Table of Contents

Chapter 1 Introduction
1.1 About This Manual
1.2 What is SNPAnalyzer?

Chapter 2 Executing SNPAnalyzer
2.1 System Requirement
2.2 Registration/Login Process
2.3 Installation

Chapter 3 SNPAnalyzer Tutorial
3.1 Analysis Modules
   3.1.1 Hardy-Weinberg Equilibrium
   3.1.2 Haplotype Estimation
   3.1.3 Linkage Disequilibrium
   3.1.4 QTL Analysis

Chapter 4 Copyright Statement and Terms of Uses

APPENDIX A

APPENDIX B
Chapter 1 Introduction

1.1 About This Manual
This is the SNPAnalyzer Version 1.0 manual which provides general instruction on SNPAnalyzer and its functions. As the SNPAnalyzer software is under constant development, the manual might also be updated frequently. Please contact your counterpart at ISTECH Inc. for the recent updates.

1.2 What is SNPAnalyzer?
SNPAnalyzer is software that performs four essential statistical analyses of SNPs in a common computational environment. It is composed of three main modules: 1) Data manipulation, 2) Analysis, and 3) Visualization. The data manipulation module is responsible for data input and output and handles genotype, phenotype, and genetic distance data. To ensure user convenience, the data format is simple. The analysis module performs statistical calculations and consists of four sub-components: 1) Hardy-Weinberg Equilibrium, 2) Haplotype Estimation, 3) Linkage Disequilibrium (LD), and 4) Quantitative Trait Locus (QTL) Analysis. The main feature of the Analysis module is multiple implementations of different algorithms and indices for haplotype estimation and for LD analysis. This enables users to compare separate results generated by different algorithms, which helps to avoid biased results acquired by applying a single statistical algorithm. The performance of all implemented algorithms has been validated using experimentally proven data sets. The Visualization module presents most of the analyzed results as figures, rather than as simple text, which aids in the intuitive understanding of complex data. The SNPAnalyzer has been developed using C and C++ and is available at http://www.istech.info/istech/board/login_form.jsp.
Chapter 2 Executing SNPAnalyzer

2.1 System Requirement
The only requirement for executing SNPAnalyzer is that user should have previously installed Microsoft Internet Explorer v5.0 or higher on his or her local pc.

2.2 Registration/Login Process
To freely use SNPAnalyzer, user’s registration and login process is required. [Figure 2.1] shows the login page at http://www.istech.info/istech/board/login_form.jsp. You can get a free access ID by completing registration process. For the convenient usage, users can freely implement SNPAnalyzer by using login ID “guest” with password “123456”, however, the user’s registered information would be a great help for us to further enhance SNPAnalyzer.

![User’s login page](http://www.istech.info/istech/board/login_form.jsp)

[Figure 2.1] User’s login page
[Figure 2.2] shows registration form to be filled out.

[Figure 2.2] User’s registration form

After finishing registration and log-in process, users should see the web page like [Figure 2.3]. Users should first setup the running environment by clicking [Environment Setup] located in the left side of the page before directly executing SNPAnalyzer by clicking [Execute SNPAnalyzer].

[Figure 2.3] Running environment setup notification
Users can see the figure like [Figure 2.4] by executing SNPAnalyzer after finishing setup process correctly.
Chapter 3 SNPAnalzyer Tutorial

3.1 Analysis Modules

3.1.1 Hardy-Weinberg Equilibrium (HWE)

The Hardy-Weinberg Equilibrium test module can estimate the overall distribution of genotypes of each locus by the chi-square statistics. [Figure 3.1] is a screen shot of HWE test.

Execution Step

1. Press [Input Data] button to import individuals’ genotype data.
2. Once individuals’ genotype data is loaded as shown in [Figure 3.1], press [Run] button to implement HWE test (For individuals’ genotype data, refer to Appendix A.1).
3. Observed alleles’ frequencies and statistics like chi square and p-values are displayed.

[Figure 3.1] A screen shot of HWE test module

Contents of Display

1. Raw data – Individual’s Genotypes
2. Frequency data - Observed allele’s frequencies, observed genotype’s frequencies, expected allele’s frequencies, expected genotype’s frequencies
3. Statistical estimate – Chi-square value, degree of freedom, p-value

**Controls**

1. [Input Data] button  
2. [Run] button  

**Analysis Method**

Refer to Appendix B.1

3.1.2. Haplotype Estimation

Because genotype data have no phase information, other cost effective and accuracy confident methods that can computationally estimate haplotypes are required. The SNPAnalyzer is incorporated with several haplotype estimation algorithms. The Haplotype Estimation module is equipped with three different algorithms: 1) Clark’s algorithm, 2) EM-based algorithm, 3) Bayesian-based algorithm (Pseudo Gibbs Sampling). The Clark’s algorithm and the Bayesian-based algorithm can analyze up to 28 SNPs and more than 100 samples. EM-based algorithm can handle up to 25 SNPs and more than 100 samples. [Figure 3.2] and [Figure 3.3] is a screen shot of Haplotype Estimation.

**Execution Step**

1. Press [Input Data] button to import individuals’ genotype data. [Figure 3.2] shows a screen shot after individuals’ genotype data is loaded.

2. Select one of the analysis algorithms, Clark’s algorithm, EM-based algorithm or Gibbs-sampling based algorithm from [Algorithm] list box. If you choose Gibbs-sampling based algorithm, [Set Parameters] button is activated. Click this button and set the sampling count (Note: 10,000 or more is recommended).

3. Press [Run] button to implement haplotype estimation. [Figure 3.3] shows a screen shot of the haplotype estimation result.

4. If you want to save result of haplotype estimation as a text file, press [Save As Text] button.

5. Phylogenetic tree shows just simple relationships between the estimated haplotypes. A haplotype is represented as a circle, of which size is proportionial to the frequency of that haplotype’s frequency. One base difference between haplotype sequence is represented as solid red line and two or more base difference as dotted black line. Press [Scaled View] button to fit the image of
the phylogenetic tree to the display and press [Save Image] button to save this image as an image file.

[Figure 3.2] A screen shot of the imported data

[Figure 3.3] A screenshot of the haplotype estimation result

**Contents of Display**

① Raw data – Individuals’ genotypes
② Observed data - Observed alleles’ frequencies
③ Estimation result – Estimated haplotypes constituting population’s genotypes, histogram of haplotypes’ frequencies, reconstructed haplotypes of each individual, and phylogenetic tree of haplotypes.

**Controls**

① [Input Data] button  
② [Algorithm] list box  
③ [Set Parameters] button  
④ [Run] button

**Analysis Method**
For more detailed information about haplotype estimation algorithm, refer to Appendix B.2.

3.1.3 Linkage Disequilibrium
The alleles of the two loci on a single chromosome are said to be in the condition of Linkage Disequilibrium (LD) when they do not segregate independently during meiosis. The LD is measured by comparing the probability of estimated frequency of each allele when they are presumed to segregate independently with the probability of the observed frequency of allele. The SNP Analyzer employs the EM-based algorithm to estimate the haplotypes and calculates 5 types of LD measure indices (D, D’, |D’|, Δ, Δ²). [Figure 3.4] is a screenshot of Linkage Disequilibrium.

**Execution Step**

① Press [Input Data] button to import individuals’ genotype data.
② Press [Run] button to calculate the LD measures for all the pair-wise combination of loci.
③ To view only the selected loci pair, select desired locus number from [Locus One] and [Locus Two] list box next to [Run] button.
④ Types of LD measures can be selected from [LD Measure] list box in the middle right of the display.
⑤ If you want to save LD or four gamete test image as image file, press [Save Image] or [Save Four Gamete Image] button.
### Contents of Display

1. **Observed data** – Individuals’ genotypes
2. **Haplotype frequency models** – Haplotypes without association and haplotypes with association
3. **Likelihood test results** – Chi-square statistics and p-values of likelihood test
4. **LD indices** – D, D’, |D’|, Δ, Δ²
5. **LD map of SNPs** – Linkage pattern of pair-wise SNPs
6. **Four gamete test** – Blue colour represents that this site genotype can be have all possible haplotypes, and green colour the opposite.

### Controls

1. **[Genotype Data] button**
2. **[Run] button**
3. **[Locus One] list box**
4. **[Locus Two] list box**

---

**Figure 3.4** A screenshot of Linkage Disequilibrium

| Observed Genotypes | Haplotypes (w/t Association) | Haplotypes (w/ Association) | Likelihood Test | LD Indices | 2-D Map of LD |
|--------------------|-----------------------------|-----------------------------|----------------|------------|---------------|
| Genotype | Observed Two Loci Genotypes | Haplotypes (Expected) | Expt Haplotypes (Frequencies) | Estimated Haplotypes (Frequencies) | 2-D Map of LD |
|         |                            | C   |     | C   |     |       |                |
|         |                            |     |     |     |     |       |                |
|         |                            | 0.5407286 | 0.5487894 | 0.3151010 | 1.0000000 |                |
|         |                            | 0.3002243 | 0.3320480 | 0.4551110 |                |
|         |                            | 0.6004327 | 0.6554384 | 0.3445616 |                |

---

The table above shows the observed genotypes and haplotypes, along with the likelihood test results and LD indices. The 2-D Map of LD visualizes the linkage pattern of SNPs.
3.1.4. QTL Analysis

QTL (Quantitative Trait Locus) indicates the locus that is closely related to the expression of a certain trait or phenotype. The SNP Analyzer consists of two main analysis methods of “Single Locus Analysis” and “Multi Loci Analysis” as depicted in [Figure 3.5]. The “Single Locus Analysis” can evaluate the influence of single locus on the expression of an observed trait or phenotype as lod score. The “Multi Loci Analysis” can evaluate the combinatorial effect of two loci on the expression of an observed trait or phenotype as F values.

[Figure 3.5] A screen shot of QTL Analysis main display
**Execution Step: Single Locus Analysis**

1. To implement QTL Analysis, three types of data are required.
2. Import individuals’ genotype data, individuals’ phenotype data, and distance data by separately pressing [Genotype Data] button, [Phenotype Data] button, and [Distance Data] button. [Figure 3.6] shows a screen shot of the imported data.
3. Press [Run] button to implement analysis
4. [Figure 3.7] shows the result of the analysis. Horizontal axis indicates markers in the specific chromosome and vertical axis indicates F-value, the effect of marker on the specific phenotype or trait.
5. If you want to change trait, select desired trait number in the [Trait No] list box.
6. Select desired number in the [Chromosome No] or [Marker No] list box to check other locus’ effect on the selected trait.

![Figure 3.6] A screen shot of imported data
Contents of Display

① Raw data – Individuals’ genotypes, phenotypes, and genetic distances of markers.
② Statistics – Mean, variance, regression parameters, likelihood test values, and old-score.
③ Line plot – The effect of markers on the specific trait represented as lod-score.

Controls

① [Genotype Data] button
② [Phenotype Data] button
③ [Distance Data] button
④ [Run] button
⑤ [Trait No] list box
⑥ [Chromosome No] list box
⑦ [Marker No] list box

Analysis Method

For more detailed information about single locus analysis, refer to Appendix B.4.

Execution Step – Multiple Loci Analysis

① Import individuals’ genotype data, individuals’ phenotype data, and distance data by separately pressing [Genotype Data] button, [Phenotype Data] button, and [Distance Data] button.
② Press [Run] button to implement analysis

③ [Figure 3.8] shows the result of the analysis. In the 2-D map, red color means that there is close relationship between markers’ genotypes and a specific phenotype and blue color the opposite.

④ If you want to change trait, select desired trait number in the [Trait No] list box.

⑤ Select desired number in the [Chromosome No] or [Marker No] list box to check the combination effect of other two loci on the selected trait.

[Figure 3.8] A screen shot of the result of the multiple loci analysis

Contents of Display

① Raw data – Individuals’ genotypes, phenotypes, and genetic distances of markers.

② Statistics – Mean, variance, regression parameters, likelihood test values, and lod-score.

③ Line plot – The effect of markers on the specific trait represented as lod-score.

Controls

① [Genotype Data] button
② [Phenotype Data] button
③ [Distance Data] button
④ [Run] button
⑤ [Trait No] list box
Analysis Method
For more detailed information about multiple loci analysis, refer to Appendix B.4.
Chapter 4 Copyright Statement and Terms of Uses

ATTENTION: PLEASE READ THIS STATEMENT CAREFULLY BEFORE USING THIS SOFTWARE. USING SNPAnalyzer INDICATES THAT YOU ACCEPT THESE TERMS. IF YOU DO NOT ACCEPT THESE TERMS, DO NOT USE SNPAnalyzer.

This demo version of software is licensed, not sold. You may use this Software only as described in this agreement. If you do not agree to the terms of this agreement, do not install the Software.

1. SOFTWARE. The capitalized term "Software" refers to the object code for the computer program known as SNPAnalyzer Demo Version 1.0, any updates, supplemental code or programs provided to you by ISTECH Inc., the user's manual and Help file, any components, any related media and printed materials, and any related "online" or electronic documentation.

2. RESTRICTIONS. You may not reduce the Software to human readable form, reverse engineer, de-compile, disassemble, merge, adapt, or modify the Software, except and only to the extent that such activity is expressly permitted by applicable law notwithstanding this limitation. You may not rent, lease, or lend the Software. You may not use the Software to perform any unauthorized transfer of information, such as copying or transferring a file in violation of a copyright, or for any illegal purpose.

3. SUPPORT SERVICES. ISTECH Inc. may provide you with support services related to the Software. Use of support services is governed by the user's manual, online documentation, and other ISTECH Inc. materials, as they may be modified from time to time. ISTECH Inc. may use any information you provide as part of obtaining support services for its business purposes, including product support and development.

4. TERMINATION. This license terminates if you fail to comply with its terms and conditions. If your license terminates, you must destroy all copies of the Software. The termination of this license does not limit ISTECH Inc.'s other rights it may have by law.

5. INTELLECTUAL PROPERTY. You acknowledge that you have only the limited,
non-exclusive right to use and copy the Software as expressly stated in this license and that ISTECH Inc. retains title to the Software and all other rights not expressly granted. You agree not to remove or modify any copyright, trademark, patent or other proprietary notices that appear, on, in or with the Software. The Software is protected by United States copyright, patent and trademark laws and international treaty provision.

6. NO WARRANTIES AND RELEASE. TO THE EXTENT PERMITTED BY APPLICABLE LAW, THE SOFTWARE AND ANY SUPPORT SERVICES ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, WITHOUT LIMITATION, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NONINFRINGEMENT. IF APPLICABLE LAW REQUIRES A WARRANTY, THE REQUIRED WARRANTY IS LIMITED TO NINETY (90) DAYS FROM YOUR RECEIPT OF A COPY OF THE SOFTWARE. COMPUTER PROGRAMS ARE INHERENTLY COMPLEX, AND THE SOFTWARE MAY NOT BE FREE OF ERRORS. THE SOFTWARE IS PROVIDED WITH ALL FAULTS AND THE ENTIRE RISK AS TO SATISFACTORY QUALITY, PERFORMANCE, ACCURACY AND EFFORT IS WITH YOU.

7. LIMITATION OF LIABILITY. ISTECH INC. IS NOT LIABLE TO YOU FOR ANY CONSEQUENTIAL, SPECIAL, INCIDENTAL, OR INDIRECT DAMAGES OF ANY KIND ARISING OUT OF THE DELIVERY, PERFORMANCE, OR USE OF THE SOFTWARE, EVEN IF ISTECH INC. HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

8. Should you have any questions concerning this license, or if you desire to contact ISTECH Inc. for any reason, please contact ISTECH Inc. by mail at: #506, Woongshin Art Plaza, 847 Janghang2-dong, Ilsan-gu, Goyang-si, Gyeonggi-do, 411-837, Republic of Korea, by telephone at : 82-31-903-1155, or by electronic mail to jino12@istech21.com or yskim@istech21.com , by visiting internet homepage at: http://www.istech21.com.
Appendix A

Input Data
SNPAnalyzer requires three types of input data: 1) Individuals’ genotype data, 2) Individuals’ phenotype data, 3) Genetic distance data. While the QTL module requires all of three data, the HWE, Haplotype Estimation, and Linkage Disequilibrium modules use only genotype data as input.

A.1 Individual’s genotype data
[Figure A.1] shows a sample of individuals’ genotype of 7 people with 5 SNPs. The rows indicate individuals and the columns indicate SNP. In the genotype data, each field should be separated by “TAB” character and missing genotype for the specific SNP should be written as “NN” character.

| No | rs989333 | rs989332 | rs1546669 | rs2879668 | rs2879665 |
|----|----------|----------|-----------|-----------|-----------|
| 1  | AA       | CC       | AA        | GT        | CT        |
| 2  | AA       | CC       | CC        | GG        | CC        |
| 3  | AA       | CC       | CC        | CC        | CC        |
| 4  | AA       | CC       | CC        | GT        | CC        |
| 5  | AA       | CC       | CC        | GT        | CC        |
| 6  | AA       | CC       | CC        | GT        | CC        |
| 7  | AA       | CC       | CC        | GT        | CC        |

[Figure A.1] Individuals’ genotype data

A.2 Individuals’ phenotype data
[Figure A.2] shows a sample of individuals’ phenotype of 7 people for 3 phenotypes. The values of phenotypes should be quantitative such as blood pressure or blood-sugar level. The rows indicate individuals and the columns indicate phenotypes. Each field should be separated by “TAB” character.

| No | Test00 | Test01 | Test02 |
|----|--------|--------|--------|
| 1  | 1 2    | 3 3    | 3 3    |
| 2  | 3 3    | 3 3    | 3 3    |
| 3  | 3 3    | 3 3    | 3 3    |
| 4  | 3 2    | 4 4    | 4 4    |
| 5  | 1 1    | 3 3    | 3 3    |
| 6  | 1 2    | 2 4    | 2 4    |
| 7  | 4 2    | 2 4    | 2 4    |

[Figure A.2] Individuals’ phenotype data
A.3 Genetic distance data

[Figure A.3] shows a sample of genetic distance data of markers located in the specific chromosome. The first line of genetic distance data should always begin with the reserved term “-NUM” that is followed by the total number of chromosomes. From the second line of genetic distance data, information of chromosomes and markers should be represented. “-CH1” indicates the name of the first chromosome that is followed by the number of markers it contains. In [Figure A.3], for example, the first chromosome contains 3 markers, of which names and values of genetic distance are represented between “-CH1” and “-CH2”. Each field should be separated by “TAB” character. (Note: values of genetic distance should be presented as cM scale)

```
-NUM  3
-CH1  3
  M1_1 5.0
  M1_2 13.4
  M1_3 15.0
-CH2  2
  M2_1 3.4
  M2_2 5.6
-CH3  2
  M3_1 3.4
  M3_2 5.6
```

[Figure A.3] Genetic distance data
Appendix B

B.1 Hardy-Weinberg Equilibrium
HWE test estimates the overall distribution of genotypes of each locus by the chi-square statistics. Following is the HWE statistic.

\[ \chi^2 = \sum_i \frac{(o_i - E_i)^2}{E_i} \]

, where \( o_i \) is the observed frequency and \( E_i \) is the expected frequency of the \( i \)th genotype.

B.2 Haplotype Estimation
1. Clark’s algorithm
   1) Rule-based algorithm
   2) Sequential haplotype inference of each individual
   3) If there exists no definitively constructible data, algorithm fails
   4) There may be individuals whose haplotypes cannot be reconstructed.
   5) Reference: Clark, A.G., Inference of haplotypes from PCR-amplified samples of diploid populations, Mol. Biol. Evol., 7:111-122, 1990

[Figure B.1] A brief schematic of Clark’s algorithm
2. EM-based algorithm

1) Maximum likelihood-based algorithm
2) Consists of two steps: E-step and M-step
3) Reference: Excoffier, L. and Slatkin, M., Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population, Mol. Biol. Evol., 12:921-927, 1995.

Following is a brief summary of EM-based algorithm.

\[ E - step: \text{Expectation step} \]
\[ Q(\theta, \theta^i) = E[\log p(\theta | Z, Y) | \theta^i, Y] \]
\[ Y: \text{observed data} \]
\[ Z: \text{unobserved data} \]
\[ \theta: \text{parameter concerned with Y} \]

\[ M - step: \text{Maximization step} \]
\[ \frac{\partial Q(\theta, \theta^i)}{\partial \theta} |_{\theta} = 0 \Rightarrow \text{estimate} \theta^{i+1} \]

B.3 Linkage Disequilibrium

1. Calculation of the LD indices

| Marker | Disease allele | Normal allele | Total |
|--------|----------------|---------------|-------|
| A1     | \(n_{11}\)    | \(n_{12}\)    | \(n_{1}\) |
| A2     | \(n_{21}\)    | \(n_{22}\)    | \(n_{2}\) |
| Total  | \(n_{11}\)    | \(n_{22}\)    | \(n\)  |

**Notation for Estimated(Expected) Haplotype, Marker Allele, and Disease Allele Frequencies in a 2 X 2 Table**

- \(\pi_{11}\) represents the disease allele frequency of marker A1
- \(\pi_{22}\) represents the disease allele frequency of marker A2
- \(\pi_{+1}\) represents the disease allele frequency of disease

**Independent Event:**
\[ P(A_iA_j) = P(A_i)P(A_j|A_i) = P(A_i)P(A_j) \]

\[ D = \pi_{11} - \pi_{1+}\pi_{+1} \]

If marker and allele are independent, \(D = 0\)
\(D \neq 0\) means specific marker has specific allele, i.e. disease allele

[Figure B.2] 2x2 table for calculating LD index, D

22
2. Several indices for LD measuring.

| Symbol | Formula |
|--------|---------|
| $D$   | $(x_i - \bar{x})^2 = x_i^2 - \bar{x}^2$ |
| $\Delta$ | $\frac{x_i - \bar{x}}{x_i + \bar{x}}$ |
| $D'$  | $\frac{\sum x_i}{n} - \bar{x}$ |
| $\delta$ | $\frac{x_i - \bar{x}}{x_i + \bar{x}}$ |
| $\delta'$ | $\frac{\sum x_i}{n} - \bar{x}$ |
| $Q$   | $\frac{\sum x_i}{n} - \bar{x}$ |

- $\Delta$: the most frequently used measure according to Hill and Weir (1994)
- $D'$: common measure introduced by Lewontin (1964)
- $\delta$: used by Ozelius et al. (1992), and Risch et al. (1991, 1995)
- $\delta'$: Under case-control sampling, $\delta \neq D'$
- $d$: recommended when case-control sampling is employed (Kaplan and Weir, 1992)
- $Q$: robust to case-control sampling (Clegg et al., 1976; Nei and Li, 1980; Olson and Wijsman, 1994)

Relationship between Recombination rate and LD

$$(1 - \theta)^2 = \frac{D_n}{D_0}$$

- $\theta$: recombination rate
- $D_n$: LD value after the nth generation
- $D_0$: LD value at the first generation

[Figure B.3] Several LD indices and their characteristics

**B.4 QTL Analysis**

1. Single locus analysis model

Following is a regression model of single locus analysis

- $\Phi_i$: observed value of quantitative phenotype of ith sample
- $g_i$: marker genotype of ith sample (number of A) $\rightarrow AA: 2$, $AB = 1$, $BB = 0$,
  - A is major allele, B is minor allele
- $\Phi_i = a + bg_i + \epsilon$
- $\epsilon \sim N(0, \sigma^2)$
- $u^2 = \frac{1}{n_i} \sum_{i=1}^{n_i} (\Phi_i - (a + bg_i))^2$
- $\Phi_i$: phenotype of ith sample
- $n_i$: total count of sample
- $g_i \in \{0, 1, 2\}$

$z(x, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$

$L_0 = \prod_{i=0}^{n_i} z(\Phi_i - a, u^2)$

$L_1 = \prod_{i=0}^{n_i} z(\Phi_i - (a + bg_i), u^2)$

$Lod\ Score = \log_{10} (L_1 / L_0) = \alpha (\ln L_1 - \ln L_0)$

$\alpha = 1 / \ln 10$
2. Multi loci analysis model
Following is an ANOVA approach of multi loci analysis

\[ \Phi_i : \text{observed value of quantitative phenotype of } i\text{th sample} \]
\[ g_i : \text{marker genotype of } i\text{th sample(number of major allele)} \rightarrow AABB : 4, AaBb : 2, aabb : 0, \]
\[ A, B \text{ are major alleles, } a, b \text{ are minor alleles} \]
\[ S_j = \{i \mid g_i = j, \ j = 0,1,2,3,4\} \]

**Inter Variance:**
\[ MS_A = \sum |S_j|(\mu_j - \mu)^2 \]

**Intra Variance:**
\[ MS_w = \frac{\sum \sum (\Phi_i - \mu_j)^2}{N - 2} \]

**F - ratio:**
\[ F = \frac{MS_A}{MS_w} \]

\[ \mu : \text{global mean} \]
\[ N : \text{total count of sample} \]