PROJECT 3

Vinh Ton

2023-03-17

ABSTRACT

INTRODUCTION

We have been provided with the BRFSS 2015 data set and introduced to the methods of hierarchical-clustering and Entropy as alternatives to the method of ANOVA in the task of determining whether or not multiple samples have the same mean or distribution.

Now, we are going to put apply these methods to our BRFSS dataset by dividing our dataset into five samples with regard to the GenHlth category, and investigating the 3-way interaction effects of three variables: sex, high cholesterol status, and high blood pressure status. Using these methods of hierarchical-clustering and entropy, we will determine what association the interactions of sex, high cholesterol, and high blood pressure, has on BMI distribution, and how it compares to the overall BMI distribution.

However, we will also test the reliability of entropy and HC-trees through the use of simulated data based on the multinomial distribution for each sub-dataset with respect to 3-way interactions. ### METHODOLOGY

OVERVIEW Each of our sub-datasets with regard to GenHlth will be further subdivided with respect to high blood pressure, high cholesterol, and sex. These datasets will be represented through contingency tables and contingecy tables of proportions. Then, through the use of hierarchical clustering we will discover those with similar distributions, and with entropy we will see which combination of variables will have the greatest predictor of BMI.

To test reliability, a contingency table of proportions will be constructed for each 3-way interaction sub-dataset. Then, we will simulate data for each contingency table from a multinomial distribution in which each row-vector of proportions will serve as the probability for the multinomial distribution.

For the entropy approach, each row will construct its own multinomial distribution with n=1000. Then we use the entropy approach for each row to create a histogram of the Shannon entropies and compare to the actual data.

For the hierarchical clustering approach, we will create 5 such contingency tables as detailed above, and construct HC-trees on each table to investigate how reliable HC is.

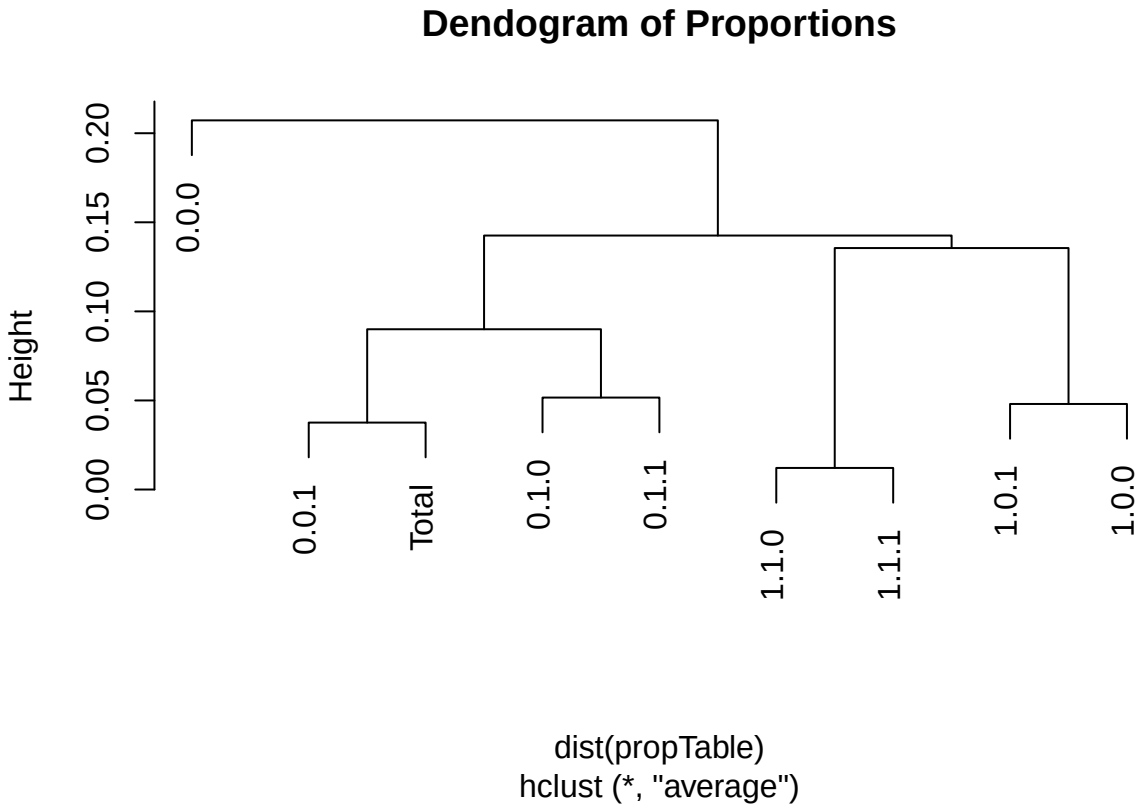
SUB-DATASET 1

```
##       labels  entropies
##  [1,] "0.1.0" "1.19736457961285"
##  [2,] "1.0.1" "1.03625153364893"
```

```
##  [3,] "1.1.0" "1.12395283770773"
##  [4,] "0.0.0" "1.16461675677489"
##  [5,] "1.0.0" "1.07468229264336"
##  [6,] "0.0.1" "1.14414161818085"
##  [7,] "0.1.1" "1.17319618494005"
##  [8,] "1.1.1" "1.11735976289452"
##  [9,] "Total" "1.15008095454969"
```

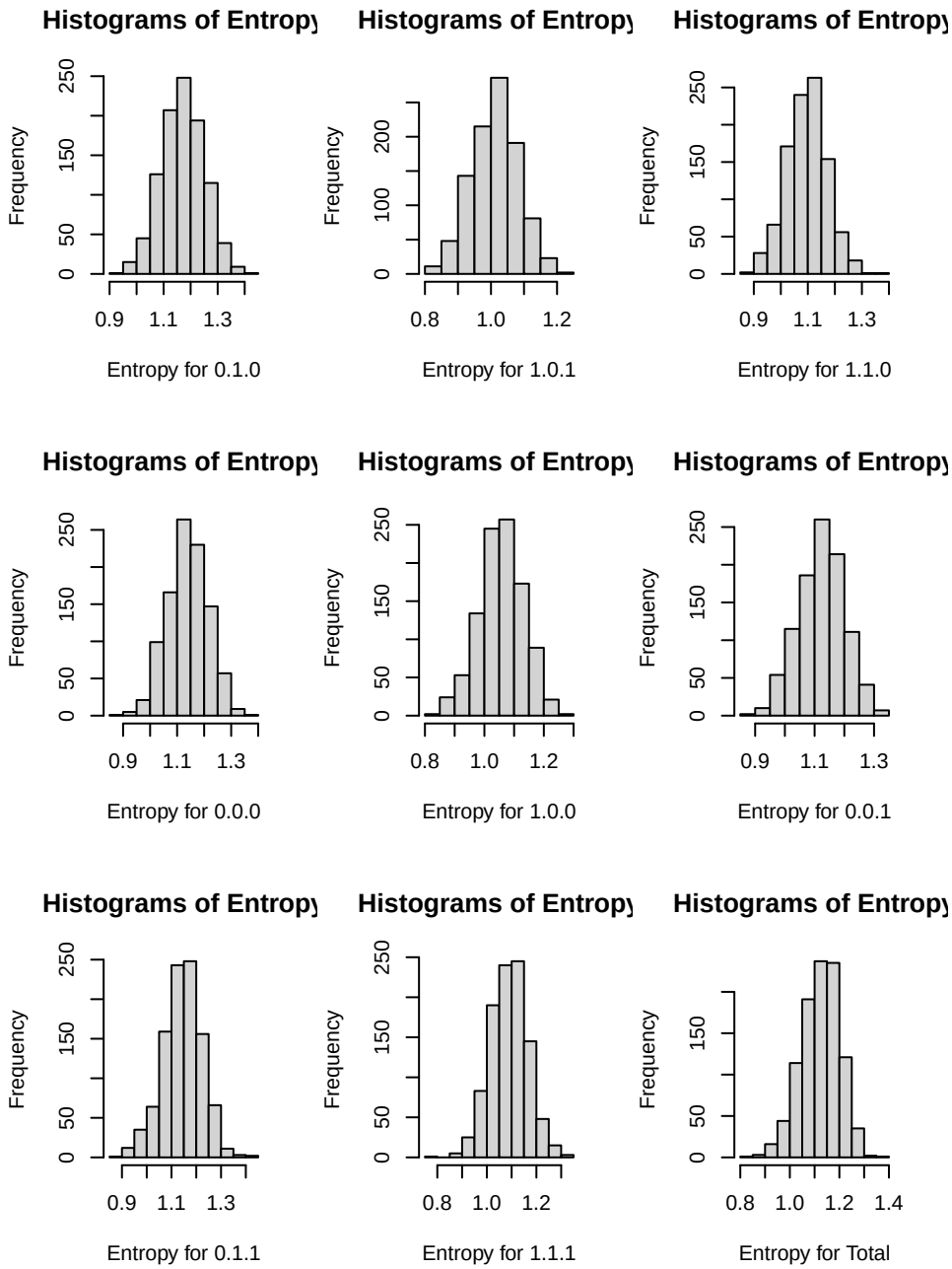
We can see from the entropy table that the best predictors when looking at interaction terms seem to be the combinations of sex being male, low blood pressure, high cholesterol, sex being male with low blood pressure and low cholesterol, and sex being male with high blood pressure and high cholesterol.

The worst predictors seem to be sex being female with high blood pressure and low cholesterol, and sex being female with low blood pressure and low cholesterol, and sex being female with high blood pressure and high cholesterol, since all three have entropies greater than the col-sum entropy.

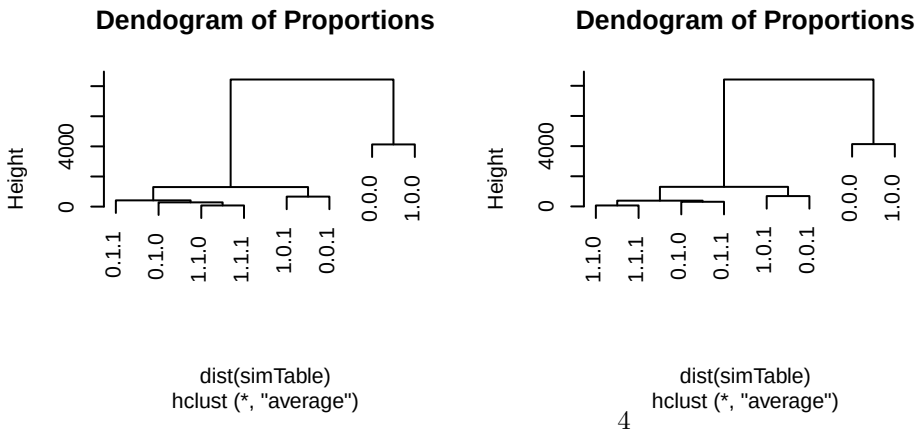


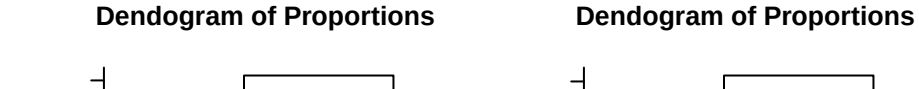
The hierarchical-clustering approach shows that the most closely related structures to the overall BMI distribution are sex being female with low blood pressure and high cholesterol, sex being female with high blood pressure and low cholesterol, and sex being female with high blood pressure and high cholesterol.

Interestingly, it seems that the status of being female sex, low blood pressure, and low cholesterol is the farthest distributed from the overall total, indicating that the interaction of these three seems to be very poor in both distribution similarity and in predictability.



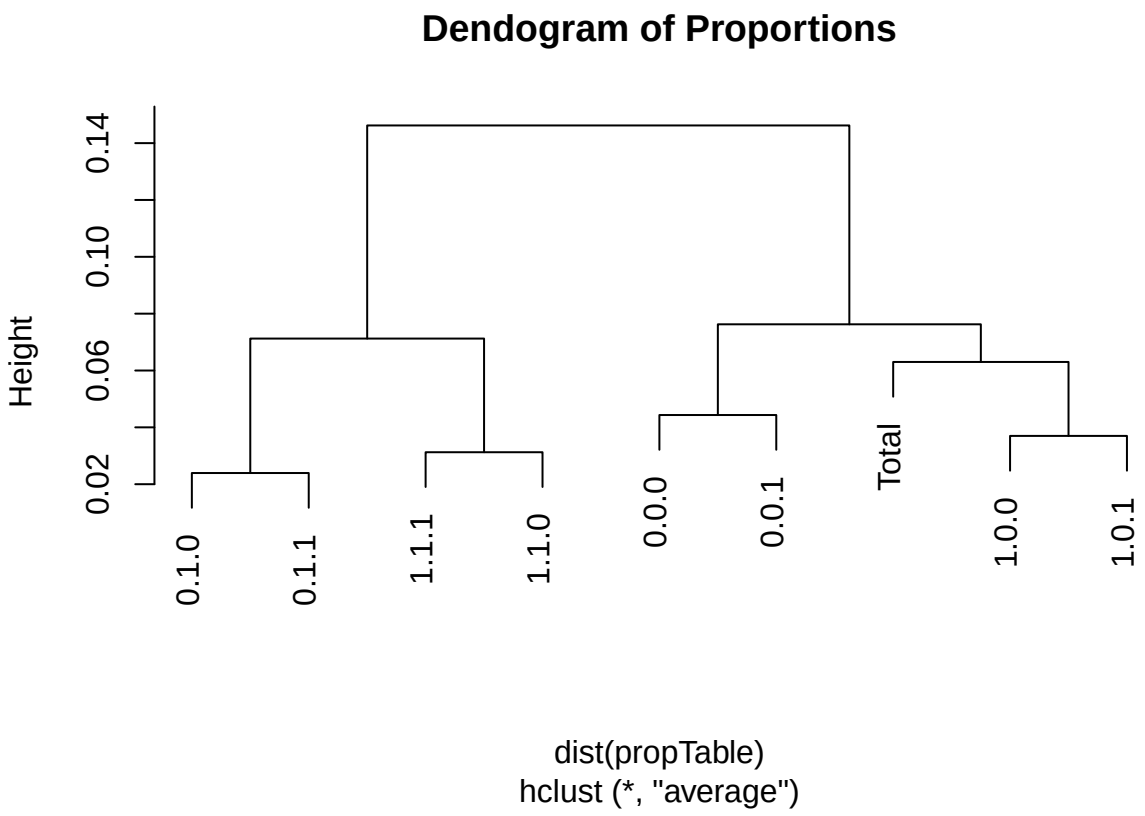
RELIABILITY

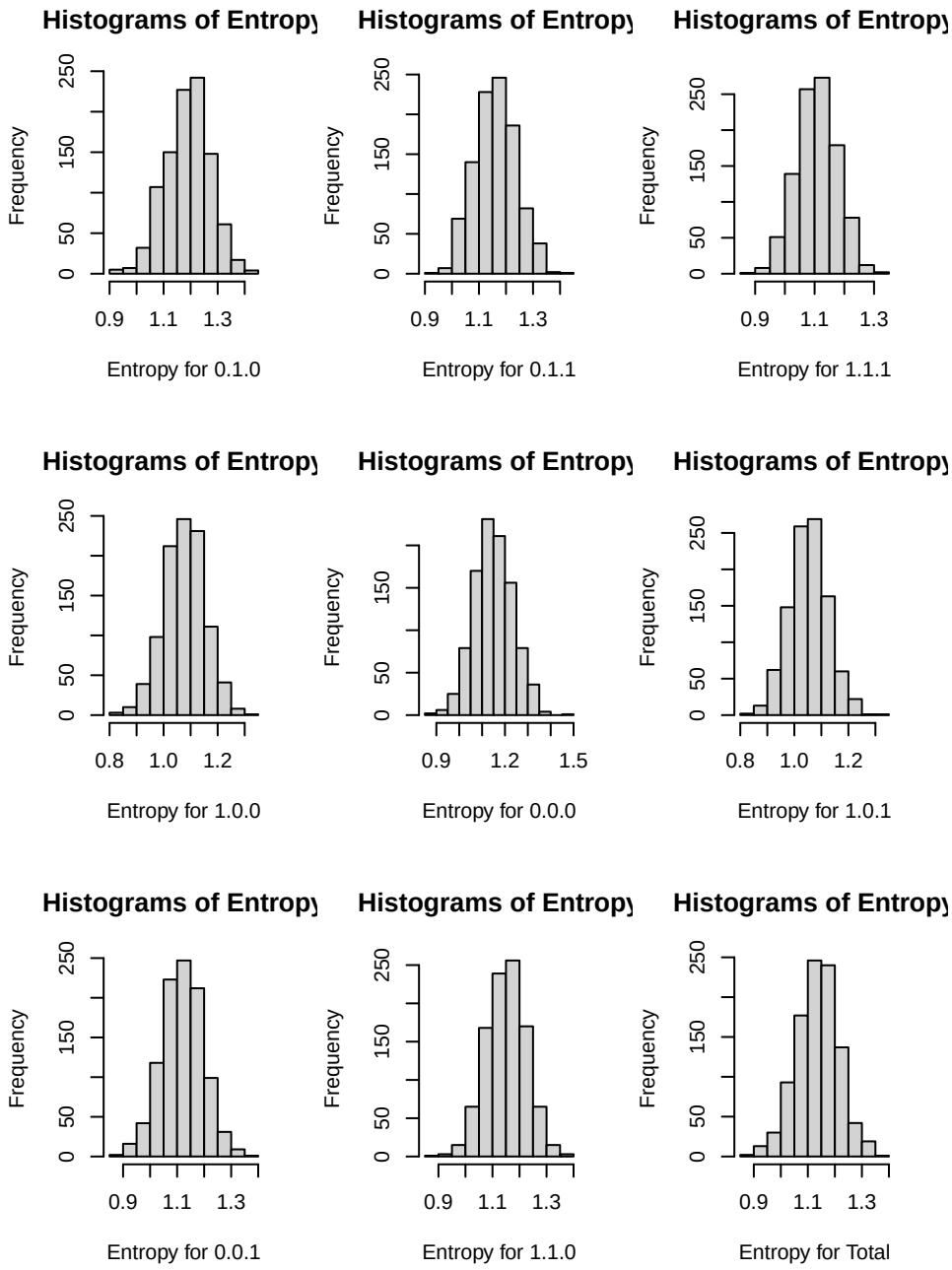




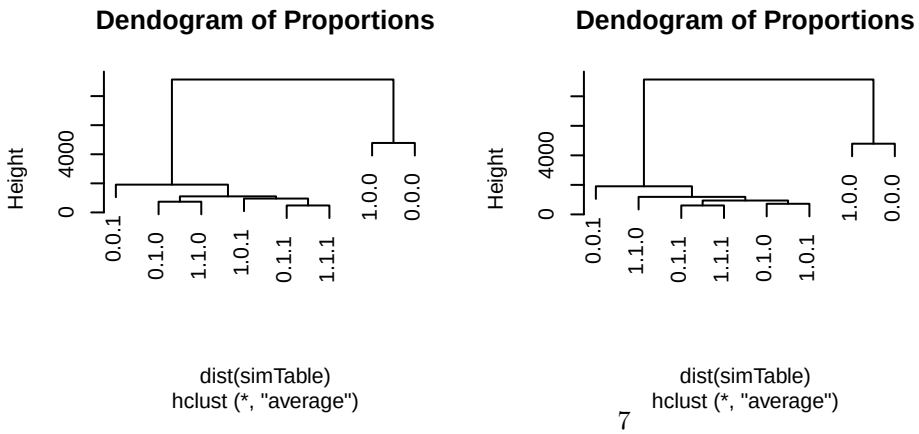
SUB-DATASET 2

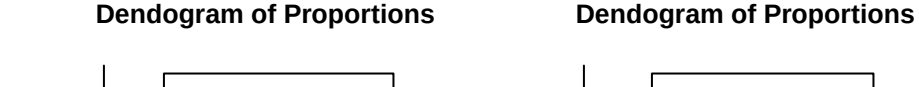
```
##       labels  entropies
##  [1,] "0.1.0" "1.22451469212958"
##  [2,] "0.1.1" "1.18974402656747"
##  [3,] "1.1.1" "1.14003323044178"
##  [4,] "1.0.0" "1.10644036644588"
##  [5,] "0.0.0" "1.17766984392017"
##  [6,] "1.0.1" "1.08006988697691"
##  [7,] "0.0.1" "1.15087904259178"
##  [8,] "1.1.0" "1.17796524239297"
##  [9,] "Total" "1.16698989908534"
```





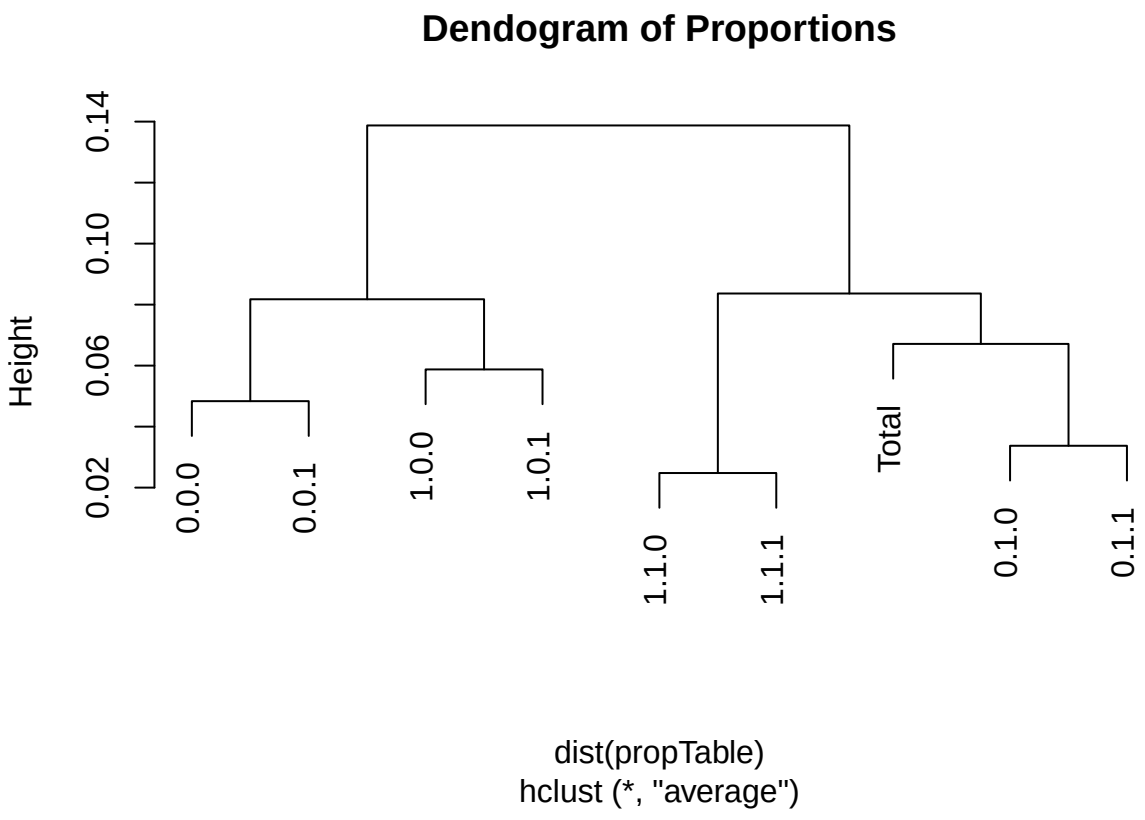
RELIABILITY

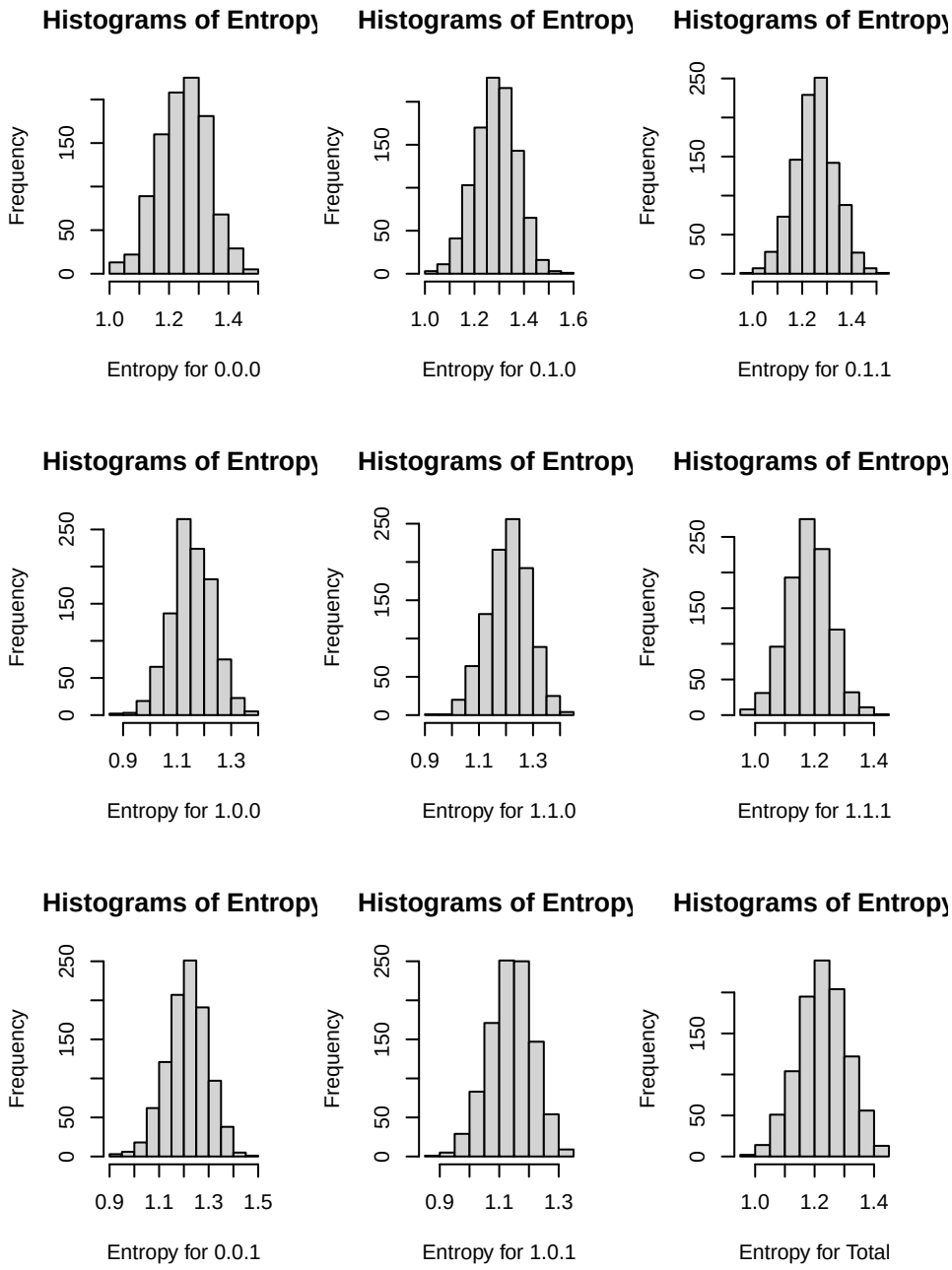




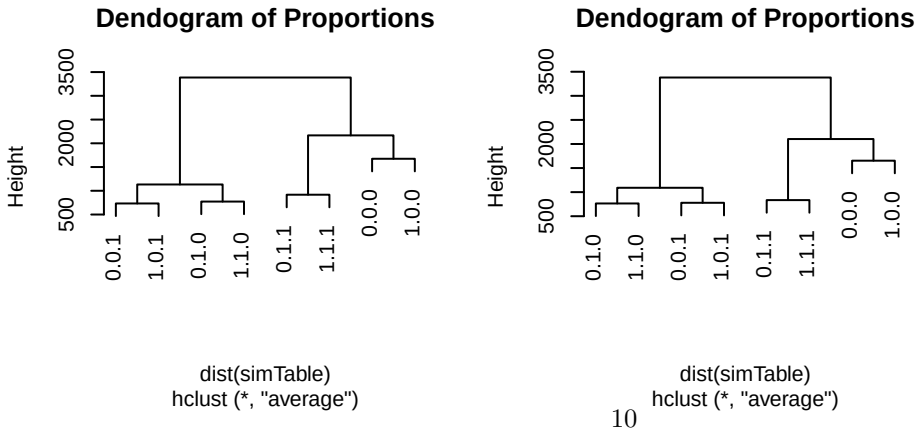
SUB-DATASET 3

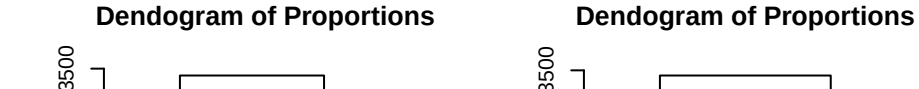
```
##       labels  entropies
##  [1,] "0.0.0" "1.27963965462985"
##  [2,] "0.1.0" "1.32172759724418"
##  [3,] "0.1.1" "1.28537080540787"
##  [4,] "1.0.0" "1.18445677926223"
##  [5,] "1.1.0" "1.24652382232116"
##  [6,] "1.1.1" "1.21071527569342"
##  [7,] "0.0.1" "1.24883011246159"
##  [8,] "1.0.1" "1.16718240007635"
##  [9,] "Total" "1.25827472958179"
```





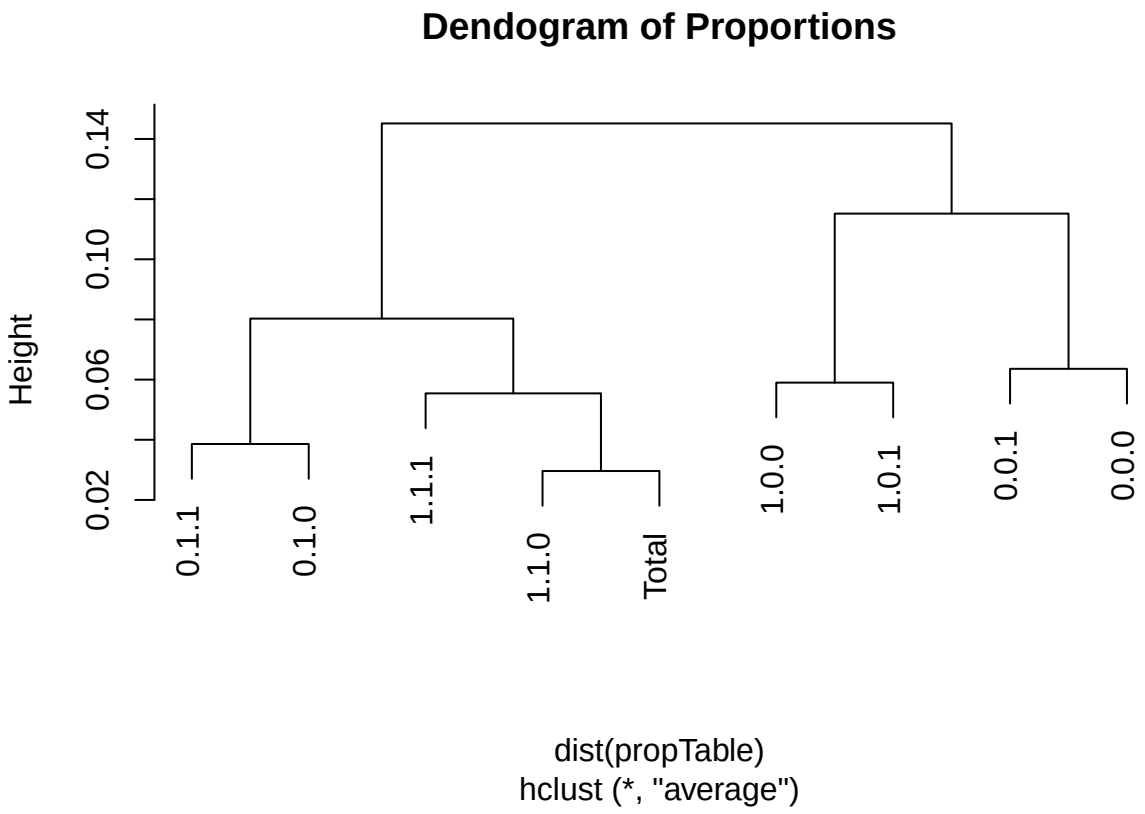
RELIABILITY

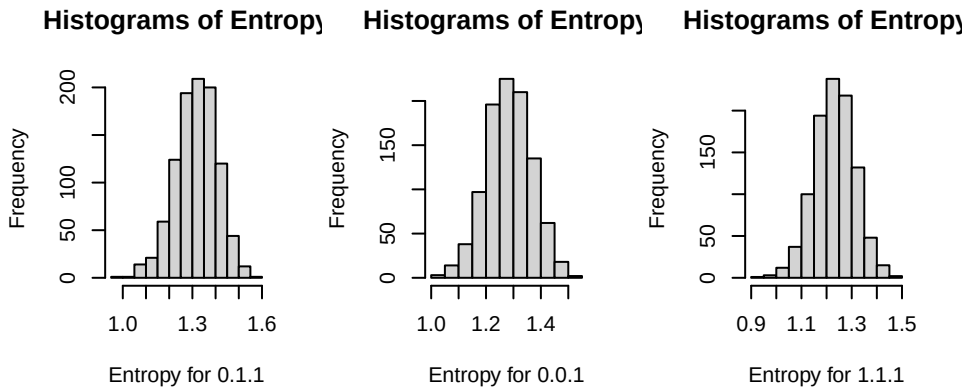


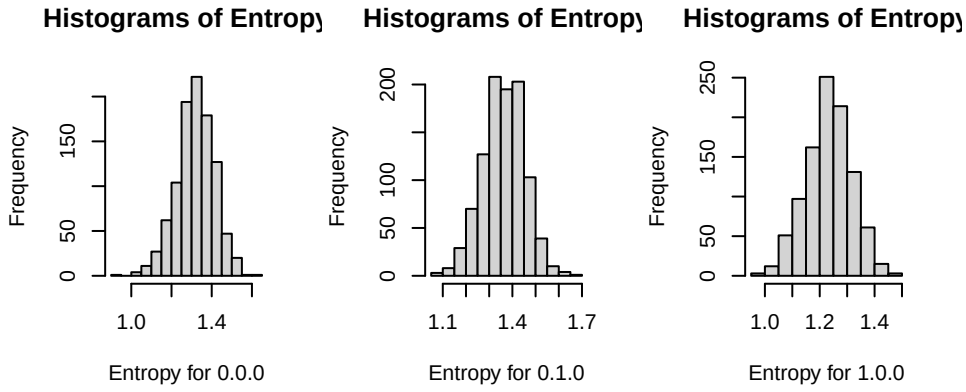


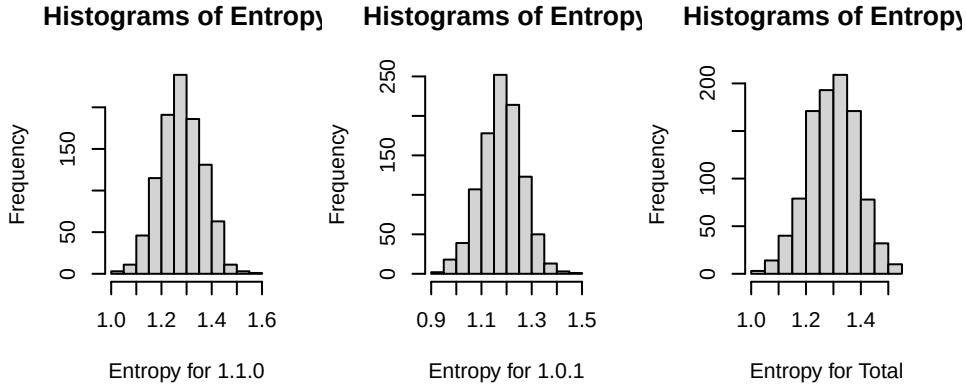
SUB-DATASET 4

```
##       labels  entropies
##  [1,] "0.1.1" "1.35581519081991"
##  [2,] "0.0.1" "1.31988971032999"
##  [3,] "1.1.1" "1.2716152942841"
##  [4,] "0.0.0" "1.35466000723481"
##  [5,] "0.1.0" "1.39991639337012"
##  [6,] "1.0.0" "1.26717850640731"
##  [7,] "1.1.0" "1.31129601036873"
##  [8,] "1.0.1" "1.2082019878063"
##  [9,] "Total" "1.3335501750471"
```

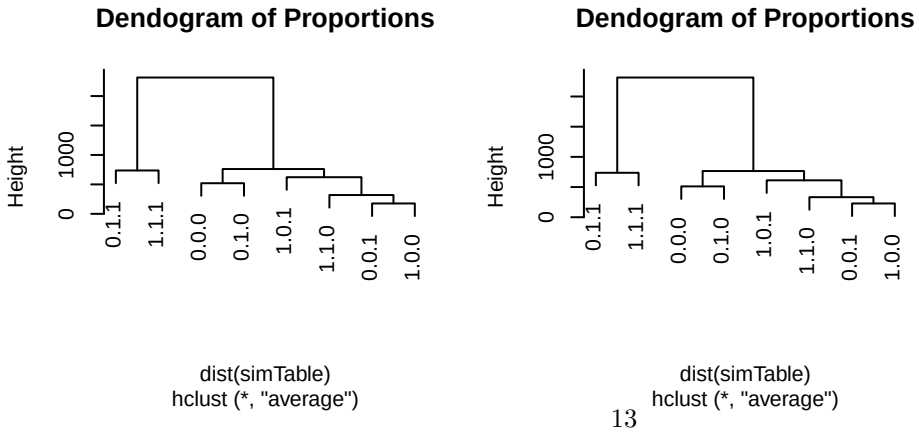


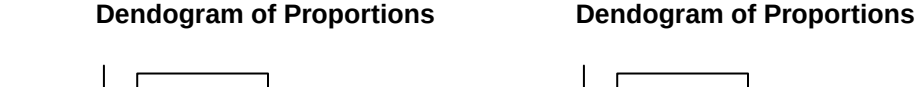






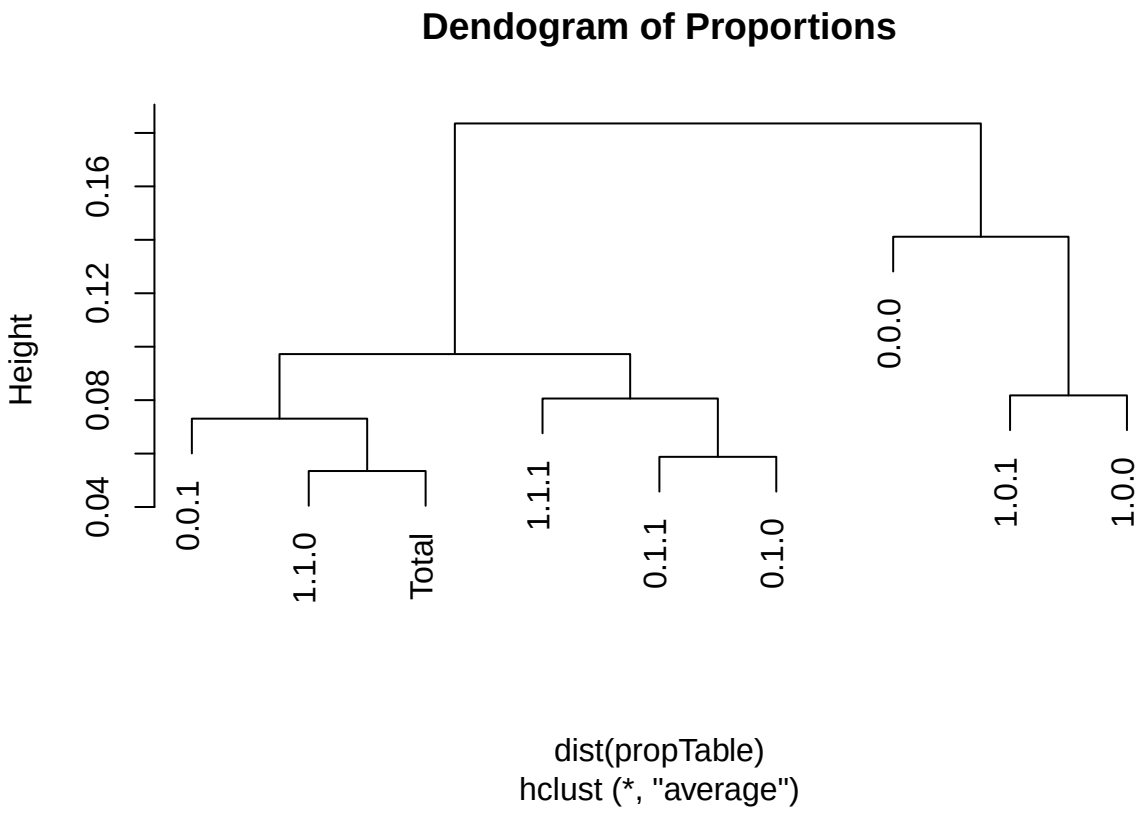
RELIABILITY

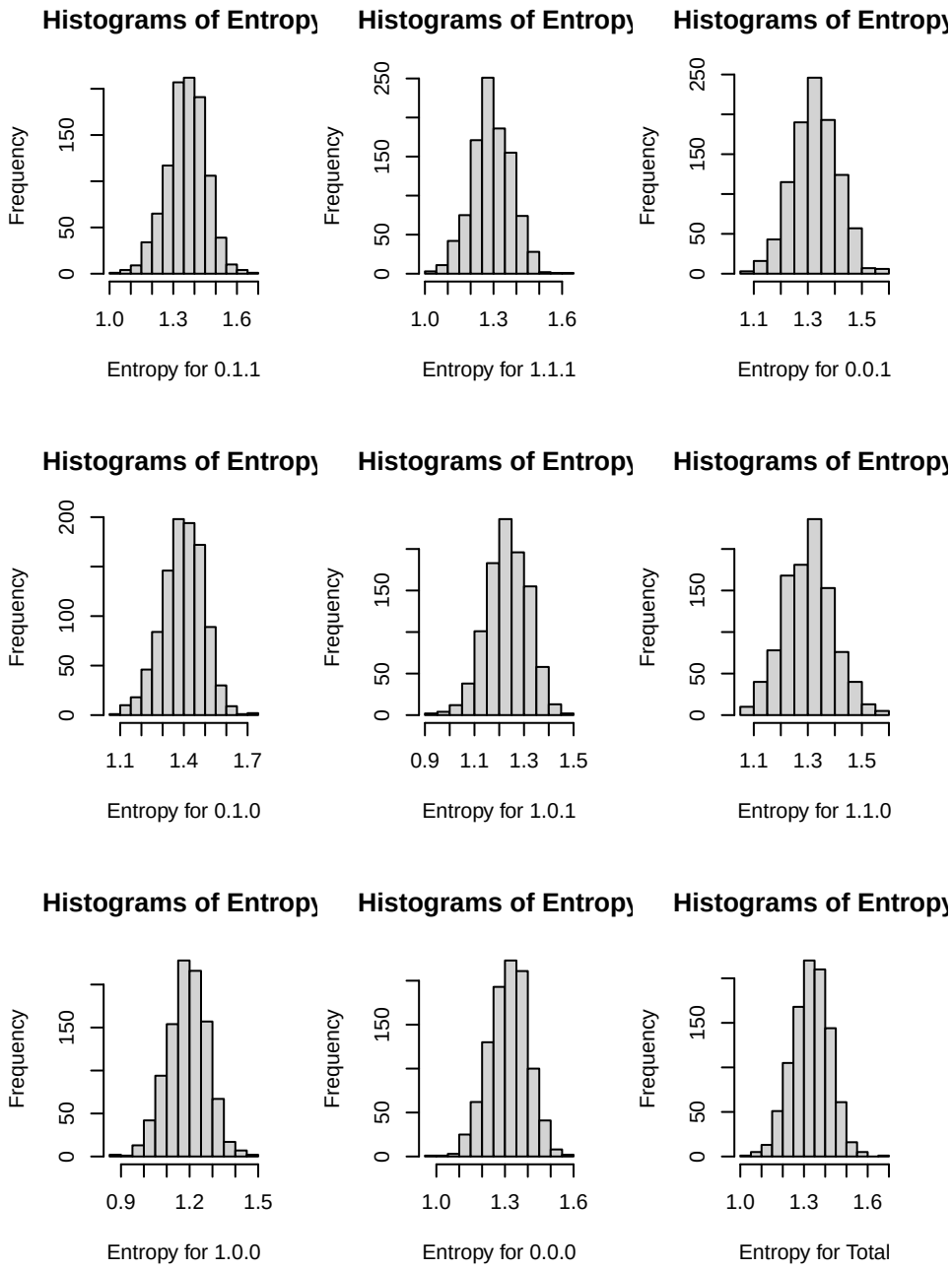




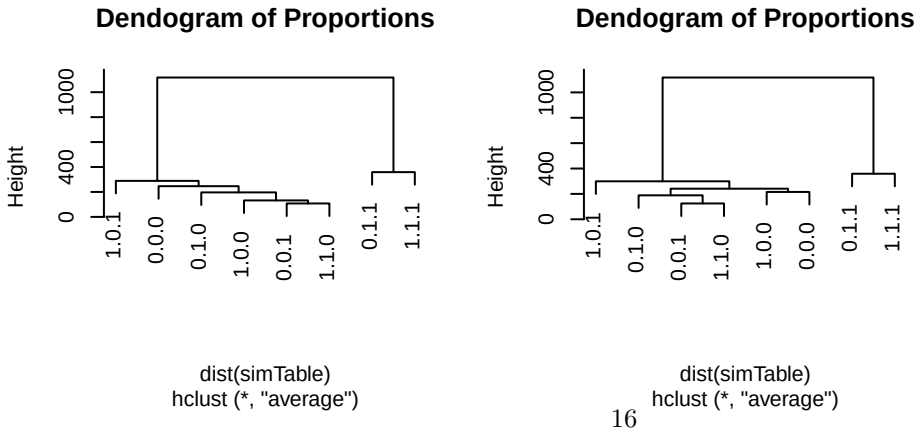
SUB-DATASET 5

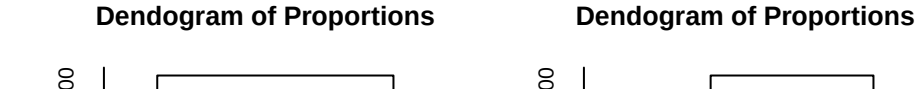
```
##       labels  entropies
##  [1,] "0.1.1" "1.40013070982235"
##  [2,] "1.1.1" "1.32593135789839"
##  [3,] "0.0.1" "1.35874124411276"
##  [4,] "0.1.0" "1.43573188594784"
##  [5,] "1.0.1" "1.25939948550824"
##  [6,] "1.1.0" "1.33959634345703"
##  [7,] "1.0.0" "1.2232122801643"
##  [8,] "0.0.0" "1.3457752499349"
##  [9,] "Total" "1.3694520013872"
```





RELIABILITY





KEY RESULTS

CONCLUSION