A chi square (X2) statistic is used to investigate whether distributions of categorical variables differ from one another. We wanted to determine the likelihood of the Geneotype at the locus which is significantly different in parkinson's disease individuals Vs Controlled individuals using Chi-sq testing methods.There is an Assumption that the SNP difference is the main cause, which could help us to determine if the people are really affected with Parkinson's disease or not.

cc\_map file : The cc subdirectory indicates data for the Caucasian control individuals.The first column indicates for the chromosome. The second column indicates the specific base-pair (nucleotide) on the chromosome for the location of this SNP. The third column indicates the dbSNP identifier for this SNP. The fourth column indicates the major allele found at this SNP, or the variant (i.e. base-pair) most commonly seen. The fifth column indicates the minor allele found at this SNP, or the variant (i.e. base-pair) least commonly seen. The sixth and seventh columns indicate the frequency of the major and minor alleles seen in this population. The eighth column indicates the number of missing genotypes (i.e. missing measurements).

ND 412 1 T C G G G G A A A A C C ...

cc\_pre file : Each row of this file represents a single individual in the study. “ND 412” indicates the code for the individual. 1 indicates an unaffected individual, while 2 indicates an affected individual. After this number, a series of A, T, C and G characters appear. Each pair of characters represents the sequenced alleles (for two chromosomes: one maternal and one paternal) at a single locus for a single individual.

For example, the first two alleles for individual “ND 412” are T and C. These first two base-pairs correspond with the first row of map file. In other words, individual “ND 412” has at locus rs5747620 a T base-pair on one chromosome, and a C basepair on the other chromosome. Individual “ND 412” has at locus rs2236639 a G base-pair on both chromosomes.

for (i in 1:nrow(cc\_map)) {

Aa=sum(cc\_map[i,4]==cc\_pre[,2\*i+2] & cc\_map[i,5]==cc\_pre[,2\*i+3])

AA=sum(cc\_map[i,4]==cc\_pre[,2\*i+2] & cc\_map[i,4]==cc\_pre[,2\*i+3])

aa=sum(cc\_map[i,5]==cc\_pre[,2\*i+2] & cc\_map[i,5]==cc\_pre[,2\*i+3])

df[i,"AA\_cc"]=AA

df[i,"Aa\_cc"]=Aa

df[i,"aa\_cc"]=aa

}

We have used the chi-squared test to determine whether the genotype distribution seen in Parkinson’s Disease patients is different than control individuals (i.e. a case-control study). A chi-squared test for these data with 2 degrees of freedom yields a p-value of 0.76066 indicating that it is likely that the Parkinson’s Disease distribution matches the control distribution.It seems there is NO such assumption.

We have the learn from the Data Science HW2 project that to see whether distributions of samples differ from each another, we need to perform chi-sq test.A chi-square test compares two samples in a contingency table to see if they are related. A very small chi square test statistic means that your observed data fits your expected data extremely well. In other words, there is a relationship. A very large chi square test statistic means that the data does not fit very well. In other words, there is not a relationship.

And data analytics point of view we have learnt that reading file from all four text file i.e. cc and pd files for all 25 chromosomes using for loop for individual chromosomes.

1

4.105

taking every row in df\_combined[,3:8], create an appropriate table for chi-square processing and then find the appropriate p value.

False

7

We have the learn from the Data Science HW2 project that to:

* How to read the Data frames for the CC and PD files.
* Creating data frames to hold integrated data.
* Using for loop to match the allele count.

and see whether distributions of samples differ from each another, we need to perform chi-sq test.A chi-square test compares two samples in a contingency table to see if they are related.

We can apply the same logic to make money in stock market by using chi-sq to assess whether stock activity of each day is independent of the stock's behavior on previous day. It will increase the accuracy of our predictions.

8

Yes, the same procedure can be used in HW2(chi-squared stat and p-value) used to determine the difference between testing and training data in an investment bank data set. By Comparing past data of stock market and testing data to determine the difference and make the necessary changes in the investment model.