Statistical Analysis of Molecular Data Using Software Packages

Lecture Notes on

Statistical analysis of genetic association studies among unrelated individuals

Second edition

Dr. Kuldeep Kumar Tyagi

Department of Animal Genetics & Breeding

College of Veterinary & Animal Sciences

Sardar Vallabhbhai Patel University of Agriculture & Technology

Modipuram, Meerut- 250 110

Published Online:

Second Edition: December, 2021
Total pages: 20

First edition: July, 2021
Total pages: 18

Published by:

Department of Animal Genetics & Breeding
College of Veterinary & Animal Sciences
Sardar Vallabhbhai Patel University of Agriculture and Technology
Meerut- 250 110, Uttar Pradesh, India

Publication No.

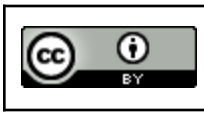
SVP/2021/06/02/249 Dated: December 13th, 2021 *(for official use)*

How to cite this Lecture note

Tyagi, K 2021, *Statistical analysis of genetic association studies among unrelated individuals*, lecture notes series *Statistical Analysis of Molecular Data Using Software Packages*, Training on "Molecular biology tools and its application in Agriculture and Allied Sciences" Sardar Vallabhbhai Patel University of Agriculture & Technology, Meerut, Uttar Pradesh- 250110, India during 01-14 December 2021. Delivered 08th December 2021. Retrieved online from https://vepub.com

Address Correspondence to:

Dr. Kuldeep Kumar Tyagi
Associate Professor & OIC
Department of Animal Genetics & Breeding
COVAS, SVPUAT, Meerut- 250110 (U.P.) India
drtyagivet@gmail.com
+91 9601283365 (M)



ABOUT

This lecture note on "*Statistical analysis of genetic association studies among unrelated individuals*" was prepared under my lecture series on "*Statistical Analysis of Molecular Data Using Software Packages*". This lecture was delivered to trainees attending 14 day training on "Molecular biology tools and its application in Agriculture and Allied Sciences" organized by CST, UP Centre of Excellence in Agriculture Biotechnology in Collaboration with DBT funded Bioinformatics Infrastructure Facility, College of Biotechnology, SVPUAT during 1-14th December 2021. I have tried to explain how molecular data of a population can be analyzed using SPSS and HW_Test software packages. Overall, population association studies in which unrelated individuals for different economic traits are typed at a number of Single Nucleotide Polymorphism (SNP) markers excluding family-based association studies, admixture mapping or linkage studies have been introduced and explained. In addition to the previous first edition Chi Square test for finding relation between various genotypes pertaining to each gene and qualitative trait had been included in this second edition. Some structural reframing of text for better understanding and rectification of some mistakes reported by users have also been undertaken in this second edition. Use of diagrams, explanatory boxes, examples and tables have deliberately been used throughout the notes to create an interest among the trainees. Once through with this lecture note readers will be able to understand the basics and application of statistical analysis of genetic association studies among unrelated individuals using statistical packages. I had tried my level best to simplify the concept in easy to understand language. Further constructive suggestions to improve this lecture notes are always welcome from readers on my email and whatsapp.

[KULDEEP KUMAR TYAGI]

DISCLAIMER

This lecture note on "*Statistical analysis of genetic association studies among unrelated individuals*" under the lecture series on "*Statistical Analysis of Molecular Data Using Software Packages*" has been compiled from various resources available in the public domain. Only excerpts from the original works have been used. This is being done for educational purposes in the interest of developing a concise and updated reading material for students and trainees with no intent of commercial benefits. References to the source of material used have been included in the footnotes. The author does not claim any ownership of any copyrighted material included by chance in the lecture. If due to inability to trace the original source any copyrighted material got included, it may please be brought to the notice of the author for rectification.

[KULDEEP KUMAR TYAGI]

OTHER LECTURE NOTES BY THE AUTHOR

As on: 13-12-2021

Animal Genetics & Breeding					Circulation		
S.No.	Title	Date of Publication	Editions	Pages	Views	Down loads	Access
1	Introduction and importance of statistics and biostatistics	12-10-2020	First	20	53.5k	476	Download
2	Probability	20-08-2020	First	17	47.2k	1.4k	Download
3	Probability	24-11-2020	Second	24	34.7k	439	Download
4	Probability Distribution	07-12-2020	First	33	17.4k	353	Download
5	Introduction of population genetics	23-02-2021	First	16	27.4k	308	Download
6	Genetic structure of population	23-02-2021	First	40	28.7k	451	Download
7	Hardy Weinberg Law	10-07-2021	First	42	2.7k	106	Download

Statistical Analysis of Molecular Data Using Software Packages					Circulation		
S.No.	Title	Date of Publication	Editions	Pages	Views	Down loads	Access
1	Statistical analysis of genetic association studies among unrelated individuals	14-07-2021	First	18	2.0k	118	Download

ACKNOWLEDGEMENT

The author wishes to thank Dr. Pankaj Kumar and the organizing committee of training "Application of Molecular and Bioinformatics Tools in Agriculture and Allied Sciences" for giving me an opportunity to prepare and deliver this lecture to the participants. The help rendered by Dr. Atul Gupta while preparing this manuscript was commendable. This manuscript may not have taken this present shape without his critical review and inputs. I also want to thank Dr. Rajbir Singh, Dean, College of Veterinary and Animal Sciences for his continuous persuasion and support. I am blessed that he considers me befitted for lecture deliveries at various platforms and always routes to me such opportunities related to my expertise. Thanks to my wife, Dr. Surbhi Tyagi; son, Om Tyagi and daughter, Somya Tyagi for their selfless love, affection, support and sparing me from my various household duties. They adjusted and spared me the time from our daily routine for preparation of this lecture note.

[KULDEEP KUMAR TYAGI]

TABLE OF CONTENTS

Statistical analysis of genetic association studies among unrelated individuals

People all around the world have been conducting genetic association studies for a long time. Still the literature to practically understand the statistical analysis of such data using software packages is scarce. Statistical analysis of molecular data is a vast topic. So we will be undertaking one of the important basic and preliminary topics *"Statistical analysis of genetic association studies among unrelated individuals"* in this lecture. Overall, we will be discussing population association studies in which qualitative and quantitative data on unrelated individuals for various economic traits is collected and these individuals are also typed at a number of Single Nucleotide Polymorphism (SNP) markers. This excludes family-based association studies, admixture mapping or linkage studies. Various software packages used in this lecture to analyze molecular data of a population are given below in Table 1.

Table 1 Software packages used in this lecture

Software package	Availability
Microsoft Excel	Paid
IBM Statistical Package for Social Sciences Ver. 20.0	Paid
HW_TEST Software	Open Source

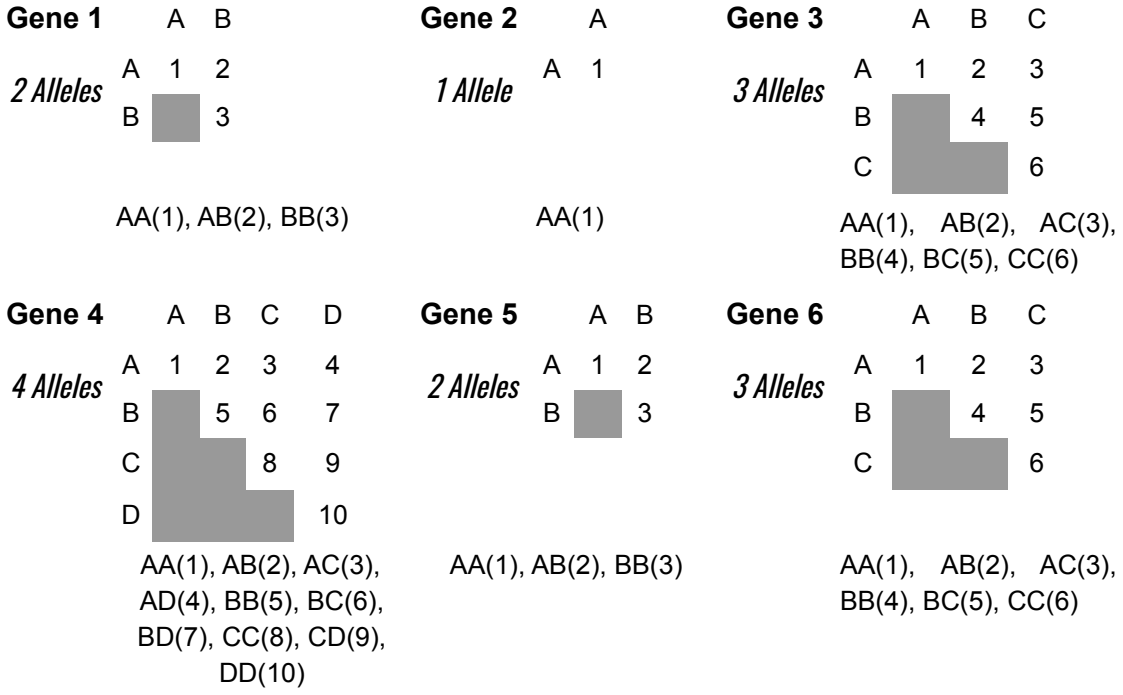
1. Coding the genotypes

To demonstrate practical hands on we will be making use of dummy data. This dummy data is available in Excel workbook and can be downloaded online. This data sheet represents SNP data of six genes (Gene 1, Gene 2, Gene 3, Gene 4, Gene 5 and Gene 6) on 550 lactating buffaloes. The associated quantitative data on traits of economic importance are Age at First Calving (AFC) in days, Service Period (SP) in days and Standard 300 day Lactation yield (SLY) in Kg. The qualitative data on mastitis has also been included with numerals 1, 2 and 3 codes for normal, subclinical and mastitic animals.

1.1. Feeding data in excel workbook

First data can be feeded in Excel workbook for the ease of data feeding. The header row will define the variables. The first column can be used to either assign the serial number or identification number of the buffalo. Columns thereafter constitute the variables like various genes and traits of economic importance. Data pertaining to each individual is assigned in a single row. Genotypes are coded for each gene based on the number of alleles identified as shown in table 2. Say for example serial number 102, the genotypes (coding) were AB(2), AA(1), AB(2), CD(9), AB(2), AB(2) for Gene 1, Gene 2, Gene 3, Gene 4, Gene 5 and Gene 6 respectively. Whereas the values for corresponding traits of economic importance Age at First Calving (AFC) is 1077 days, Service Period (SP) is 121 days and Standard 300 day Lactation yield (SLY) is 2225 Kg respectively. Additionally this animal is not suffering from mastitis hence categorized under normal (1). Similarly, data pertaining to each buffalo is feeded into the excel workbook. Save your excel workbook on your computer in a folder designated for this analysis.

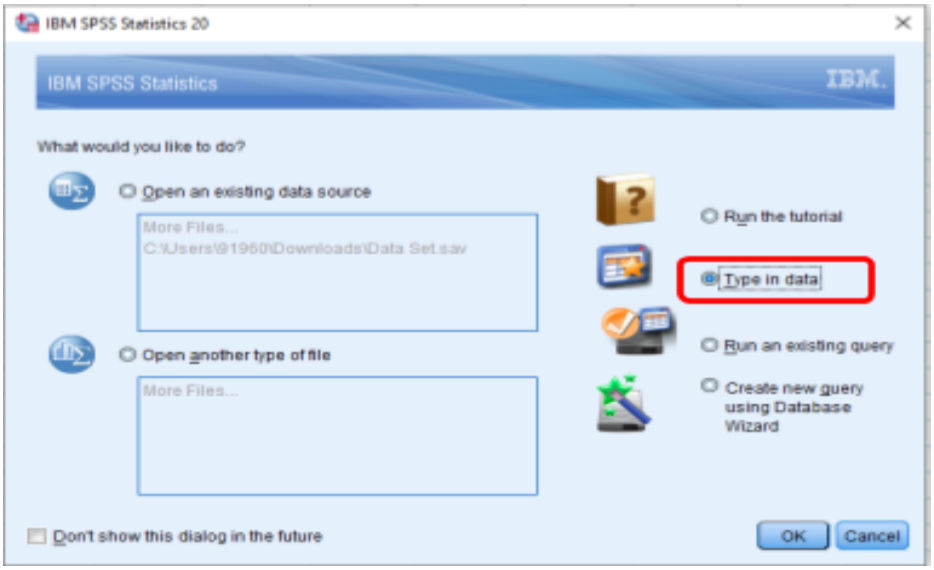
Table 2: Possible genotypes for different genes and their corresponding coding.

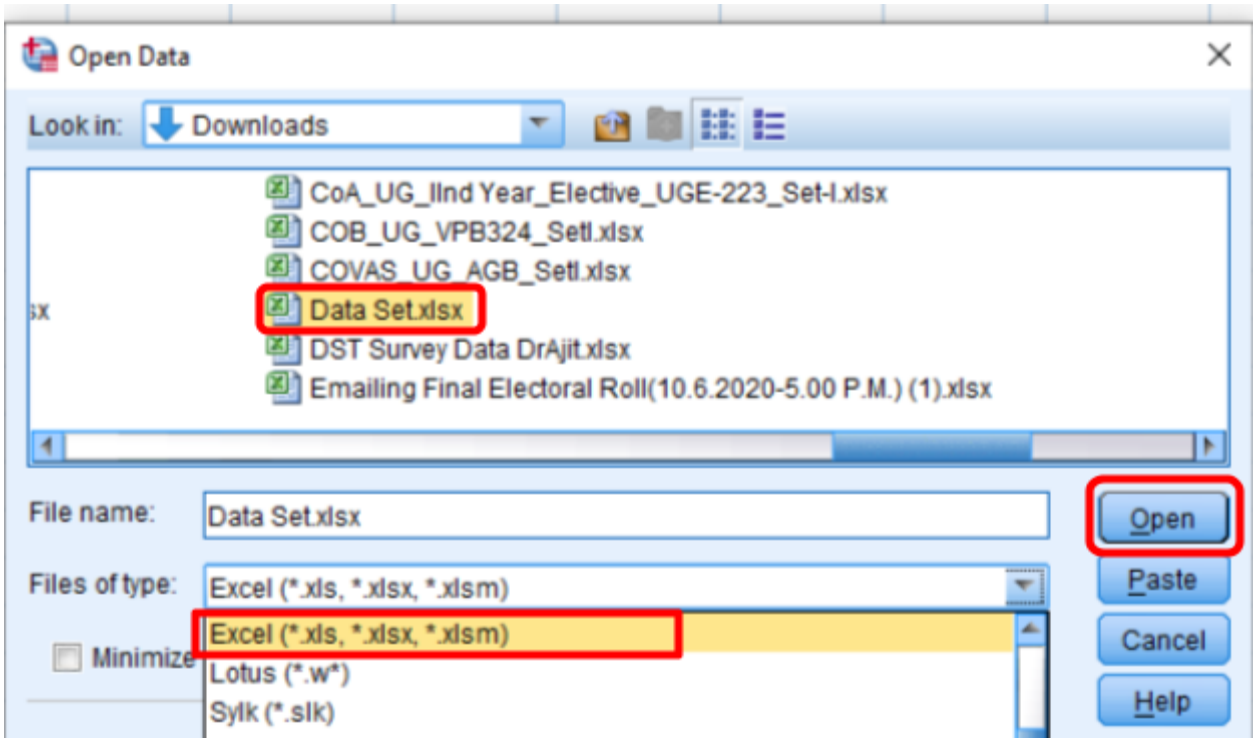


2.SPSS data sheet

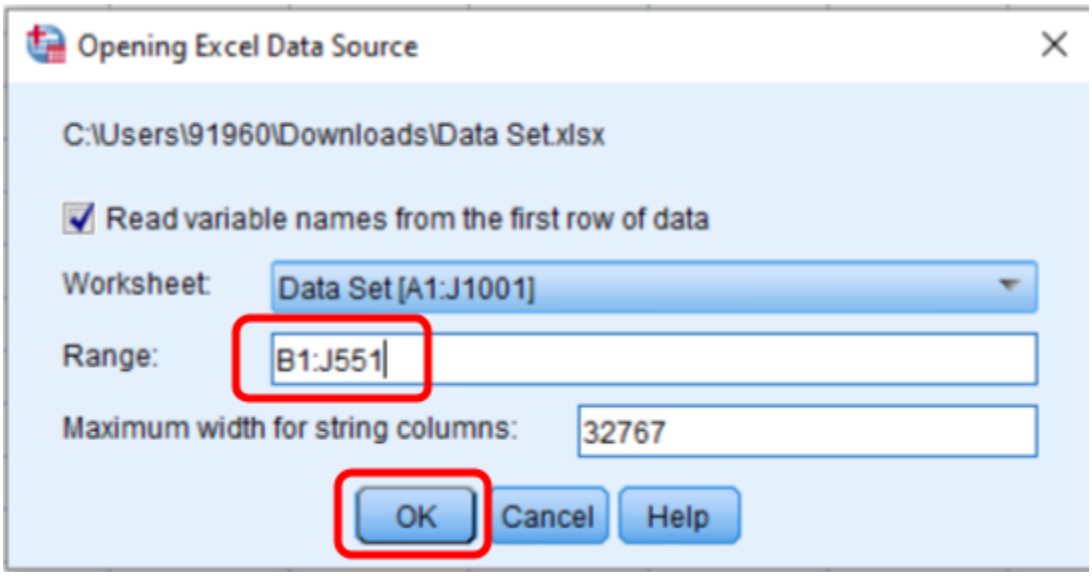
2.2. Importing data to SPSS

We will be importing data to SPSS once our excel workbook is ready. Open SPSS program installed on your computer. A dialog box as shown in the figure below will open.

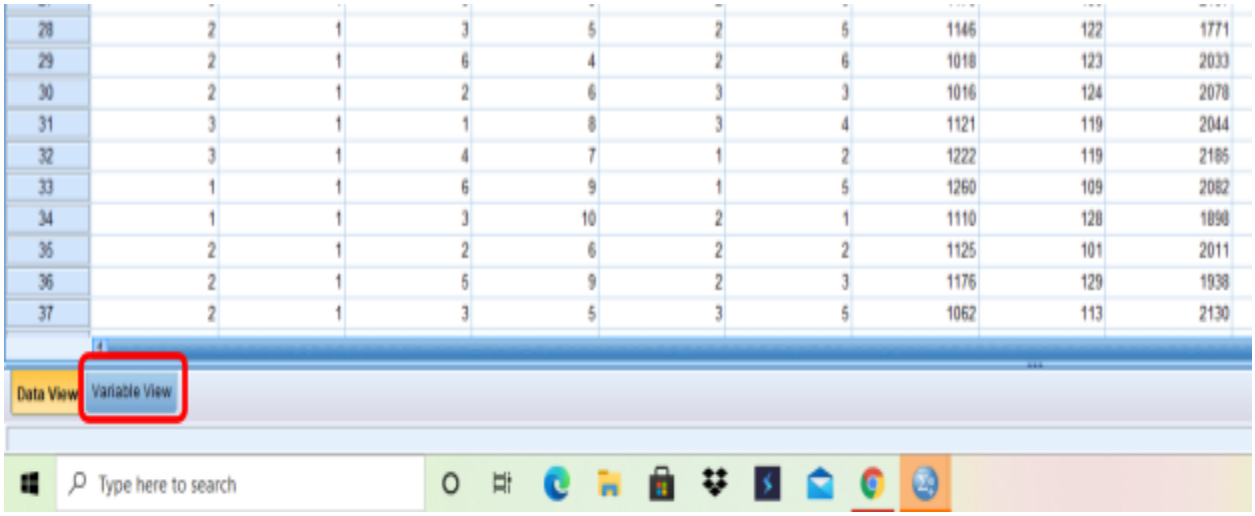




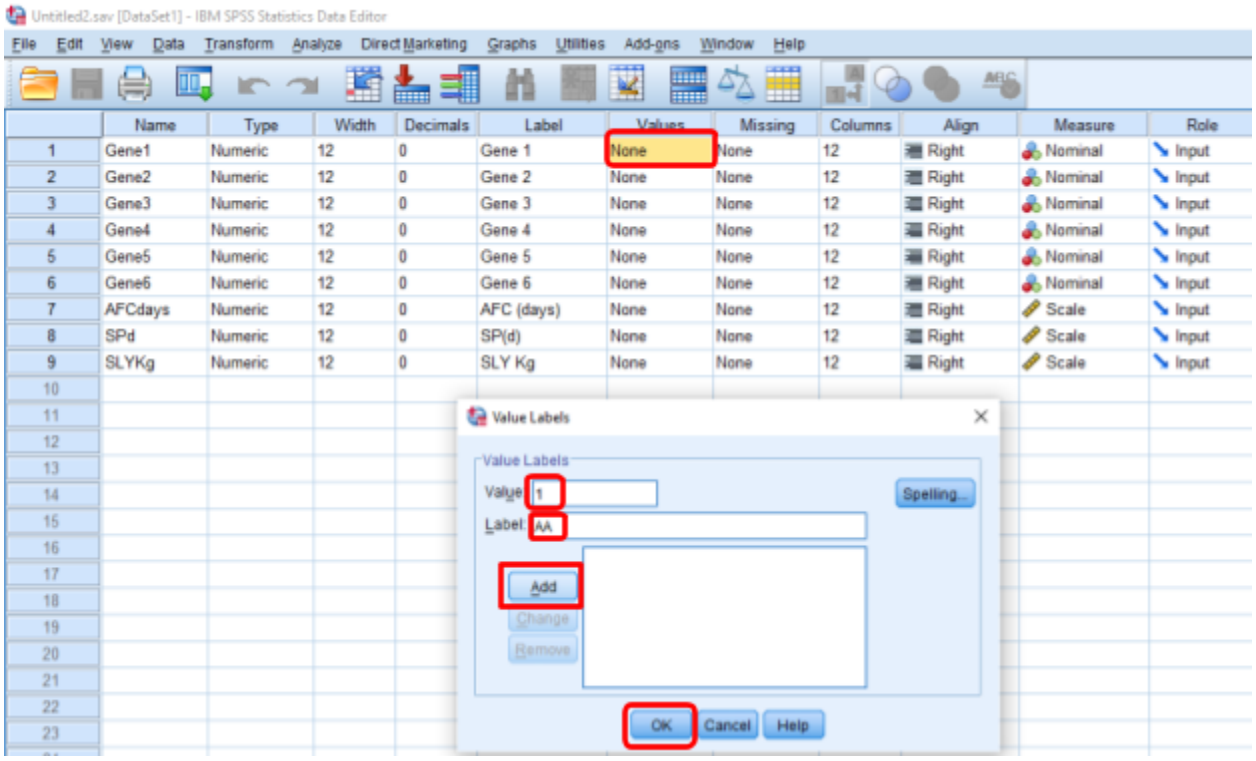
Click on type in data to open the SPSS data view. Go to *File ⇒ Read Text Data*..It will prompt you to locate your data file as shown below.
Select the File type from the drop down menu and then locate your file and click open. This will open another dialog box as shown below.



Define your data range (B1:J551) and then click OK. This will import all your excel data into the SPSS data view. Now we will be defining variables by clicking the variable view tab given at the bottom left of SPSS.



This will open the variable view of the SPSS as shown below.



Click in the cell of the column labeled "Values". This will open a dialog box in which value and their associated genotype (Label) can be added. Click on Ok once you complete feeding all the genotype labels corresponding to a gene. The final dataset file can be downloaded here.

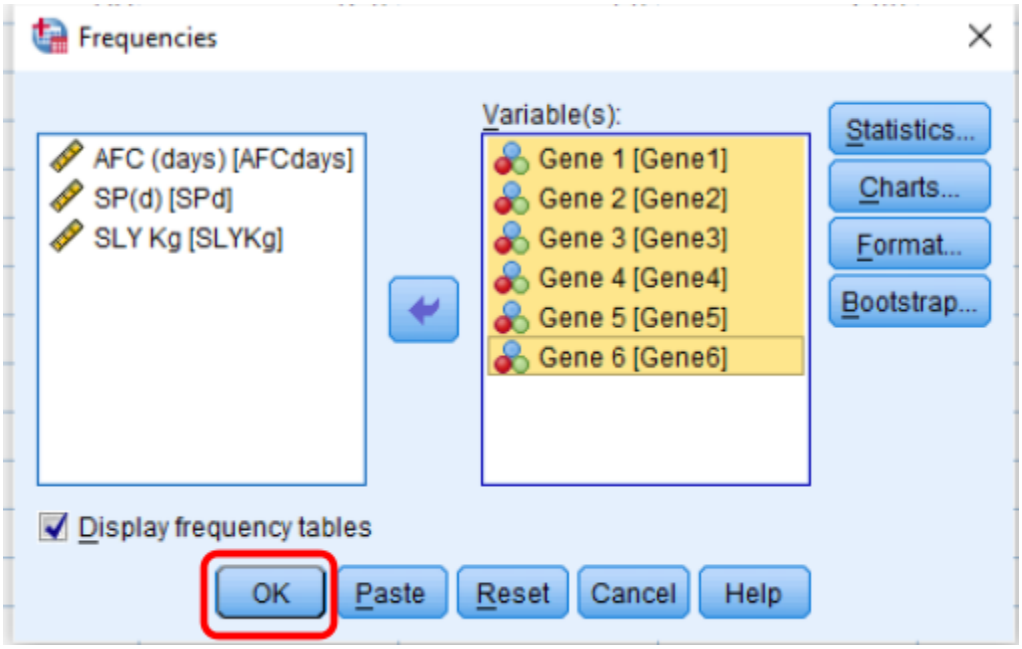
Now each genotype will be visible in the data view using the toggle button provided at the top menu ribbon as shown in the figure below. So basically if you consider gene 1, which has two alleles (A and B) will be having three genotypes AA, AB and BB and their corresponding coding will be 1, 2 and 3 respectively. Similarly, coding patterns for other genes are also followed in a similar fashion. Coding our data in this way will help us to analyse our data with ease using above stated softwares.

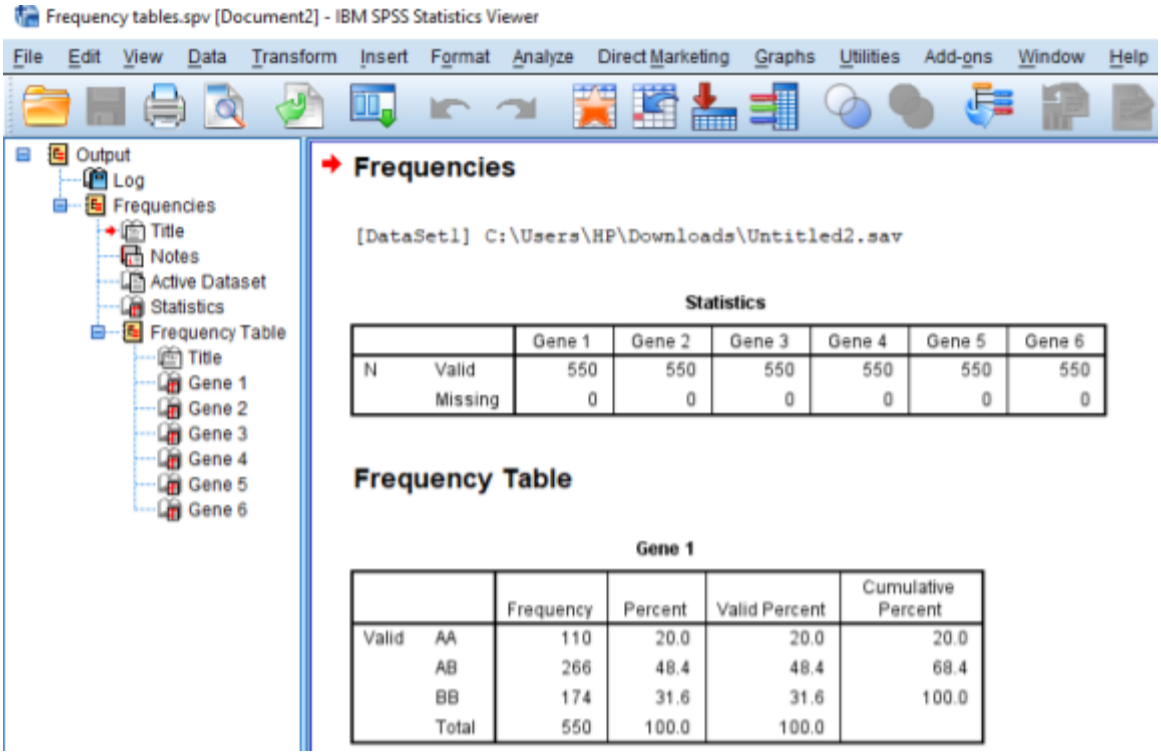
3. Analysis

3.1. Frequency table using SPSS

Go to *Analyze* ⇒ *Descriptive Statistics* ⇒ *Frequencies*. This will open a dialog box, select and put all the variable named genes into the variables box and click OK.

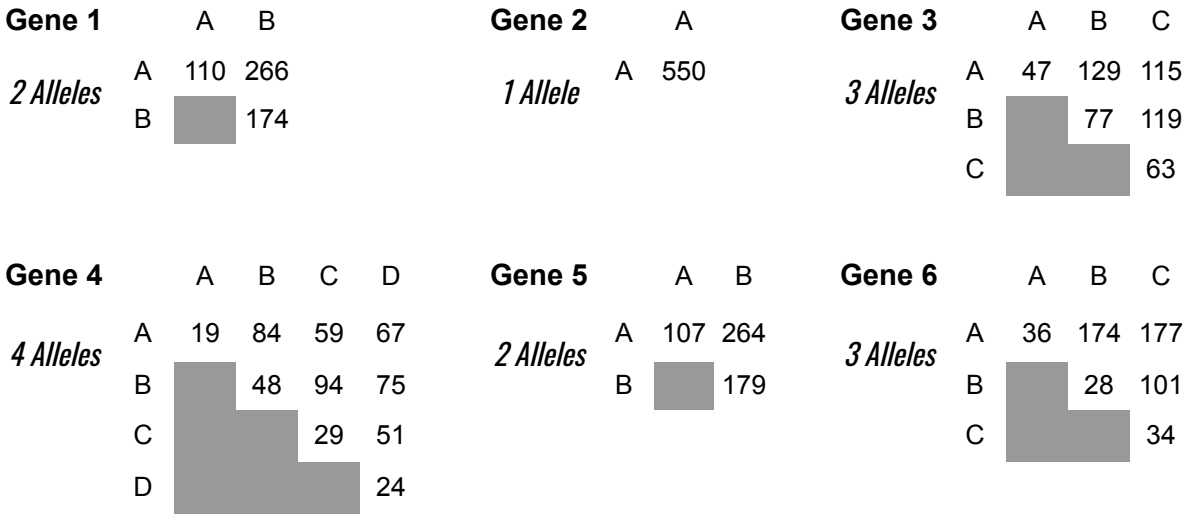
This will give you an output file with frequency of genotypes for all the genes. The file can be [downloaded online]





Above is the screenshot of the output file. These frequencies are being feeded in table 3 and will be used subsequently in Hardy Weinberg equilibrium testing. The benefit of making this table is to have handy absolute frequencies for each genotype ready to be feeded in HW_TEST software. The format prevents mistakes while feeding the value of each genotype because it simulates the data feeding design available in the HW_TEST software.

Table 2: Frequency of various genotypes obtained for different genes.



3.2. Testing for Hardy Weinberg Equilibrium (HWE)

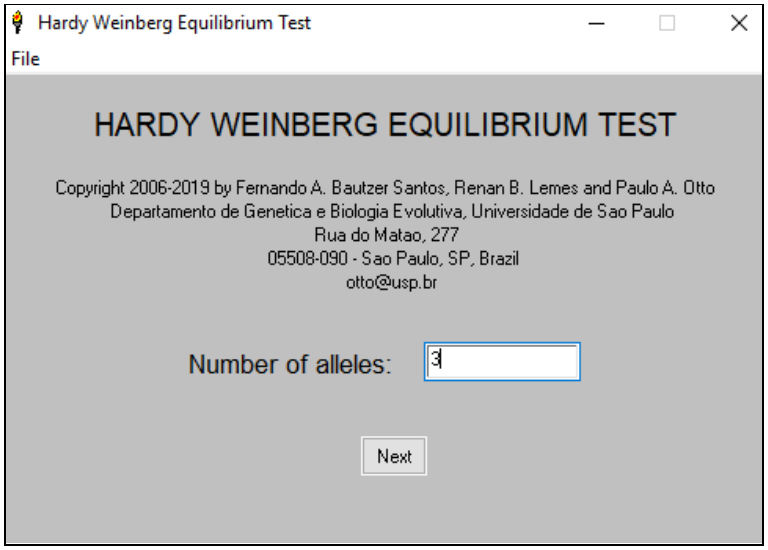
Molecular data on a population are generally subjected to be tested for Hardy Weinberg equilibrium (HWE) as a measure to find preliminary anomalies. HWE test is used extensively on a routine basis to exclude samples with gross molecular typing defects from the usually very large sets of genetic markers presently used in various types of population genetic analyses. Natural or domesticated populations are generally subjected to inbreeding, population stratification, selection or sometimes disease association may lead to deviations from HWE. Sometimes deviation from HWE may be an indication of presence of a common deletion polymorphism, a mutant PCR-primer site or because of a tendency to miscall heterozygotes as homozygotes. So far researchers[1]

[1] Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006 Oct;7(10):781-91. doi: 10.1038/nrg1916. PMID: 16983374.

have tested for HWE primarily as a data quality check and have discarded loci that, for example, deviate from HWE among controls at significance level $\alpha = 10^{-3}$ or 10^{-4}. However, the possibility that a deviation from HWE is due to a deletion polymorphism or a segmental duplication that could be important in disease causation should now be considered before discarding loci.

We will be using user friendly, window based executable open source software HW_TEST for testing HWE[2]. The program can be obtained free of charge directly from the github internet repository as an standalone executable zip file https://github.com/Lemes-RenanB/HardyWeinbergTesting. The downloadable zip file contains a well defined user manual. The Pearson test is easy to compute, but the $\chi 2$ approximation can be poor when there are low genotype counts, and it is better to use a Fisher exact test, which does not rely on the $\chi 2$ approximation.

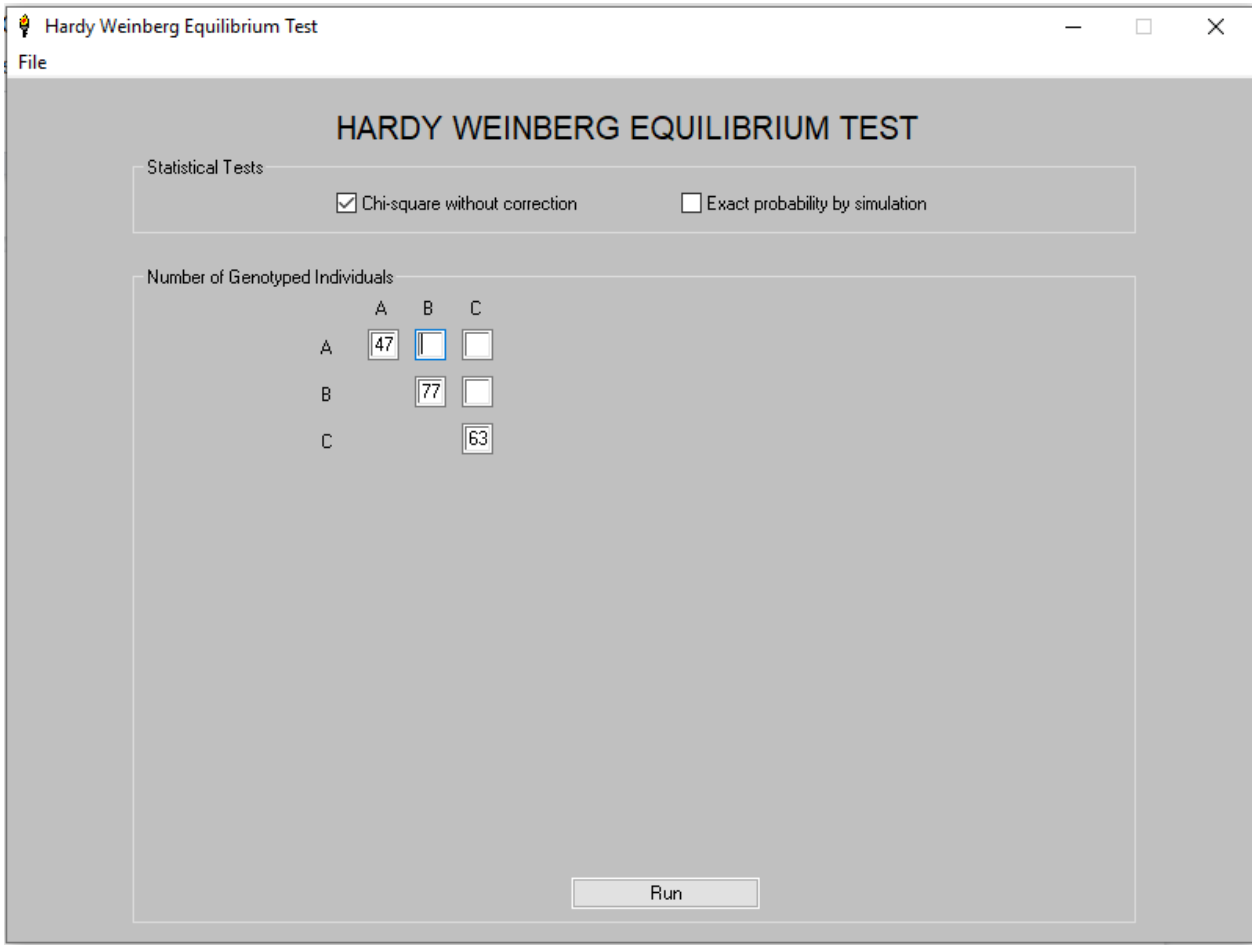
3.2.1 Analysis using HW_TEST software



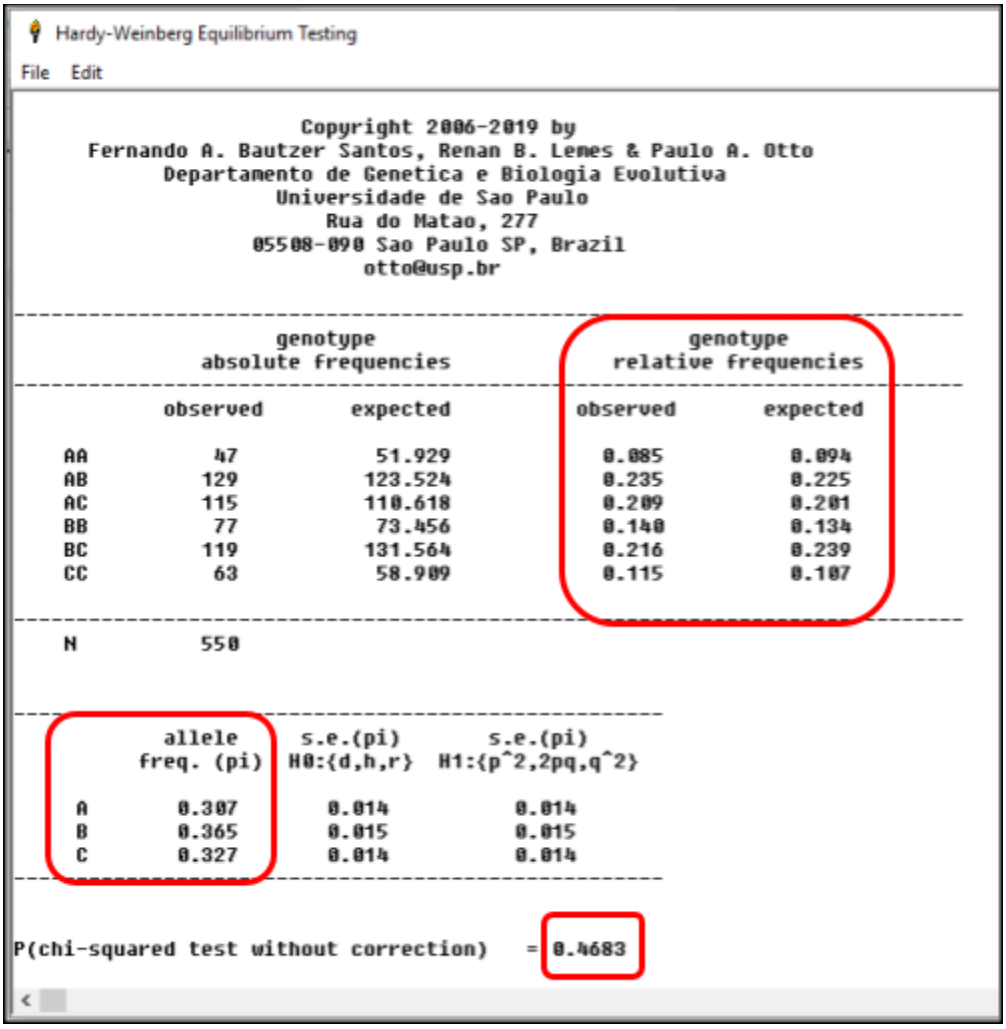
The opening window of software HW_TEST asks for the number of alleles in your data set. Let us consider a case of gene 3 which has three alleles. Put this value into the designated field and press next.

[2] Santos, F., Lemes, R. B., & Otto, P. A. (2020). HW_TEST, a program for comprehensive HARDY-WEINBERG equilibrium testing. Genetics and molecular biology, 43(2), e20190380. https://doi.org/10.1590/1678-4685-GMB-2019-0380

Now refer to the structural representation of gene 3 and their corresponding genotypes given in Table 1. The values corresponding to each genotype need to be filled in designated space provided in HW_TEST software.



Once all the values are entered correctly, click on "Run". This will open the output file with results. The first result output provides us an idea about the observed and expected genotypes frequencies. One can see that the sum of these genotypic frequencies equals unity. Now, the second output provides us with the estimates of allelic frequencies. Finally the P value for χ^2 estimate is provided. If the value of "P" is greater than 0.05 indicates that our null hypothesis of no difference between observed and expected genotypic frequencies is accepted. Therefore, for the given gene the population is in Hardy Weinberg equilibrium. We can assume for the given gene evolutionary forces like mutation, migration and selection are not operating in the given population at the time of screening.

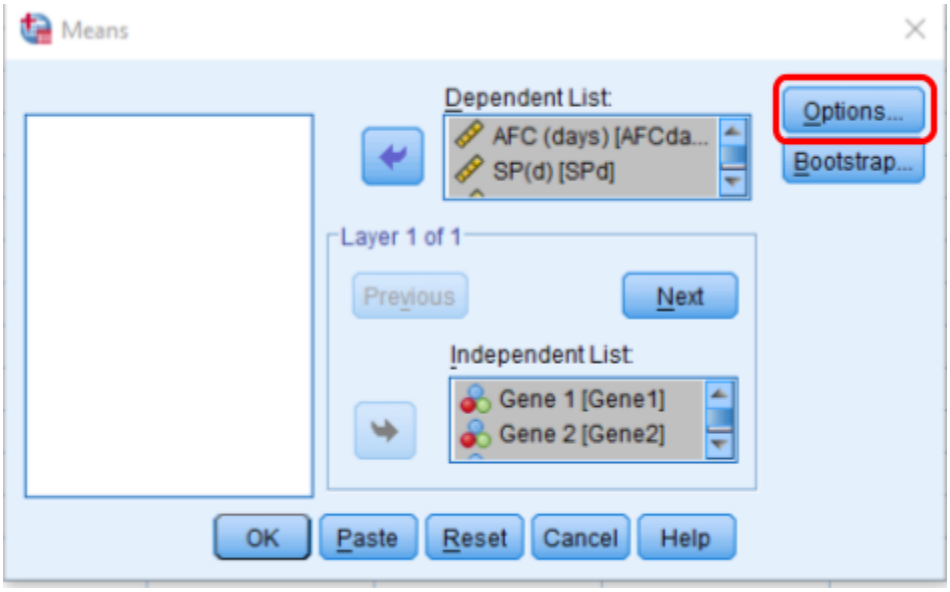


3.3 Analyzing association of genes with quantitative traits

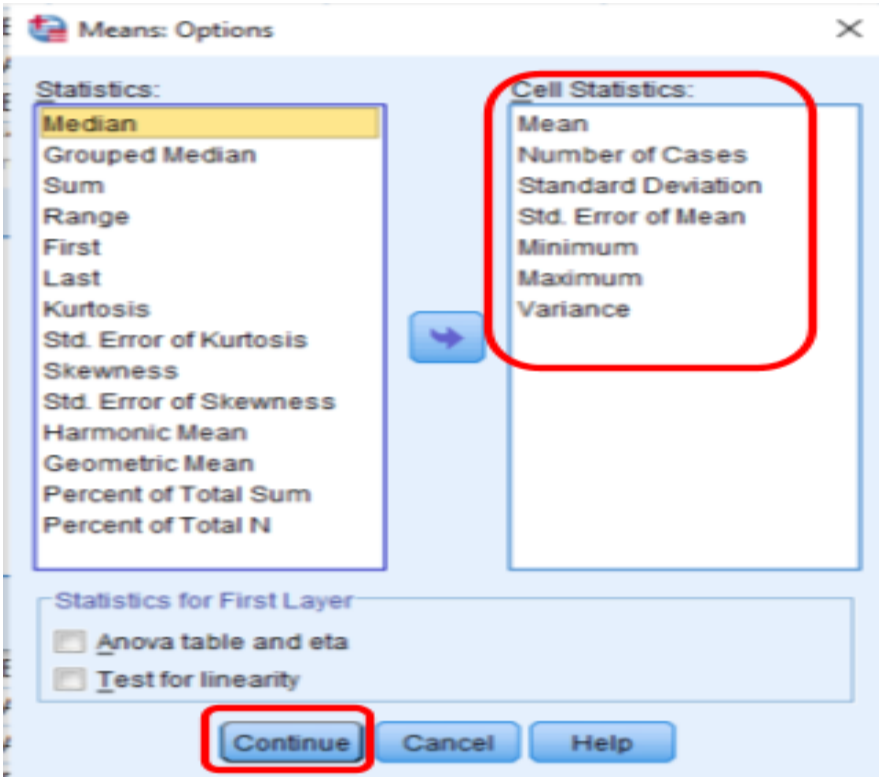
Now first we will describe how to analyze association of genes with quantitative data on traits of economic importance like Age at First Calving (AFC) in days, Service Period (SP) in days and Standard 300 day Lactation yield (SLY) in Kg.

3.3.1 Descriptive Statistics using SPSS

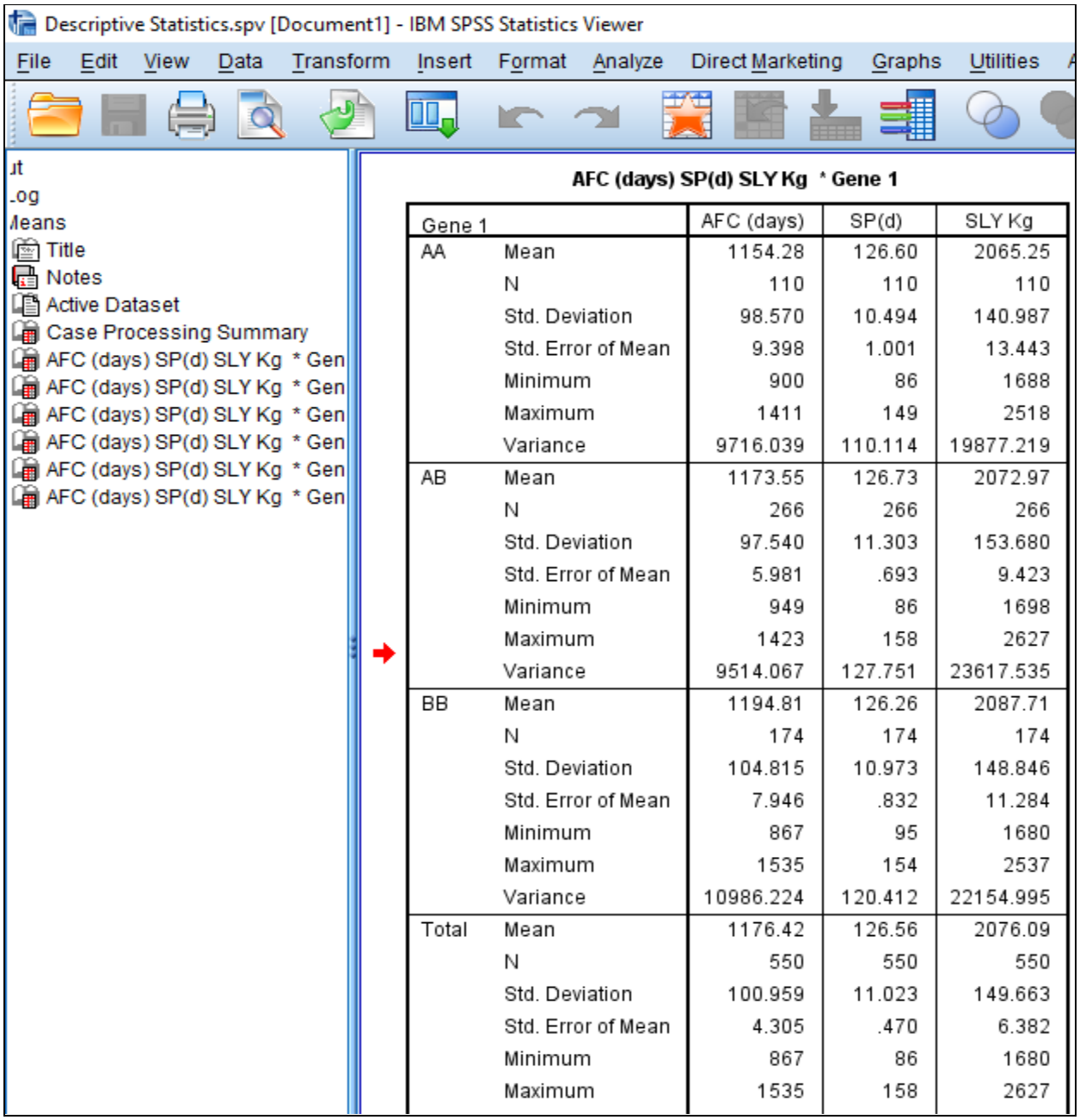
To obtain the descriptive statistics we click on Analyze followed by comparing means and then means ie. $Analyze \Rightarrow Comparing\ Means \Rightarrow Means$. This will open a dialog box. In the dialog box put all the traits of economic importance into the dependent variable box and all the genes into the independent list as shown in the figure given below.



Now this will open another dialog box in which various parameters for descriptive statistics can be defined.



Click on continue to save the changes and then Ok to obtain the results of descriptive statistics. The output file can be downloaded here for your ready reference to match your results when you practice this exercise.



AFC (days) SP(d) SLY Kg * Gene 1

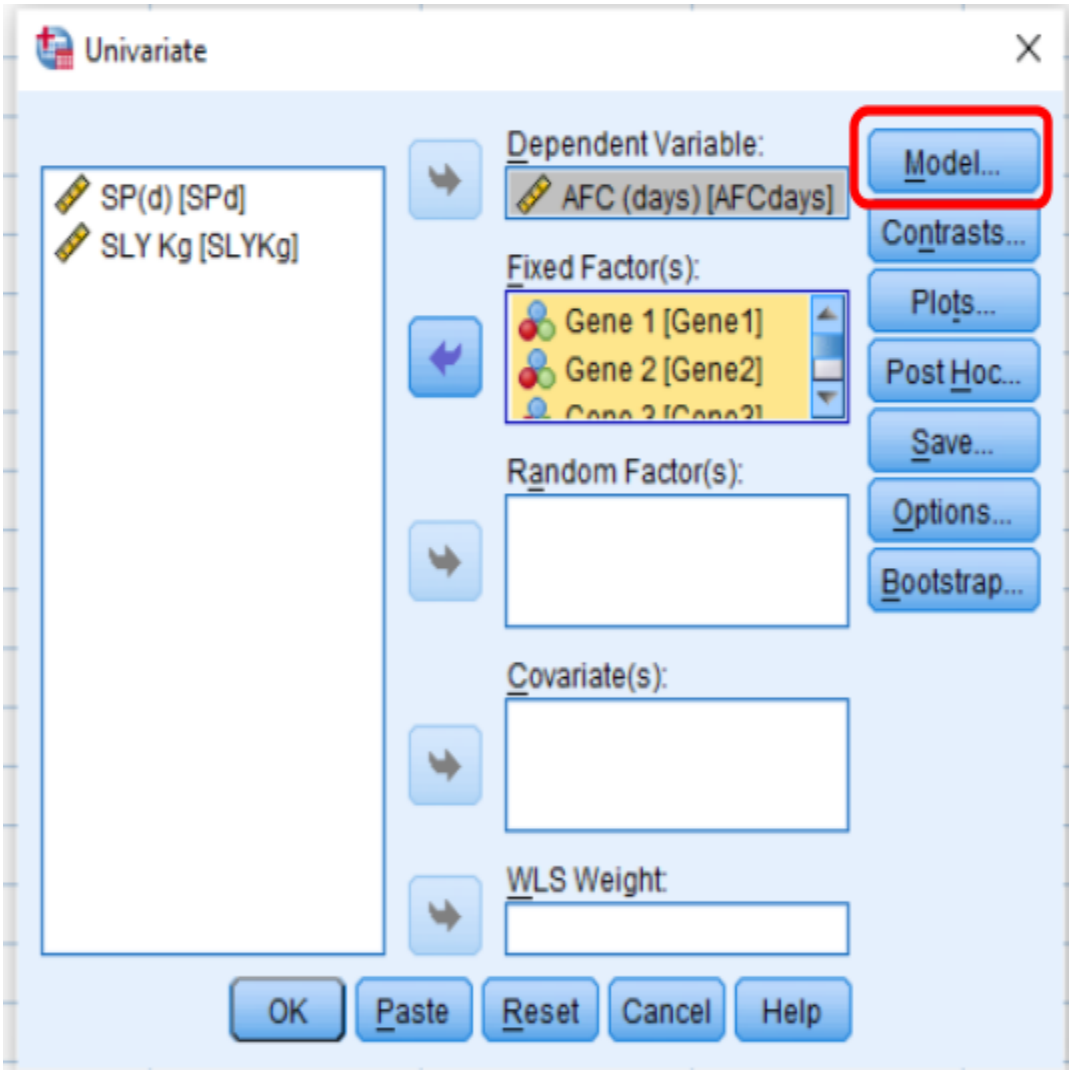
Gene 1		AFC (days)	SP(d)	SLY Kg
AA	Mean	1154.28	126.60	2065.25
	N	110	110	110
	Std. Deviation	98.570	10.494	140.987
	Std. Error of Mean	9.398	1.001	13.443
	Minimum	900	86	1688
	Maximum	1411	149	2518
	Variance	9716.039	110.114	19877.219
AB	Mean	1173.55	126.73	2072.97
	N	266	266	266
	Std. Deviation	97.540	11.303	153.680
	Std. Error of Mean	5.981	.693	9.423
	Minimum	949	86	1698
	Maximum	1423	158	2627
	Variance	9514.067	127.751	23617.535
BB	Mean	1194.81	126.26	2087.71
	N	174	174	174
	Std. Deviation	104.815	10.973	148.846
	Std. Error of Mean	7.946	.832	11.284
	Minimum	867	95	1680
	Maximum	1535	154	2537
	Variance	10986.224	120.412	22154.995
Total	Mean	1176.42	126.56	2076.09
	N	550	550	550
	Std. Deviation	100.959	11.023	149.663
	Std. Error of Mean	4.305	.470	6.382
	Minimum	867	86	1680
	Maximum	1535	158	2627

The output provides, mean, number of observations, standard deviation, standard error of mean, minimum, maximum and variance pertaining to each genotype.

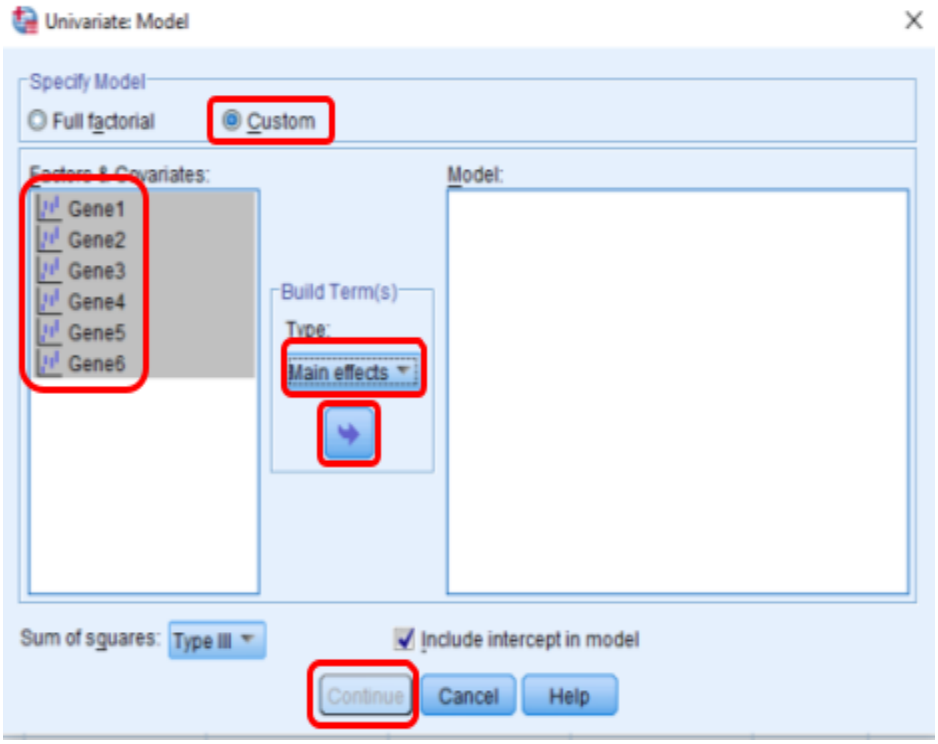
3.3.2 Analysis of Variance (ANOVA) and Duncan test using SPSS

Now we will be subjecting our data for finding any association between various genes studied and traits of economic importance available with us for each individual.

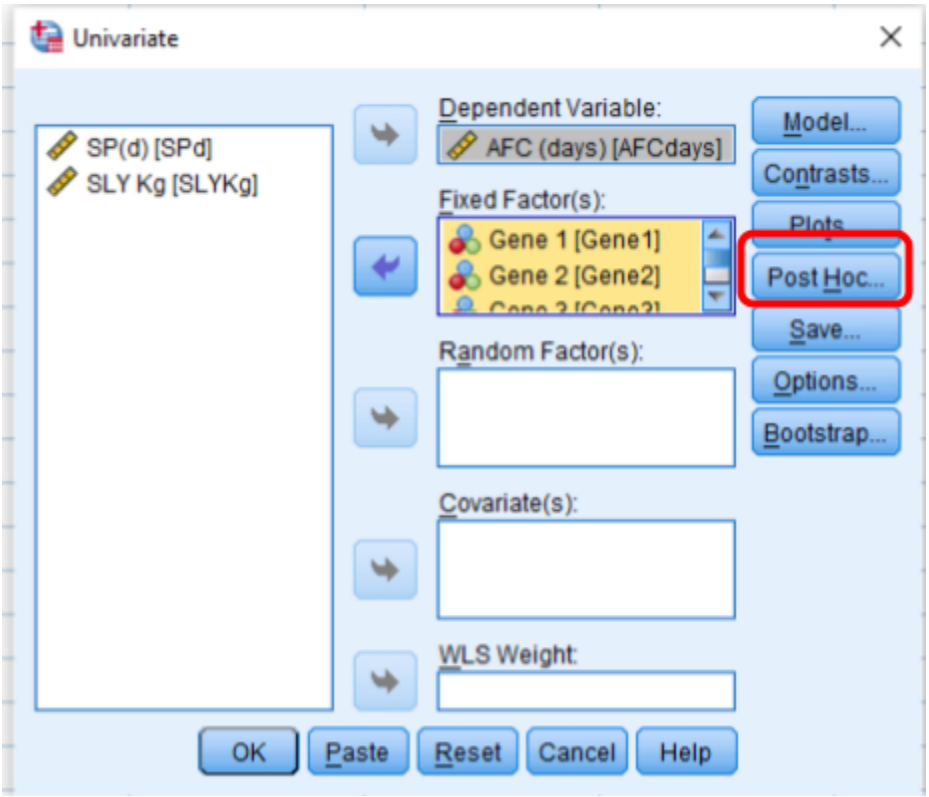
Go to *Analyze ⇒ General Linear Model ⇒ Univariate*. This will open another dialog box. Put one quantitative trait (say AFC) into the dependent variable box and all the genes in the Fixed factor box as shown below. Then click on the model which will open another dialog box to define our model for ANOVA.



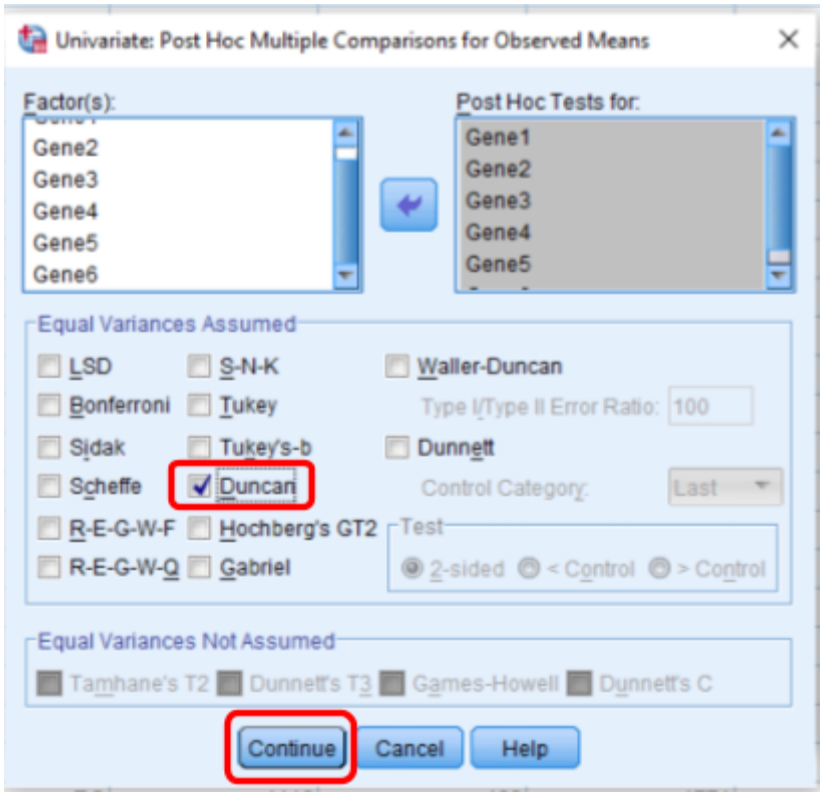
Click on the custom radio button, select all the fixed factors and choose main effects type from the drop down menu of build terms. Click on the arrow to put all fixed factors into the model box and to activate the continue button. Click on continue to save your changes.



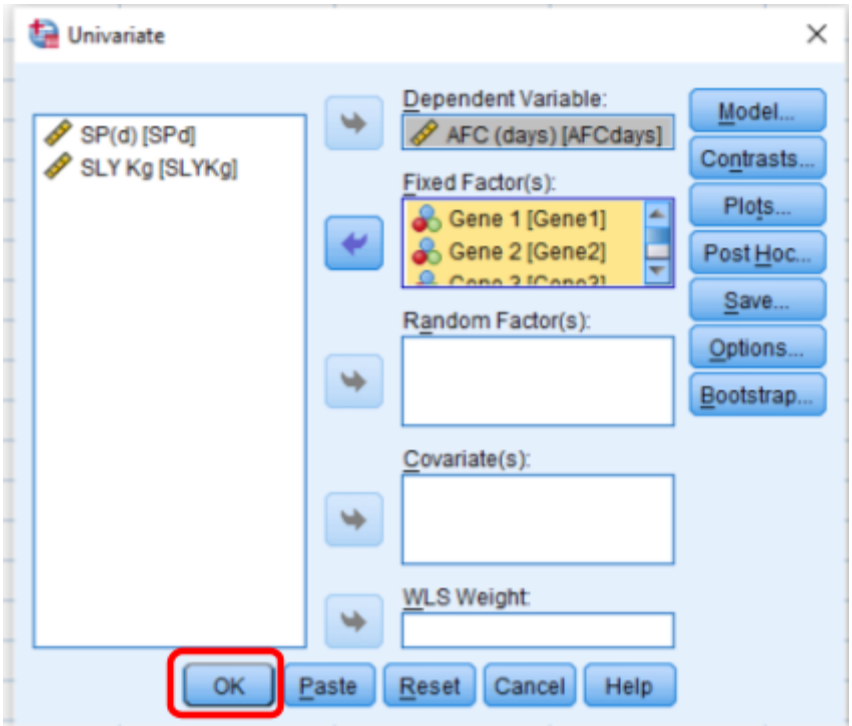
Now click on the Post Hoc button to carry out post hoc tests in your analysis.



This will open another dialog box as shown below.

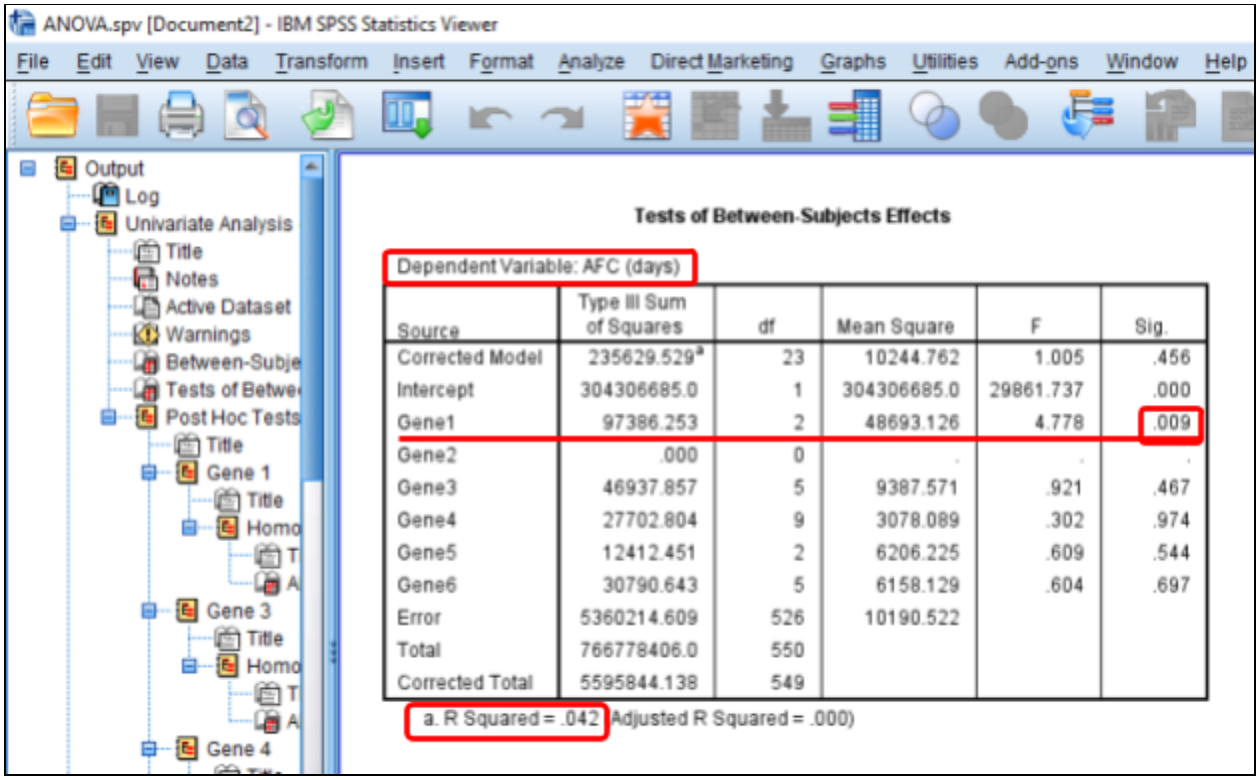


Put all the fixed factors into the post hoc box, select Duncan checkbox and click on continue to save your changes. Now click on Ok to obtain the results.



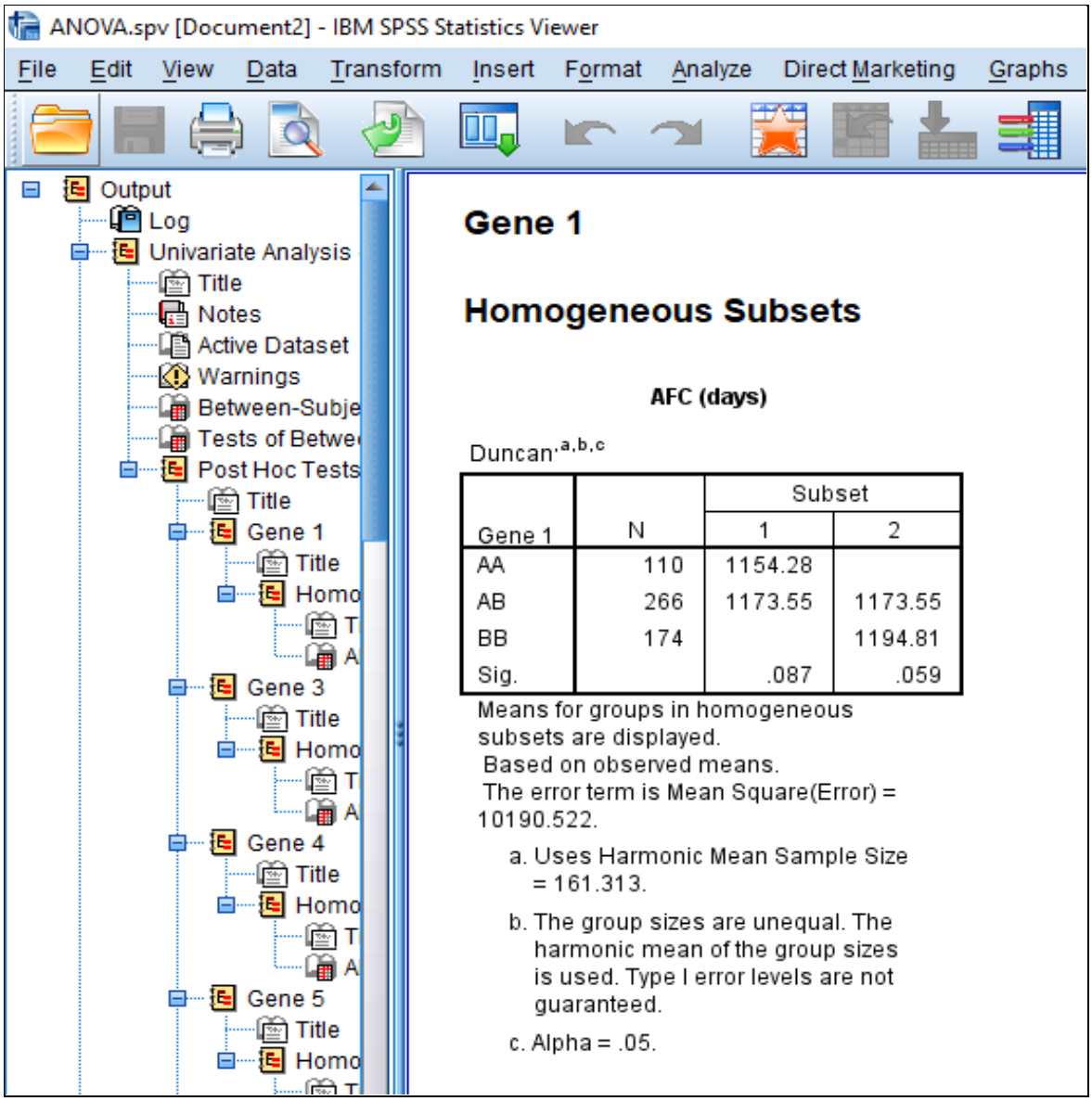
This will open the output window with desired results. Similarly other two traits of economic importance will be put in the dependent variable box one by one and click ok to obtain the results. The output file can be downloaded here for your ready reference to match your results when you practice this exercise.

The first table we need to lookup in the output file is "Test of between subject effect". This is the ANOVA table.



The table shown above is for the dependent variable AFC. The first column shows the sources of variation which in our case are different genes. Now in this table we have to look for the source of variation which is significantly affecting our dependent variable. The gene 1 in this case is contributing a significant effect with p = 0.009 which is less than 0.05. No other gene is causing a significant effect on our dependent variable AFC. The R squared value of 0.042 represents that around 4.2 % variation in our dependent variable is explained by our model. Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

Since only gene 1 is contributing significant effect on dependent variable the post-hoc tests pertaining to this gene will only be considered. The output results of post-hoc Duncan test are summarized under the heading of "Homogeneous Subsets".

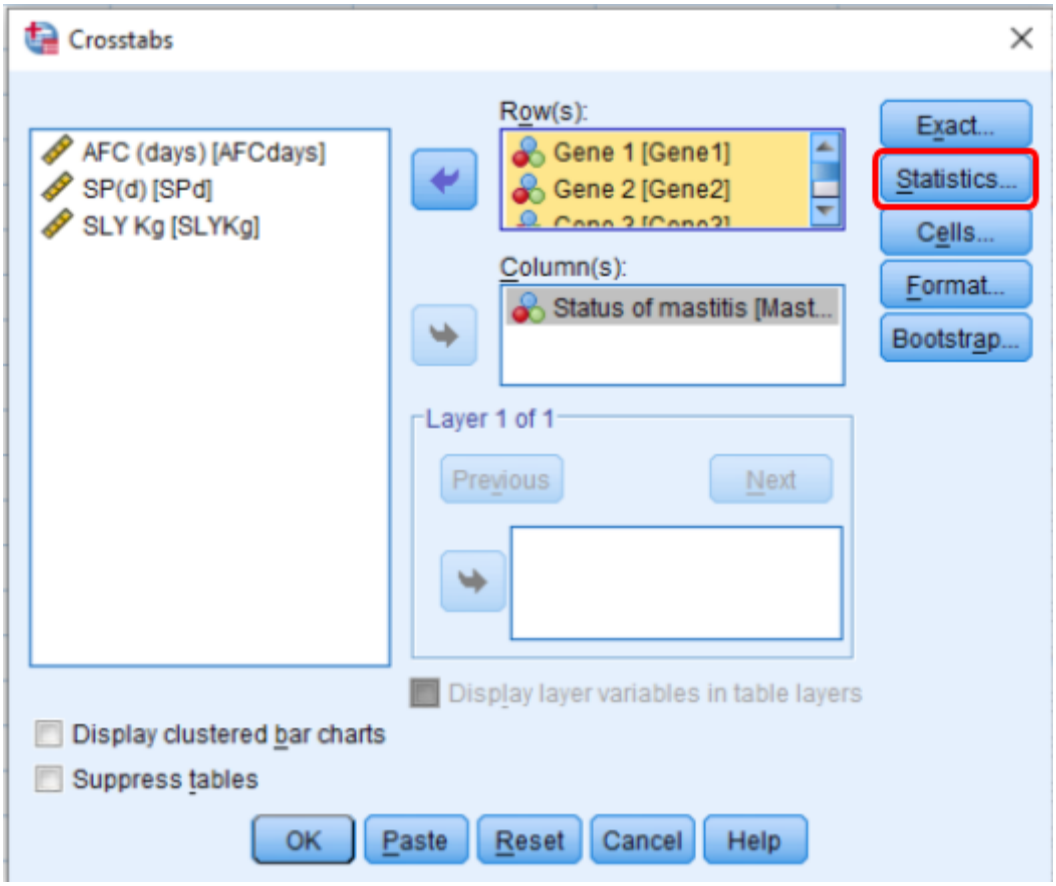


The output table for Duncan test is quite helpful in determining the superscripts of means that differed significantly ($\alpha = 0.05$). Thus the mean AFC (days) for genotype AA, AB and BB will bear superscripts a, ab and b respectively. Similarly results for all the genes can be reported in similar fashion. This enables us to analyze molecular data for testing Hardy Weinberg equilibrium along the identification of genes affecting dependent variables significantly. The method also gives us an idea about the percent variation explained by our predictor variables in the model.

3.4 Analyzing association of genes with qualitative trait

The animals in the present example have been characterized qualitatively as normal, subclinical and mastitic. The possible significant association of various genes with this qualitative attribute can be analyzed using χ^2 Chi square test.

Go to $Analyze \Rightarrow DescriptiveStatistics \Rightarrow Crosstabs$. This will open another dialog box. Put the qualitative trait (say status of mastitis) into the columns box and all the genes in the rows box as shown below.



Then click on the statistics which will open another dialog box to mark tick on χ^2 Chi square test. Click on continue to save your changes. Now click on the OK button to obtain the results. This will open the output window with desired results. The output

file can be downloaded here for your ready reference to match your results when you practice this exercise.

The first table we need to lookup in the output file is Crosstab tables obtained for various genes. These tables depict significant cause-effect relationship between various genotypes of a gene and status of mastitis in concerned animals. If we look at Pearson Chi Square test for gene 6 the P= 0.210, which is greater than the cut off value of 0.05 and hence considered to be non significant. At the same time when we look at Pearson Chi Square test for gene 4 the P=0.026, which is lesser than 0.05 and hence the effect is considered to be significant. Therefore we can say that various genotypes corresponding to gene 4 have significant effects on the status of mastitis in our case study. The results for remaining genes can also be interpreted in the same way.

Gene 6 * Status of mastitis

Crosstab

Count

		Status of mastitis			Total
		Normal	Sub Clinical	Clinical	
Gene 6	AA	8	23	5	36
	AB	43	106	25	174
	AC	46	106	25	177
	BB	3	25	0	28
	BC	24	69	8	101
	CC	10	19	5	34
Total		134	348	68	550

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	13.251[a]	10	.210
Likelihood Ratio	17.108	10	.072
Linear-by-Linear Association	.667	1	.414
N of Valid Cases	550		

a. 3 cells (16.7%) have expected count less than 5. The minimum expected count is 3.46.

Gene 4 * Status of mastitis

Crosstab

Count

		Status of mastitis			Total
		Normal	Sub Clinical	Clinical	
Gene 4	AA	4	12	3	19
	AB	25	47	12	84
	AC	11	39	9	59
	AD	10	50	7	67
	BB	10	28	10	48
	BC	21	62	11	94
	BD	29	37	9	75
	CC	10	18	1	29
	CD	13	33	5	51
	DD	1	22	1	24
Total		134	348	68	550

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	31.419[a]	18	.026
Likelihood Ratio	33.358	18	.015
Linear-by-Linear Association	1.466	1	.226
N of Valid Cases	550		

a. 4 cells (13.3%) have expected count less than 5. The minimum expected count is 2.35.

4. Summary

This lecture on “Statistical analysis of genetic association studies among unrelated individuals” teached us one of the methods of coding our genotype data. Thereafter data is feeded as each column representing a variable and each row representing observations made on an individual. It has been observed that HW_TEST software is user friendly and proved to be an indispensable tool for testing a population for Hardy Weinberg equilibrium. This software also provides us the estimates of genotypic and allelic frequency in a matter of a few seconds. Further usage of SPSS software has been discussed for finding absolute frequencies of various genotypes pertaining to each gene. SPSS also proved to be user friendly menu driven software to obtain descriptive statistics and ANOVA with post - hoc tests for dependent quantitative variables pertaining to various genotypes. Final interpretation of ANOVA helped us to determine genes causing significant variation in dependent variables. Similarly, SPSS had also been used to find possible effects of various genotypes pertaining to each gene on qualitative variables using crosstab functionality to obtain Chi Square results. Thus, the conjunction of HW_TEST and SPSS softwares can successfully be used in statistical analysis of genetic association studies among unrelated individuals for various qualitative and quantitative variables.

****

ABOUT THE AUTHOR

Dr. Kuldeep Kumar Tyagi had completed his B.V.Sc & A.H. in the year 2006 from Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana, Punjab India. He got admission in a master program in the subject of Animal Genetics and Breeding at Indian Veterinary Research Institute, Bareilly, Uttar Pradesh, India after securing 6th rank in All India ICAR-JRF examination. He had completed his Masters in the year 2008 and carried out research on competent fibroblast cells used in somatic cell nuclear transfer. He qualified CSIR Net in his first attempt during the final semester of the masters program itself. He got selected as Assistant Professor in the year 2009 at College of Veterinary Science & A.H. at Navsari Agricultural University, Navsari, Gujarat, India. He enriched his practical knowledge and expertise in the subject of Animal Breeding while disbursing his duties as Scheme Incharge at Livestock Research Station of the same university for 9 years. During the same tenure he also accumulated practical expertise on various aspects of field level breeding programs while heading "All India Coordinated Research Project on Goat Improvement - Surti Field Unit" as Principal Investigator. He completed his Ph.D. in the year 2016 from the same university as an inservice candidate. He had worked on gene expression studies on mammary epithelial cells of buffaloes during his Ph.D. degree program. He had been selected as Associate Professor in the department of Animal Genetics & Breeding, College of Veterinary and Animal Science, Sardar Vallbhbhai, Patel University of Agriculture & Technology, Meerut, Uttar Pradesh, India in the year 2018. Since then he has been heading the same department as Officer-Incharge. He had handled 5 externally funded and 27 institutionally funded research projects. He had coguided two masters students. He has in his credit 64 research papers, 14 research recommendations, 8 lecture notes and 4 success stories. He is a member of 4 professional societies and attended 21 conferences/ symposiums/ workshops. He has remained on a panel of experts for framing question papers for National level, State level examination bodies and various Universities. He is hosting a google site for online teaching https://sites.google.com/view/learnagb and can be reached at drtyagivet@gmail.com for initiating a conversation.

Published by:

Department of Animal Genetics & Breeding
College of Veterinary & Animal Sciences
Sardar Vallabhbhai Patel University of Agriculture and Technology
Meerut- 250 110, Uttar Pradesh, India

To cite this lecture notes:

Tyagi, K 2021, *Statistical analysis of genetic association studies among unrelated individuals*, lecture notes series *Statistical Analysis of Molecular Data Using Software Packages*, Training on "Molecular biology tools and its application in Agriculture and Allied Sciences" Sardar Vallabhbhai Patel University of Agriculture & Technology, Meerut, Uttar Pradesh- 250110, India during 01-14 December 2021. Delivered 08th December 2021. Retrieved online from https://vepub.com