12} \cdot p_{21} = (p_1 q_1 + p_2 q_2)D + (p_1 q_2 + p_2 q_1)D$$

rearranging terms, we get:

$$p_{11} \cdot p_{22} - p_{12} \cdot p_{21} = p_1(q_1 + q_2)D + p_2(q_2 + q_1)D = D$$

Last step is true because $p_1 + p_2 = q_1 + q_2 = 1$

□

D can be expressed in terms of joint probabilities and allele probabilities as well. From Table 2.2, we can see that

$$\boxed{D = p_{11} - p_1 q_1 = p_1 q_2 - p_{12} = p_2 q_1 - p_{21} = p_{22} - p_2 q_2}$$

2.1.2 Upper and lower bounds for D

1. From Table 2.2 we can see that first row elements add up to p_1. So $p_1 q_1 + D \leq p_1$. Then $D \leq p_1 - p_1 q1 = p_1(1 - q_1) \leq p_1$. So $D \leq p_1$. Similarly, since first column elements add up to q_1 , we can again say that $p_1 q_1 + D \leq q_1$. Following steps as in above equation , we get $D \leq q_1$.

   Using the same idea for second column and second row, we get $D \leq p_2$ and $D \leq q_2$.

   So for $i \in \{1, 2\}$, we have

   $$\boxed{D \leq p_i} \quad \text{and} \quad \boxed{D \leq q_i}$$

2. Each entry of Table 2.2 are probabilities, so each term has to be less than or equal to one.

   - $p_1 q_2 - D \geq 0 \implies D \leq p_1 q_2$ and $p_2 q_1 - D \geq 0 \implies D \leq p_2 q_1$.
   - $p_1 q_1 + D \geq 0 \implies D \geq -p_1 q_1$ and $p_2 q_2 + D \geq 0 \implies D \geq -p_2 q_2$.

   $$\boxed{-\min\{p_1 q_1, p_2 q_2\} \leq D \leq \min\{p_1 q_2, p_2 q_1\}}$$

2.1.3 The drawback with D and its modification to D'

The value of D is bounded by functions which depend on allele frequencies and this allele frequency related bounds of D disqualify it as a general measure of association. The reason why this happens is given in detail below.

One would like to

- compare the gametic disequilibrium for *same* loci in *different* populations.
- compare the gametic disequilibrium for *different* loci in the *same* population.

In both the cases, the range of values D can take will depend on allele frequencies which change with change in population or change in site. In other words, different populations can have different allele frequencies at the same locus and similarly different loci within the same population have different allele frequencies. Hence the range of valued D will take will differ from one case to another. Hence comparing two quantities whose range of D is not same won't help us draw proper conclusions to the question - "Which one seems to have more departure from equilibrium? ".

Due to this problem, measures related to D are "normalized" with allelic frequencies. Normalized D has the advantage that it is independent of allele frequencies and always between 0 and 1. This helps us overcome the comparison problem we earlier had. One of many possible normalizations of D' is:

$$\boxed{D' = |D|/D_{\max}} \tag{2.3}$$

$$D_{\max} = \begin{cases} \min\{p_1q_2, q_1p_2\} & \text{if} D > 0 \\ \min\{p_1p_2, q_1q_2\} & \text{if} D < 0 \end{cases} \tag{2.4}$$

Now, I would like to summarize a few important properties of D' that have been discussed so far.

2.1.4 Some important properties of D'

- Range of D' is independent of allele frequencies.
- $0 \leq D' \leq 1$
- Values close to 1 are assumed to indicate high levels of LD and the values close to 0 indicate low departure from equilibrium.

2.1.5 r2 : a measure of LD

Another intuitively appealing measure of LD is given by r^2. This measure is based on Pearson's chi-squared test of independence. It is given by

$$r^2 = \frac{\chi_1^2}{N}$$

where N is the total sample size. This quantity r^2 also has a connection with D which we mentioned earlier. In this section, we will try to derive r^2 in terms of D

Lemma 2.2. *r^2, D and χ_1^2 are related to each other and the relation is given as follows*

$$r^2 = \frac{\chi_1^2}{N} = \frac{D^2}{p_1p_2q_1q_2} \tag{2.5}$$

Proof. we know that the chi-squared statistic looks as follows,

$$\chi_1^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

From Table 2.2 and Table 2.1a , we can see that $O_{ij} - E_{ij} = \pm ND$

$$(ND)^2 = (O_{ij} - E_{ij})^2 \ \forall \ (i,j) \in \{1,2\} \times \{1,2\}$$

From tables Table 2.2 and Table 2.1a,

$$\chi_1^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = (ND)^2 \times \left[\frac{1}{Np_1q_1} + \frac{1}{Np_2q_2} + \frac{1}{Np_1q_2} + \frac{1}{Np_2q_1}\right]$$

$$\chi_1^2 = ND^2 \times \left(\frac{p_2q_2 + p_1q_1 + p_2q_1 + p_1q_2}{p_1p_2q_1q_2}\right)$$

Rewrite the numerator as $p_2q_2 + p_1q_1 + p_2q_1 + p_1q_2 = (p_1 + q_1)(q_1 + q_2) = 1$. We get

$$\chi_1^2 = \frac{ND^2}{p_1p_2q_1q_2}$$

$$\text{Hence, } \boxed{r^2 = \frac{\chi_1^2}{N} = \frac{D^2}{p_1p_2q_1q_2}} \tag{2.6}$$

□

Points to note

1. Observe that both the measures that we have talked about are different "normalizations" of the same quantity D.
2. $0 \leq r^2 \leq 1$
   - If $r^2 = 0$, then loci are in complete linkage equilibrium
   - If $r^2 = 1$, then loci are in complete linkage disequilibrium
3. The measure r^2 is preferred over D' due to its straightforward relationship to the usual chi-squared test for association.

Remark 2.2. The adjustment in both the cases (D' and r^2) involves marginal allele frequencies since the value of D depends on these.

2.1.6 Examples in R

Since we have understood the concepts of LD and and HWE, the next step should be to use our knowledge of LD and HWE to test for their presence in different data sets. A few examples have been given below which show the use of R code to measure the amount of departure from equilibrium using D' and r^2.

Example 2.1. Measuring LD using D' for a pair of SNPs.
Here we use the FAMuSS study which contains information about the genotypes at specified SNPs within various genes. Our focus will be on certain SNPs within the gene "alpha actinin 3" (actn3) and those in "estrogen receptor gene 1" (esr1).

Aim: We wish to calculate and compare the amount of LD using D' in the two cases given below:

1. LD as measured by D' for the SNPs actn3_r577x and actn3_rs540874 which are within the *same* actn gene.
2. Calculate D' between actn3_r577x and esr1_rs1801132 which are within *different* genes, using the following function in R

`LD() function`

The following R code has been used to calculate D'

```
fms <- read.delim(choose.files(), header=T, sep="\t")
library(genetics)
attach(fms)
actn3_r577x[1:10]
```

```
[1] CC CT CT CT CC CT TT CT CT CC
Levels: CC CT TT

actn3_rs540874[1:10]

[1] GG GA GA GA GG GA AA GA GA GG
Levels: AA GA GG

%create genotype vectors
Actn3Snp1 <- genotype(actn3_r577x,sep="")
Actn3Snp2 <- genotype(actn3_rs540874,sep="")
Actn3Snp1[1:10]

[1] "C/C" "C/T" "C/T" "C/T" "C/C" "C/T" "T/T" "C/T" "C/T" "C/C"
Alleles: C T

LD(Actn3Snp1,Actn3Snp2)$"D'"
[1] 0.8858385

Esr1Snp1 <- genotype(esr1_rs1801132,sep="")

LD(Actn3Snp1,Esr1Snp1)$"D'"
[1] 0.1122922
```

Conclusion : The D' value for SNPs within the same actn3 gene (actn3_ r577x and actn3_ rs540874) is closer to 1 and hence it indicates high levels of linkage disequilibrium between the SNPs. On the other hand, the SNPs actn3_ r577x and esr1_ rs1801132 lying in different genes have D' values closer to 0 and hence it indicates low levels of linkage disequilibrium between the SNPs.

Hence we can expect LD measure to be close to 1 for SNPs within the same gene and LD measures to be closer to 0 for SNPs across different genes. △

Example 2.2. Measuring pairwise LD for a group of SNPs
In this example , we will use the same set up as in Example 2.1. We will try to see how we can calculate pairwise LD for a group of SNPs. Let us consider a group of SNPs within the actn3 gene to discuss this example. The following R code is used in this example:

```
fms <- read.delim(choose.files(), header=T, sep="\t")
library(genetics)
attach(fms)
Actn3Snp1 <- genotype(actn3_r577x,sep="")
Actn3Snp2 <- genotype(actn3_rs540874,sep="")
Actn3Snp3 <- genotype(actn3_rs1815739,sep="")
Actn3Snp4 <- genotype(actn3_1671064,sep="")
```

Now,we create a data frame which contains all the four SNPs mentioned above. We name this dataframe object as "Actn3AllSnps"

```
Actn3AllSnps <- data.frame(Actn3Snp1,Actn3Snp2,Actn3Snp3,Actn3Snp4)
```

Applying the LD() function in R to this data frame gives an upper triangular matrix consisting of all possible pairwise LD measures.

```
LD(Actn3AllSnps)$"D'"
          Actn3Snp1 Actn3Snp2 Actn3Snp3 Actn3Snp4
Actn3Snp1        NA 0.8858385 0.9266828 0.8932708
Actn3Snp2        NA        NA 0.9737162 0.9556019
Actn3Snp3        NA        NA        NA 0.9575870
Actn3Snp4        NA        NA        NA        NA
```

Conclusion: We see that there is a high level of LD for all possible pairs i.e. the presence of allele at one of the SNP loci seems to be associated with the presence of a specific alleles at other SNP loci. This conclusion is expected because all SNPs are within the same gene. △

Example 2.3. LD based on r^2 and chi-squared statistic.
Just like in Example 2.1,we are interested in measuring LD between the SNPs "r577x" and "rs540874" within the "actn3" gene based on the FAMuSS data. We use the LD() function in R just like the way we used it for Example 2.1 and Example 2.2 to obtain the value of r^2.

Just like we did earlier, we first need to convert the SNP variables into genotype variables as shown below

```
fms <- read.delim(choose.files(), header=T, sep="\t")
attach(fms)
library(genetics)
Actn3Snp1 <- genotype(actn3_r577x,sep="")
Actn3Snp2 <- genotype(actn3_rs540874,sep="")
```

We can either simply use the LD() function to get lots of information as output or we can choose to use the LD()$ r^2 code to extract the value of r^2 alone.

```
LD(Actn3Snp1,Actn3Snp2)$"R^2"
[1] 0.6179236

LD(Actn3Snp1,Actn3Snp2)

Pairwise LD
-----------
                    D        D'      Corr
Estimates: 0.1945726 0.8858385 0.7860811

               X^2 P-value   N
LD Test: 895.9891       0 725
```

The above output states that the result is based on 725 individuals with complete data on both the SNPs. We obtained, $r^2 = 0.6179236$. Hence, the chi-squared statistic is $\chi^2 = r^2 \times N = 0.6179236 \cdot 725 \cdot 2 \approx 896$ which is the same as what is given in the output above.

Conclusion: The value of r^2 gives us enough evidence to believe that there is a LD between the two SNP loci "r577x" and "rs540874". △

In the Example 2.2 and Example 2.1, we found out estimates for the amount of association between the sites of interest using the quantities r^2 and D'. These quantities do give us some insight into the association between the loci but conclusions should be drawn with some caution because of the following reasons:

1. the χ^2 statistic obtained using the LD() function in R is based on data from a two way table that includes *correlated* data.

2. An additional layer of estimation is performed in the process of calculating the measures $r^2 and D'$, since the haplotypic phase and thus the cell counts are not observable.

These two problems make our tests we considered so far not very trustworthy. Hence one has to account for these issues for more accurate conclusions. Note that accounting for the estimation error mentioned in second point above happens since the haplotypes are not observed and hence need to be estimated.

2.1.7 LD blocks and SNP tagging

We have already considered measuring pairwise LDs between alleles in an earlier example which is Example 2.2. But practically, this does not always happen because we are generally interested to know if a *group of alleles* are in LD or not. One very simple and intuitive measure of LD across a region with multiple SNPs is taking the average of all pairwise measures of D'. Let L be the set of loci within a region of interest. And let D'_{ij} is a measure of LD between loci i, j for $i, j \in L$. Then the measure mentioned earlier given by

$$\bar{D}' = \frac{1}{n_L} \cdot \sum_{i,j \in L} D'_{ij}$$

Where $n_L = \binom{|L|}{2}$ = number of ways of choosing two loci from the set L with cardinality $|L|$ and the summation is over all possible pairs of i and j. An example illustrating how this quantity is measured is given below :

Example 2.4. In this example, we will continue with the FAMuSS data in which we have information about the SNPs within the "actn" gene. (Just like the way we considered it in Example 2.1.) R code to obtain all possible pairwise D' values for SNPs within the actn gene using LD() was shown earlier in Example 2.1. We initially follow the same R code as in Example 2.1 to obtain all pairwise D' values.

```
fms <- read.delim(choose.files(), header=T, sep="\t")
attach(fms)
library(genetics)
Actn3Snp1 <- genotype(actn3_r577x,sep="")
Actn3Snp2 <- genotype(actn3_rs540874,sep="")
Actn3Snp3 <- genotype(actn3_rs1815739,sep="")
Actn3Snp4 <- genotype(actn3_1671064,sep="")
Actn3AllSnps <- data.frame(Actn3Snp1,Actn3Snp2,Actn3Snp3,Actn3Snp4)

LDMat <- LD(Actn3AllSnps)$"D'"
```

LDMat calculates all pairwise LD values and Actn3AllSnps is a dataframe with each element as a genotype object representing an SNP within actn.

```
 LDMat
          Actn3Snp1 Actn3Snp2 Actn3Snp3 Actn3Snp4
Actn3Snp1        NA 0.8858385 0.9266828 0.8932708
Actn3Snp2        NA        NA 0.9737162 0.9556019
Actn3Snp3        NA        NA        NA 0.9575870
Actn3Snp4        NA        NA        NA        NA

mean(LDMat,na.rm=T)
[1] 0.9321162
```

To calculate the average LD, we have used the "mean()" function in R. Note that the phrase "na.rm=T" specifies that missing values should be removed during calculation.

Conclusion: From the mean value of LD, we can say that there is high measure of LD among the SNPs within the actn gene. △

Note that more precise measures of LD can be obtained through fine mapping studies. Through characterizing regions of high LD, the human genome can be divided into **LD blocks**. These blocks are separated by **hotspots**, regions in which recombination events are more likely to occur. In general, alleles tend to be more correlated within LD blocks than across LD blocks. Once the regions of high LD are identified, investigators aim to determine the smallest subset of SNPs that characterizes the variability in this region, a process referred to as **SNP tagging**. Here the goal is to reduce the redundancies in the genetic data. Well defined tag SNPs will capture a substantial majority of the genetic variability within the LD block. The tag SNPs are correlated with the true disease causing variant but are not typically functional themselves. Since these tag SNPs are correlated to the functional SNP, we can use the tag SNPs to derive conclusions about the actual functional variant.

It is important to note that LD blocks differ substantially across race and ethnicity groups. Due to this reason, a set of tag-SNPs may capture information on the true disease causing variant in one racial or ethnic group but may not work for another group. So this phenomenon has to be considered while devising appropriate analytic methods in analysis of population based association studies involving people with different racial and ethnic backgrounds.

2.1.8 LD and Population Stratification

Often, while doing genetic association studies, we would want to consider people from throughout the world in order to get unbiased conclusions about our questions of interest. In such cases our population will be both *stratified* and *admixed*. The emphasized terms have been explained below. The following subsection discusses the association of LD and population *substructure* (i.e. admixture and stratification).

Population stratification refers to the presence of multiple subgroups between which there is minimal mating or gene transfer. Ignoring the presence of stratification in population can lead to erroneous conclusions about the presence of LD between SNPs. This problem has been shown in the example below.

Example 2.5. Consider two SNPs, say SNP-1 and SNP-2. Assume that SNP-1 has alleles A and a and the second SNP has alleles B and b. Let there be two different populations which have both SNP-1 and SNP-2. In the first population let the dominant allele frequencies be $p_A = p_B = 0.8$ while it is $q_A = q_B = 0.2$ in the second population. Assume that the two SNPs are not associated i.e. there is no LD between them. Under this assumption of independence, the observed counts will be similar to what is given in Table 2.3. We assume a total of 100 individuals in each population which means a total of 200 haplotypes for each population.

Population-1	B	b	Total
A	$200 * 0.8 * 0.8 = 128$	$200 * 0.8 * 0.2 = 32$	160
a	$200 * 0.8 * 0.2 = 32$	$200 * 0.2 * 0.2 = 8$	40
Total	160	40	$N = 200$

(a) Population-1

Population-2	B	b	total
A	$200 * 0.2 * 0.2 = 8$	$200 * 0.8 * 0.2 = 32$	40
a	$200 * 0.8 * 0.2 = 32$	$200 * 0.8 * 0.8 = 128$	160
total	40	160	200

(b) Population-2

Table 2.3

△

Assume that we started off with a population which is a combination of the above two populations. So the observed counts for this new combined population is given by simply adding up the number of haplotypes in each group. This combined population data is given Table 2.4.

Combined population	B	b	total
A	$128 + 8 = 136$	$32 + 32 = 64$	200
a	$32 + 32 = 64$	$8 + 128 = 136$	200
total	200	200	$N = 400$

Table 2.4: Combined data of population 1 and 2

We now use R code to do a chi-squared goodness of fit test for our set up explained above. We first create a matrix containing observed values of number of haplotypes. Then we use the "chisq.test()" function in R to obtain the expected values under independence.

```
ObsCount <- matrix(c(136,64,64,136),2)
ObsCount
     [,1] [,2]
[1,]  136   64
[2,]   64  136

ExpCount <- chisq.test(ObsCount)$expected
ExpCount
     [,1] [,2]
[1,]  100  100
[2,]  100  100
```

Note that we estimate the allele frequencies using the observed data at hand. From the values in Table 2.4 we can estimate the allele frequencies as $p_A = p_a = p_B = p_b = 0.5$. Now since $D = p_{AB} - p_A \cdot p_B$ we can find $|D|$ to be

$$|D| = |p_{AB} - p_A \cdot p_B| = |\frac{136}{400} - 0.5 \cdot o.5| = 0.09$$

$$\text{now, } D_{\max} = 0.25D' = \frac{0.09}{0.25} = 0.36$$

Note that we have used the formulae (2.3) and (2.4) in the above calculation. Our final value for D' is 0.36.

Conclusion: The value of D' we obtained hints at presence of mild linkage disequilibrium between the two SNP sites which is completely wrong. This misleading conclusion about the presence of association between SNPs is because we did not consider the underlying substructure (stratification in this case) for our population. Hence population stratification is an important factor that has to be taken into account while doing population based studies.

Another point to note is about the effect of population admixture on LD. **Population admixture** is said to occur when mating occurs between two populations for which the allele frequencies differ. It is a well known fact that population admixture does effect LD values if proper corrections are not made. I have not elaborately explained the effects of population admixture and population substructure or the methods to correct them in this report.

2.2 Hardy-Weinberg Equilibrium (HWE) and its measures

The discussions in this section are based on [7] and [2]. This is an important concept in the context of population based association studies. LD and HWE are different types of measures

of allelic association. In this section, we wish to focus on HWE as we have already given a brief introduction to LD in the previous section. HWE is said to exist at a site of the chromosome when the probability of an allele occurring on one homolog does not depend on which allele is present at the second homolog. This definition might sound similar to that of LD but there is a clear difference between the two terms. The difference is highlighted below:

- LD refers to allelic association across different sites on a single homolog.
- HWE denotes independence of alleles at a single site between two homologous chromosomes.

As we saw earlier, linkage equilibrium is said to exist when law of independent assortment is followed. Similarly HWE is said to exist when the **law of segregation** is followed. HWE has been explained in further detail in the following paragraphs.

As we already know, a diploid individual carries two individual copies of each autosomal gene (i.e., one copy on each member of a pair of homologous chromosomes), one copy from mother and other from father. Each gamete produced by a diploid individual receives only one copy of each gene, which is chosen at random from the two copies found in that individual. Mendel's Law of Segregation states that each of the two copies in an individual has an equal chance of being included in a gamete.

Hence, under HWE, genotype frequency is the product of corresponding allele frequencies i.e. for e.g. $p_{Aa} = p_A \times p_a$ due to independence of the occurrence of alleles on homologs. If certain conditions are met, then it can be shown that a consequence of HWE is to have constant allele frequencies over generations. When allele frequencies get constant over the generations, we say that the population is "not evolving". Hence, HWE can also be defined as the phenomenon of finding constant allele frequencies over the generations. But under what conditions does this constancy occur? This question has been answered below along with why these assumptions are necessary.

The Hardy-Weinberg Equilibrium occurs only when the population conforms to the following assumptions:

1. Natural selection is not acting on the locus in question (i.e., there are no consistent differences in probabilities of survival or reproduction among genotypes).
2. Neither mutation (the origin of new alleles) nor migration (the movement of individuals and their genes into or out of the population) is introducing new alleles into the population.
3. Population size is infinite, which means that genetic drift is not causing random changes in allele frequencies. In fact ,all natural populations are finite and thus subject to drift, but we expect the effects of drift to be more pronounced in small than in large populations.
4. Individuals in the population mate randomly with respect to the locus in question.

If all the above conditions are met, then it is shown in the next subsection that the allele frequencies will remain constant over generations. If any one of these assumptions is not met, then the population will evolve and hence not be in Hardy-Weinberg equilibrium. Populations are usually not in Hardy-Weinberg equilibrium (at least, not for all of the genes in their genome) because population size can't be infinite. Instead, populations tend to evolve. Hence there is never a perfect HWE. [1]

Having observed that HWE is impossible to exist, our next question should be about why there is a need to understand and check for HWE?
The answer is that the population geneticists check to see if a population is in Hardy-Weinberg

[1]For more details, see:`https://www.khanacademy.org/science/biology/her/heredity-and-genetics/a/hardy-weinberg-mechanisms-of-evolution`

equilibrium because they suspect other forces may be at work. So if the population is evolving, then they try to look for reasons behind this evolution. Search if for why there is no equilibrium between the loci. Hence HWE is a useful concept.

Consequences of HWE

In this subsection, we look at a few consequences of HWE. The first consequence has been mentioned in previous ubsection as a fact. We now wish to prove the statement.

Lemma 2.3. *Allele frequencies will not change in a population from generation to generation under HWE, provided above conditions are met.*

Proof. Consider a locus with alleles A and a. Let the the generation we are looking at now be the "First generation". Assume the genotype frequencies to be as follows:

Genotype	AA	Aa	aa
Frequency	u	v	w

Table 2.5: Arbitrary genotype frequency in 1st generation

Clearly, $u + v + w = 1$ since we assume that there is no way new alleles can occur in our population. From above values, it is easy to calculate that

$$P(A) = u + \frac{1}{2} \times v$$
$$P(a) = w + \frac{1}{2} \times v$$

Assume that the locus is in HWE, then the second generation will look as follows

Mating type	Mating frequency	Diff. Progeny ratio
$AA \times AA$	u^2	only AA
$AA \times Aa$	$2uv$	$AA : Aa = 1 : 1$
$AA \times aa$	$2uw$	only Aa
$Aa \times Aa$	v^2	$AA : Aa : aa = 1 : 2 : 1$
$Aa \times aa$	$2vw$	$Aa : aa = 1 : 1$
$aa \times aa$	w^2	only aa

Table 2.6: 2nd generation frequency under HWE

Then for this generation, define p to be the frequency of AA genotype, q for Aa and r for aa. Then

$$p = P(AA) = u^2 + 0.5 \times (2uv) + \frac{1}{4} \times v^2 = (u + o.5v)^2$$
$$q = P(Aa) = uv + 2uw + 0.5 \times v^2 + vw = 2(u + o.5v)(0.5v + w)$$
$$r = P(aa) = 0.25v^2 + 0.5 \times (2vw) + w^2 = (w + 0.5v)^2$$

Now, we try to calculate the third generation genotype frequencies Since, $P(AA) = p$, $P(Aa) = q$ and $P(aa) = r$, we can say that

$$P(A) = p + \frac{1}{2} \times q \text{ and } P(a) = r + \frac{1}{2} \times q$$

Then just like we did earlier,

$$P(AA) = (p + 0.5q)^2 = [(u + 0.5v)^2 + 0.5 \times 2(u + 0.5v)(0.5v + w)]^2$$

$$= [(u + 0.5v)[(u + 0.5v) + (0.5v + w)]]^2$$
$$= [(u + 0.5v)(u + v + w)]^2$$
$$= [(u + 0.5v)(1)]^2$$
$$= [u + 0.5v]^2 = p$$

Similarly we can calculate $P(Aa)$ and $P(aa)$ and get q and r respectively. Hence in the third generation, the genotype frequencies are p, q and r. Similar calculations show that the same genotype frequencies are continued every generation from now on. □

Corollary 2.1. *One generation of random mating under HWE is sufficient to make all genotype frequencies constant.*

Proof. This is clearly observed from the previous theorem. □

2.2.1 Measures of HWE

Having understood what HWE is, one needs to move to the next step of quantifying the amount of departure from HWE. The tests which help us quantify the level of departure from HWE are:

1. Pearson's chi-squared test
2. Fisher's Exact Test

These tests have been discussed in detail in 1, one can go back to see what they are. A contingency table giving the genotype counts is given below. However, its important to note that the genotypes Aa and aA are indistinguishable in population based investigations. We observe the sum $n_{21}+n_{12}$ which we denote by n_{12*}. In the tests for HWE, we always test the null hypothesis that there exists a perfect Hardy-Weinberg Equilibrium. Given below are some examples which test for the presence of HWE.

Genotype	A	a	Total
A	n_{11}	n_{12}	$n_{1\cdot}$
a	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Table 2.7: Genotype counts for two homologous chromosomes

Note that the first row represent alleles on one homologous chromosome and column represents the same on the second homologous chromosome.

Before giving the examples in R, I would first like to briefly explain about the data set that will be used in the examples that follow. In the following two examples we use the HGDP (Human Genome Diversity Project) data set. The HGDP project began in 1991 with the aim of documenting and characterizing the genetic variation in humans worldwide. Genetic and demographic data are recorded on n = 1064 individuals across 27 countries. We consider genotype information across four SNPs from the "v-akt murine thymoma viral oncogene homolog 1"(AKT1) gene. In addition to genotype information, each individual's country of origin, gender and ethnicity are also recorded.

Example 2.6. Aim: Here we consider the HDGP data. Our aim is to test for presence of HWE for the SNP labelled "AKT1.CO756A" using the Pearson's chi-squared test. This is done as follows :

```
hgdp <- read.delim(choose.files(), header=T, sep="\t")
attach(hgdp)
```

```
Akt1Snp1 <- AKT1.C0756A
ObsCount <- table(Akt1Snp1)
Nobs <- sum(ObsCount)

ObsCount
Akt1Snp1
 AA  CA  CC
 48 291 724

FreqC <- (2 * ObsCount[3] + ObsCount[2])/(2*Nobs)
ExpCount <- c(Nobs*(1-FreqC)^2, 2*Nobs*FreqC*(1-FreqC),Nobs*FreqC^2)

ExpCount
Akt1Snp1
[1]  35.22319 316.55362 711.22319
ChiSqStat <- sum((ObsCount - ExpCount)^2/ExpCount)

ChiSqStat
[1] 6.926975
```

The above chi-squared statistic has a single degree of freedom. We now find the quantile corresponding to $1 - \alpha$, where $\alpha = 0.05$.

```
qchisq(1-0.05,df=1)
[1] 3.841459
```

Since $6.93 > 3.84$, we can reject the null hypothesis in favour of the alternate hypothesis.

Conclusion: We rejected the null hypothesis of HWE at this SNP locus and concluded that alleles on the two homologous chromosomes are associated with one another.

Alternate method to test the null hypothesis is to use following function in R that is available in the "genetics" package:

```
HWE.chisq()
```

In this, we first create a genotype object using genotype() function.

```
library(genetics)
Akt1Snp1 <- genotype(AKT1.C0756A, sep="")

HWE.chisq(Akt1Snp1)
Pearson's Chi-squared test with simulated p-value (based on 10000 replicates)

data:  tab
X-squared = 6.927, df = NA, p-value = 0.0104
```

Again based on the above result as well, we will have to reject the null hypothesis because the p-value $(0.0104) < 0.05$. Hence we can conclude the existance of HWD at this site using this R code as well. △

Before going to the next example, where we use the Fisher's test, we need to know about a small calculation. As it was mentioned earlier in 1.3.3, p-value from fisher's exact test is based on summing the exact probabilities of seeing the observed count or something more extreme in

the direction of alternate hypothesis. Also, we showed that the exact probabilities of seeing a contingency table like in Table 2.7 is given by

$$p_A = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{N}{n_{\cdot 1}}}$$

Since n_{12} and n_{21} values can't be found out, we need to find an expression in above equation in terms of the observed values - n_{11}, n_{22} and n_{12}^*. This new expression in terms of observable values is given below.

$$p_A = \frac{\binom{n}{n_{11},n_{12}^*,n_{22}}}{\binom{2n}{n_{1\cdot}}}$$

In the following example, Fisher's exact test is used to arrive at a conclusion about the presence of HWD.

Example 2.7. Here we again use the HGDP data like we used in Example 2.6. Like in previous example, we wish to test for a departure from HWE for the same "AKT1.CO756A" SNP but now we will only consider people within the "Maya" population. First we calculate the observed and expected genotype counts as follows:

```
attach(hgdp)
Akt1Snp1Maya <- AKT1.CO756A[Population=="Maya"]
ObsCount <- table(Akt1Snp1Maya)

ObsCount
Akt1Snp1Maya
AA CA CC
 1  6 18

Nobs <- sum(ObsCount)
FreqC <- (2 * ObsCount[3] + ObsCount[2])/(2*Nobs)

FreqC
 CC
0.84

ExpCount <- c(Nobs*(1-FreqC)^2, 2*Nobs*FreqC*(1-FreqC),Nobs*FreqC^2)
ExpCount
[1] 0.64  6.72  17.64
```

The expected count for one of the cells is less than 5, hence we should use the Fisher's exact test to test the null that there is HWE. Following R code finds out the exact probability of seeing the observed counts.

```
n11 <- ObsCount[3]
n12 <- ObsCount[2]
n22 <- ObsCount[1]
n1 <- 2*n11+n12
Num <- 2^n12 * factorial(Nobs)/prod(factorial(ObsCount))
Denom <- factorial(2*Nobs) / (factorial(n1)*factorial(2*Nobs-n1))
FisherP1 <- Num/Denom
FisherP1
[1] 0.4011216
```

Since Fisher's exact p-value is given by summing over all probabilities of seeing something as extreme as or more extreme than the observed data, we need to perform the above calculation for the more extreme situations as well, given by $n_{11} = 19$, $n_{11} = 20$ and $n_{11} = 21$. But this will be a very long R code. Instead of all this calculation, we can simply use the function "HWE.exact()" to get the required p-value. The R code for this is given below:

```
library(genetics)
Akt1Snp1Maya <- genotype(AKT1.C0756A[Population=="Maya"], sep="")

HWE.exact(Akt1Snp1Maya)
Exact Test for Hardy-Weinberg Equilibrium

data:  Akt1Snp1Maya
N11 = 18, N12 = 6, N22 = 1, N1 = 42, N2 = 8, p-value = 0.4843
```

Conclusion: The p-value is too high to reject the null hypothesis. Hence there seems to be enough evidence to suggest that HWE exists for the given SNP within this population. △

2.2.2 HWE and population substructure

Population substructure and HWE are closely related to each other. Knowing these relations are important to devise strategies which consider these associations and do not give erroneous conclusions. The population substructure and HWE relationships that will be highlighted in this subsection are :

1. HWE is violated in the presence of population admixture.
2. HWE is violated in the presence of population stratification.
3. HWE implies constant allele frequencies over generations.

It has already been shown that allele frequencies remain constant over generations as a consequence of HWE. Using examples we will see the effect of population substructure and admixture on HWE. Recall that population admixture is said to occur when mating occurs between two populations for which allele frequencies differ.

Example 2.8. Assume an admixed population such that $P(A) = p_A$ in one population and $P(A) = q_A (p_A \neq q_A)$ in the other population for a given site. Now suppose that two people , one from each population, reproduce. Note that the mating here is random and hence the alleles the offspring gets from each parent is totally random. So the expected number of each genotype when there is random mating among n people of an admixed population such that every couple has one from each population s given below.

Population 1 / 2	A	a	Total
A	np_Aq_A	$np_A(1-q_A)$	np_A
a	$n(1-p_A)q_A$	$n(1-p_A)(1-q_A)$	$n(1-p_A)$
Total	nq_A	$n(1-q_A)$	n

Table 2.8: Population admixture

For simplicity assume that $n = 200$, $p_A = 0.8$ and $q_A = 0.4$ for the given site. In a scenario where there is random mating, the observed counts will be *similar* to that in Table 2.8. So substituting n, p_A, q_A values in Table 2.8, should give us the approximate observed counts under random mating. Table 2.9a gives those observed counts.

After finding the observed counts, we need to calculate the expected number of counts. To find the expected counts, we need to estimate the allele frequency in the population. So frequency of allele A in population is estimated as

$$p_0 = \frac{2 \cdot 64 + 96 + 16}{2 \cdot 200} = 0.6$$

So the expected counts will be found by substituting p_0 for allele frequency of A and $1 - p_0$ for allele frequency of a. The expected counts for n people is given in Table 2.9b.

Population 1 / 2	A	a	Total
A	64	96	160
a	16	24	40
Total	80	120	200

(a) Allele frequency

Population 1 / 2	A	a	Total
A	Np_0^2	$Np_0(1-p_0)$	Np_0
a	$Np_0(1-p_0)$	$N(1-p_0)^2$	$N(1-p_0)$

(b) Genotype counts under random mating in admixed population

Table 2.9

Note that the observed data consists of $n_1 1 = 64$, $n_{12}^* = 96 + 16 = 112$ and $n_2 2 = 24$. So the corresponding expected counts given by E_{11}, E_{12}^* and E_{22} respectively are given by

$$\begin{aligned} E_{11} &= Np_0^2 = 72 \\ E_{12}^* &= 2Np_0(1-p_0) = 96 \\ E_{22} &= N(1-p_0)^2 = 32 \end{aligned}$$

So the chi-squared statistic for this example is given by

$$\chi^2 = \frac{(72-64)^2}{72} + \frac{(96-112)^2}{96} + \frac{(32-24)^2}{32} = 5.56$$

The critical value corresponding to chi-squared distribution with one degree of freedom is 3.841 at level of significance 0.05. Hence at this level of significance, we have to reject the null hypothesis of HWE. **Conclusion:** There is not enough evidence to support the hypothesis , so this indicates that there exists a departure from HWE when population admixture happens. △

The next phenomenon that will be discussed in brief is population stratification. **Population stratification** is the combination of populations in which breeding occurs within but not between sub-populations. Again we use an example to understand the phenomenon.

Example 2.9. Suppose that there are two populations for which genotype counts are as given in the table below. In one population the allele frequency of A is 0.8 and in the other it is 0.4. Observe in Table 2.10 that each individual population given below is in HWE because the observed counts match with expected counts.

Hence our population has two sub-populations which breed within their own populations but don't breed with someone outside their population. Now consider both the populations combined together, then clearly our combined population is *stratified* into two groups. The combined data looks as follows.

For the above combined data, we will test for the presence of HWE using R code given below.

Population 1 / 2	A	a	Total
A	$200 \cdot (0.8)^2 = 128$	$200 \cdot 0.8 \cdot 0.2 = 32$	160
a	$200 \cdot 0.8 \cdot 0.2 = 32$	$200 \cdot (0.2)^2 = 8$	40
Total	160	40	200

(a) Genotype counts when P(A) = 0.8

Population 1 / 2	A	a	Total
A	$200 \cdot (0.4)^2 = 32$	$200 \cdot 0.4 \cdot 0.6 = 48$	80
a	$200 \cdot 0.4 \cdot 0.6 = 48$	$200 \cdot (0.6)^2 = 72$	120
Total	80	120	200

(b) Genotype counts when P(A) = 0.4

Table 2.10

Population 1 / 2	A	a	Total
A	$128 + 32 = 160$	$32 + 48 = 80$	240
a	$32 + 48 = 80$	$8 + 72 = 80$	160
Total	240	160	$n = 400$

Table 2.11: Population admixture

```
ObsDat <- matrix(c(160,80,80,80),2)

chisq.test(ObsDat,correct=F)
Pearson's Chi-squared test

data:  ObsDat
X-squared = 11.111, df = 1, p-value = 0.0008581
```

The p-value above is less than 0.05, hence we can reject the null hypothesis and say that there exists a departure from HWE in the combined stratified population. In the above code we mentioned "correct=F" which indicates that Yates' continuity correction is not necessary. This correction is needed if one of the expected cell counts is less than 5.

Conclusion: The combined population exhibits departure from HWE which is wrong as each individual population is under HWE.Departure from HWE while considering stratified data while each individual group is in HWE. This is commonly referred to as **Wahlund effect**. △

In the above examples, we started off with admixed or stratified populations and arrived at a conclusion that such populations exhibit a departure from HWE. But this is not usually the case. We do not always know if our population is admixed or stratified. Hence a test for HWE is used to asses whether either population admixture or stratification is present. Note that the phrase "population substructure" encompasses both population admixture and population stratification.

2.2.3 Geographic origin and HWE

It feels intuitive to guess that there might be an association between HWE and geographic origin. People from different geographic regions have different genetic makeup due to environmental factors. Hence we can expect that some of the populations have a HWE at a particular site while others don't. To see this association, we use the HGDP data in the following example.

Example 2.10. HGDP data mentions the geographic region each person of the study belongs

to under the label "geographic.origin". We first use the "table()" function in R to summarize the data at hand on the basis of geographic origin.

```
attach(hgdp)
table(Geographic.area)

Geographic.area
 Central Africa   Central America         China            Israel          Japan
        119              50                184               148             31
    New Guinea   Northern Africa    Northern Europe      Pakistan        Russia
        17               30                16               200             67
 South Africa     South America      Southeast Asia    Southern Europe
        8                58                11               125

library(genetics)
Akt1Snp1 <- genotype(AKT1.C0756A, sep="")
```

Tests of HWE within each region are calculated using "tapply()" and "HWE.chisq()" functions in R. In the code below we extract the results for two regions for AKT1.CO756A SNP.

```
HWEGeoArea <- tapply(Akt1Snp1,INDEX=Geographic.area,HWE.chisq)

HWEGeoArea$"Central Africa"

Pearson's Chi-squared test with simulated p-value (based on 10000 replicates)

data:  tab
X-squared = 0.23224, df = NA, p-value = 0.6669

 HWEGeoArea$"South America"

Pearson's Chi-squared test with simulated p-value (based on 10000 replicates)

data:  tab
X-squared = 27.239, df = NA, p-value = 9.999e-05
```

Conclusion: From the p-values above, we can say that South Americans have good evidence to show a departure from HWE. Hence it suggests the presence of admixture or stratification in the population. On the other hand, we do not have evidence to reject null hypothesis of HWE among people from Central Africa. Hence geographic region and HWE do have an association among each other. △

2.2.4 Genotyping Errors and HWE

The discussion in this subsection is mainly based on [3] and [5]. A **genotyping error** is defined as a deviation between the true underlying genotype and the genotype that is observed through the application of a sequencing approach. The most common statistical approach to identifying genotyping errors in population based studies of unrelated individuals is testing for a departure from HWE at each of the SNPs under investigation.

This method is based on the assumption that in a large, randomly mating population, genotype frequencies should comply with HWE proportions. As we already saw, deviation from HWE can be caused by many factors, (like stratification, admixture etc) one of which is

genotyping errors. In the study discussed in [3], they explored HWE departure across a large multi-ethnic data set and associated HWE departure with different SNP characteristics, in order to find specific characteristics of HWE departure due to genotyping error.

According to the article, HWE-departure was associated with an excess of heterozygotes (GoH d-HWE) in 93 variants, and with a loss of heterozygosity (LoH d-HWE) in 41 variants. According to the study conducted, genotyping error appeared to be specifically associated with GoH d-HWE but not with LoH d-HWE. LoH d-HWE, on the contrary, was associated with real existing biological phenomenon including deletion polymorphisms and population substructure. Hence if a departure from HWE is detected and additional investigation confirms that there is truly a genotyping error, then the entire SNP is typically removed from analysis.

Few Causes and consequences of genotyping error
Genotyping error can be caused due to

1. sample mishandling,
2. errors introduced by genotyping process,
3. inconsistencies within family pedigree, etc.

Following might be a few consequences of genotyping error.

1. False conclusions and reduction of power to map fine loci.
2. It has been demonstrated in a simulation study that genotyping error rates as low as three percent can adversely affect LD measures.
3. This could limit attempts to identify complex disease genes because it has been demonstrated that genotyping errors always decrease power of certain statistical tests for linkage and association.
4. Genotyping errors can increase both type-1 and type-2 errors.

Though this is a very commonly used procedure to detect genotyping errors, this procedure has certain drawbacks of its own. The drawbacks have not been mentioned here.

Bibliography

[1] Altman, N. and Krzywinski, M. (2015). Association, correlation and causation. *Nature Methods*, 12(10):899–900. https://www.nature.com/articles/nmeth.3587.

[2] Andrews, C. A. (2010). The Hardy-Weinberg Principle. *Nature Education Knowledge*, 3(10):Published online. https://www.nature.com/scitable/knowledge/library/the-hardy-weinberg-principle-13235724/.

[3] Chen, B., Cole, J. W., and Grond-Ginsbach, C. (2017). Departure from Hardy Weinberg Equilibrium and Genotyping Error. *Frontiers in Genetics*, 8(167):Published online. https://doi.org/10.3389/fgene.2017.00167.

[4] Dick, D. (2012). *Candidate Gene Studies*. YouTube video posted by NIHOD on 26 Nov 2012. https://youtu.be/vwk25MwGjV4. Last visited on 19 July 2019.

[5] Foulkes, A. S. (2009). *Applied Statistical Genetics with R*. Use R! Springer-Verlag, New York, 1 st edition.

[6] Jones, M., Fosbery, R., Taylor, D., and Gregory, J. (2007). *Biology*. Cambridge University Press, Cambridge, 2nd edition.

[7] Kerr, K. and Thornton, T. A. (2013a). *Allele Frequencies and Hardy-Weinberg Equilibrium*. Lecture slides for Topic 2, Module 8, Summer Institute in Statistical Genetics. https://faculty.washington.edu/tathornt/sisg2013/Kerr/2HWE_Kerr.pdf. Last visited on 15 August 2019.

[8] Kerr, K. and Thornton, T. A. (2013b). *Linkage and Linkage Disequilibrium*. Lecture slides for Topic 3, Module 8, Summer Institute in Statistical Genetics. https://faculty.washington.edu/tathornt/sisg2013/Kerr/3LD_Kerr.pdf. Last visited on 04 July 2019.

[9] Khan, M. A. (2012). *Introduction to different measures of linkage disequilibrium (LD) and their calculation*. Lecture slides. http://pbgworks.org/sites/pbgworks.org/files/measuresoflinkagedisequilibrium-111119214123-phpapp01_0.pdf. Last visited on 04 July 2019.

[10] LaMorte, W. W. and Sullivan, L. (2016). *Confounding and Effect Measure Modification*. MPH online learning modules, Office of Teaching and Digital Learning, Boston University School of Public Health. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713_Confounding-EM/BS704-EP713_Confounding-EM_print.html. Last visited on 04 July 2019.

[11] Lane, D. M. (2010). *Online Statistics Education: A Multimedia Course of Study*. Developed by Rice University (Lead Developer), University of Houston Clear Lake, and Tufts University. http://onlinestatbook.com/2/index.html. Last visited on 19 July 2019.

[12] Lengerich, E. (2018). *STAT 507: Epidemiological Research Methods*. Eberly College of Science, Department of Statistics Online Programs, The Pennsylvania State University. https://newonlinecourses.science.psu.edu/stat507/node/20/. Last visited on 04 July 2019.

[13] Lewontin, R. C. (1988). On measures of gametic disequilibrium. *Genetics*, 120(3):849–852. https://www.genetics.org/content/120/3/849.

[14] McDonald, J. H. (2014). *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, 3 rd edition. The online edition of this textbook is avialable here: http://www.biostathandbook.com/.

[15] Mueller, J. C. (2004). Linkage disequilibrium for different scales and applications. *Briefings in Bioinformatics*, 5(4):355–364. https://doi.org/10.1093/bib/5.4.355.

[16] Pulst, S. M. (1999). Genetic Linkage Analysis. *Archives of neurology*, 56(6):667–672. https://doi.org/10.1001/archneur.56.6.667.