Darling Judah Hsu

An Investigation into Hierarchical Trees vs. Entropy

Abstract

The primary purpose of this project is to investigate the effects of varying factors, as well as their interactions, on BMI. By using a dataset found on kaggle by the Center for Disease Control and Prevention (CDC), filtered by user Alex Teboul, we will use different BMI data of experimental subjects, grouped according to certain categories, and estimate the effect of the groupings on BMI. We'll be using the Shannon entropy approach, as well as a hierarchical clustering approach, to determine the categorical variables' effects- as well as to determine which approach may be more useful.

Method

To investigate the effects of certain categorical variables on BMI, we will primarily be focusing on the effects of variables "GenHlth" (general health), "Smoker," "HvyAlcoholConsump" (heavy alcohol consumption), and "HighBP" (high blood pressure). The general health variable contains ordinal data spanning from values 1-5, where 1 indicates excellent health and 5 indicates poor health. To preface, the smoker variable contains binary data where a 1 indicates they are a smoker while a 0 indicates the opposite. Likewise, the heavy alcohol consumption and high blood pressure variables also contain binary data, where a 1 indicates that they do consume a lot of alcohol or have high blood pressure, while a 0 indicates the opposite. As for chart names, the naming convention is as follows: GenHlth_Smoker_HvyAlcoholConsump_HighBP.

In order to estimate the effects of the categorical variables, we'll be using two approaches: the shannon entropy approach and the hierarchical clustering approach.

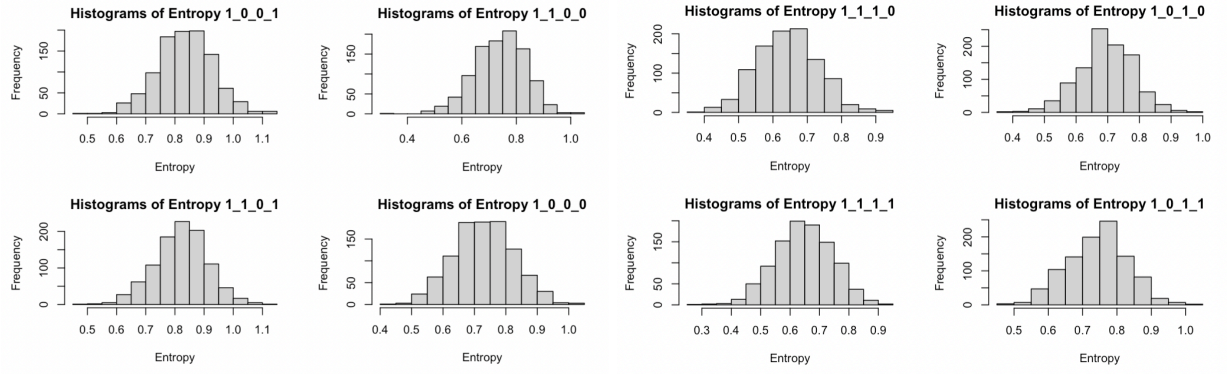
In the shannon entropy approach, we first split the original data into various groups based on the categorical variable selected. Since we've chosen GenHlth, Smoker, HvyAlcoholConsump, and HighBP, we split the data into new samples representing the interactions of the variables. For example, we create a new sample of BMI data for people who scored a 1 on GenHlth, 1 on Smoker, 1 on HvyAlcoholConsump, and 1 on HighBP- this group would represent people with poor general health, who smoke, who drink often, and who have high blood pressure. In total we create 40 different samples of BMI data based on the aforementioned characteristics. From there, we create contingency tables of 10 bins for proportions of BMI, within each group. For example, if BMI ranges from 1-40 then the following bins will be 1-4, 5-8, 9-12, 13-16, 17-20, etc… Within each category bin, we find the proportion of people, within the group, whose BMIs fall within the bins. As a result, we obtain a vector of proportions for each of the 40 interaction subgroups. Next, we use the code in r "multinom" to randomly sample BMI values many times, using the interaction as a multinomial variable where each the probability of getting any observation is equal to the previously calculated proportion of its bin. After getting a large multinomial sample for each interaction's proportion vector, we find the Shannon entropies of the sample and represent it as a histogram. As a result, we obtain 40 histograms of shannon entropy- each representing the entropy of an interaction. Shannon entropy generally measures "disorder" or how much information some measurement (or interaction in this case) tells us; because a smaller entropy implies that the measurement tells us more information, we generally look for smaller entropy values. However, in this case, our focus shifts to a comparison of entropy values rather than looking at them on

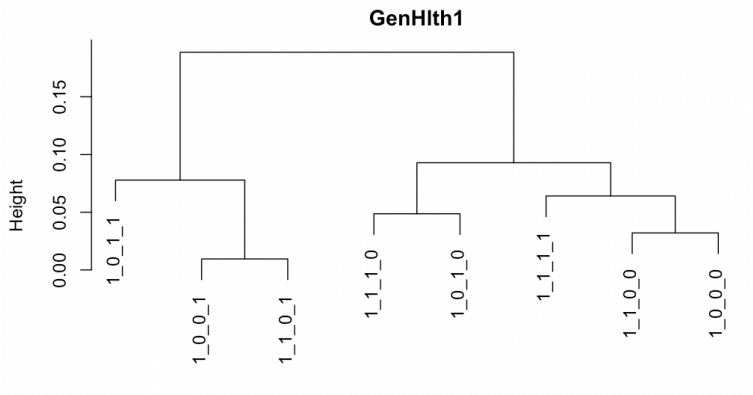
their own. For this entropy approach, we will create histogram entropies for more general cases of the interactions (like for the interaction of general health vs just being a smoker) to compare with the interactions in order to determine whether or not the interactions give us more information than their broader counterparts. For efficiency and readability, we'll primarily be comparing interactions with their broadest cases unless entropies indicate that the interaction gives noticeably more information than its broadest case.

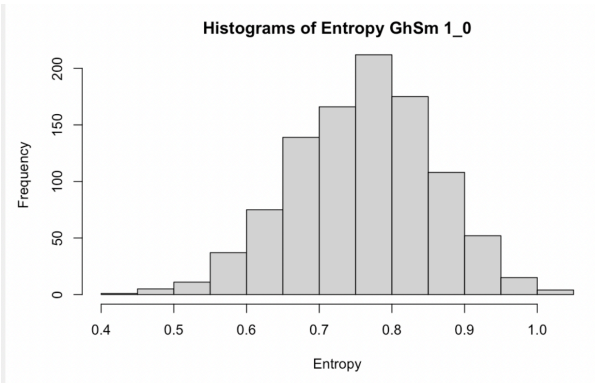
For the hierarchical tree approach, the logic is rather similar though the effect is not. Similar to the entropy approach, we'll still find a vector of proportions for each interaction and sample from the multinomial distribution. This time, however, instead of finding entropy histograms, we'll create a hierarchical tree showing the distances to better visualize differences in the information given to us by the distributions.

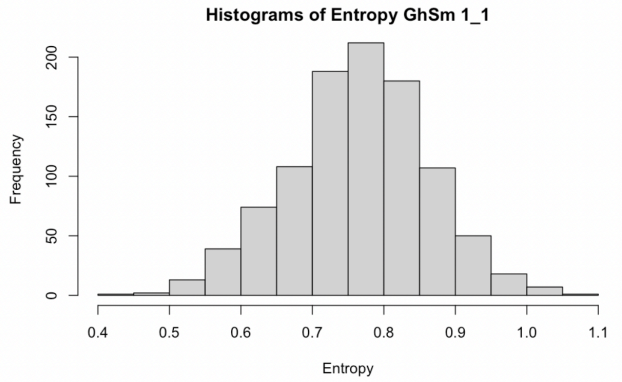
Results (charts of histograms and hierarchical trees listed before)

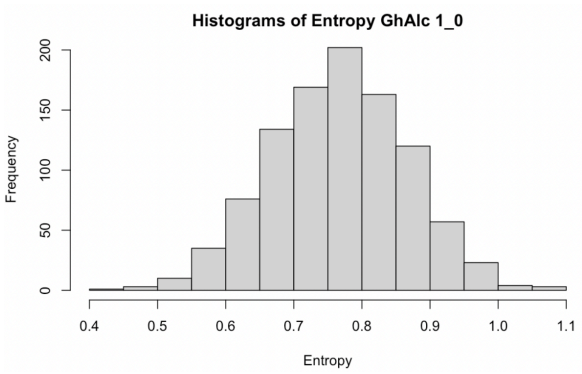
Charts for General Health lv. 1

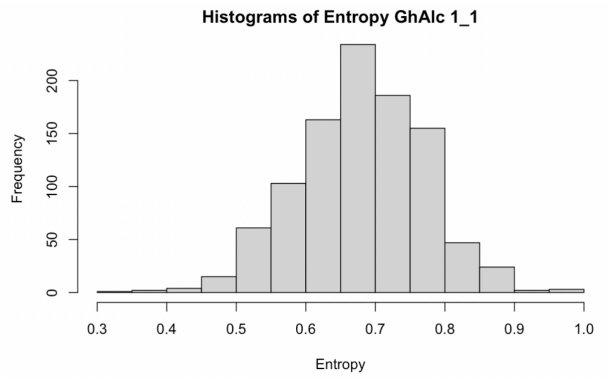


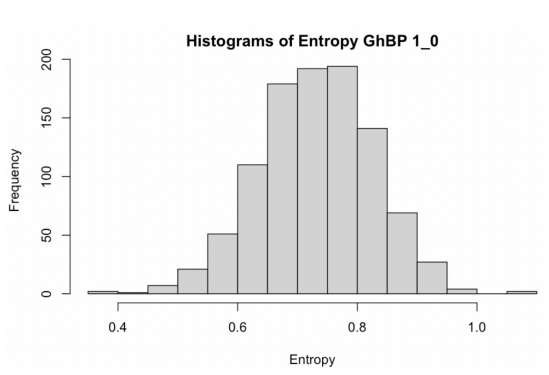


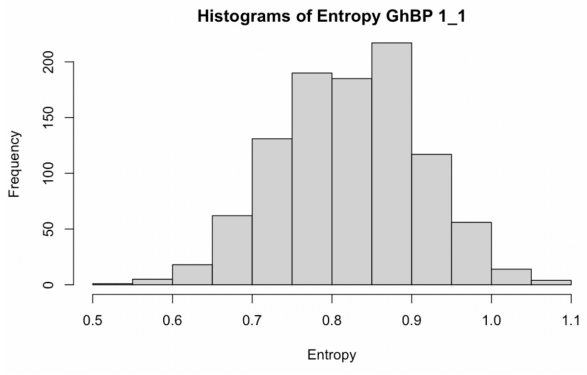




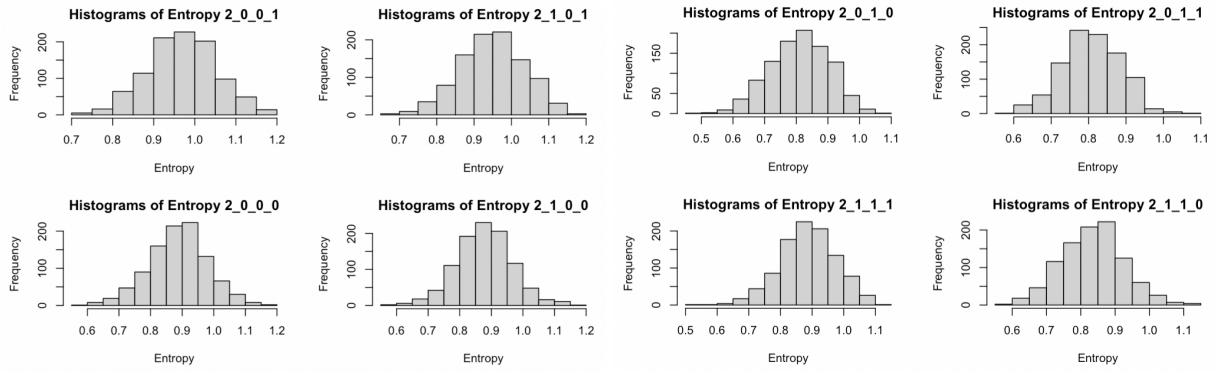

Histograms of Entropy GhAlc 1_0
Frequency
0
50
100
150
200
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
Entropy

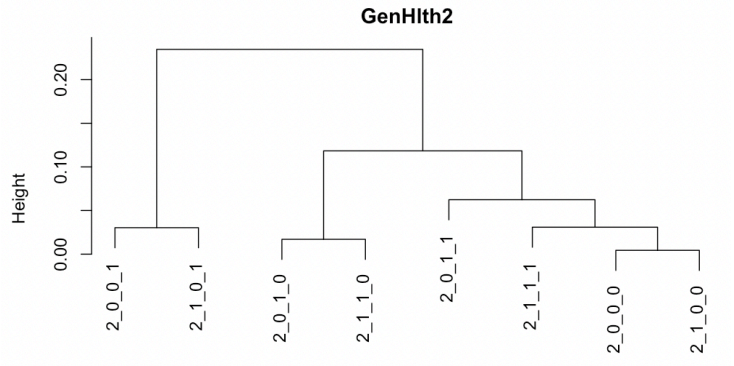

Histograms of Entropy GhAlc 1_1
Frequency
0
50
100
150
200
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Entropy

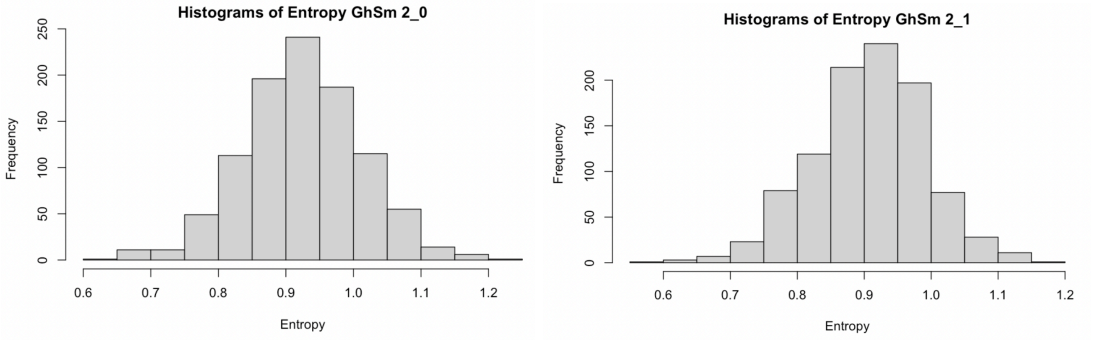

Histograms of Entropy GhBP 1_0
Frequency
0
50
100
150
200
0.4
0.6
0.8
1.0
Entropy

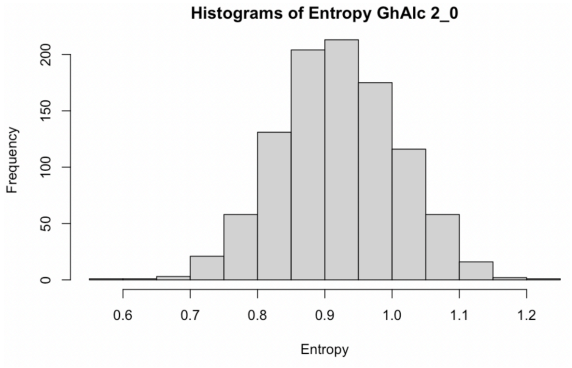

Histograms of Entropy GhBP 1_1
Frequency
0
50
100
150
200
0.5
0.6
0.7
0.8
0.9
1.0
1.1
Entropy

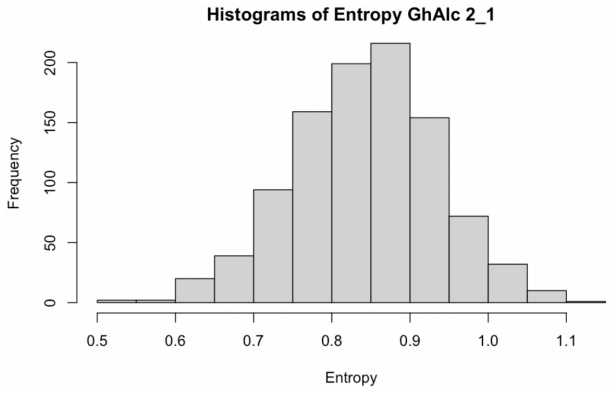
Charts for General Health lv. 2

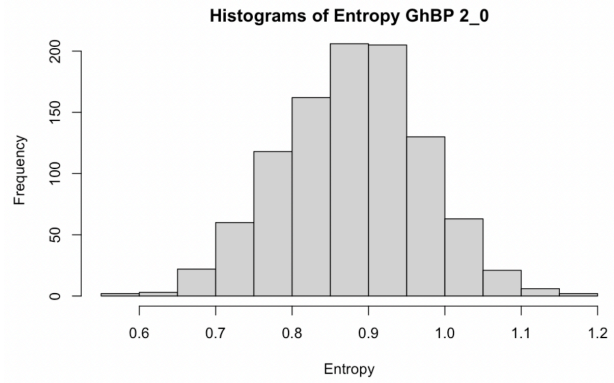


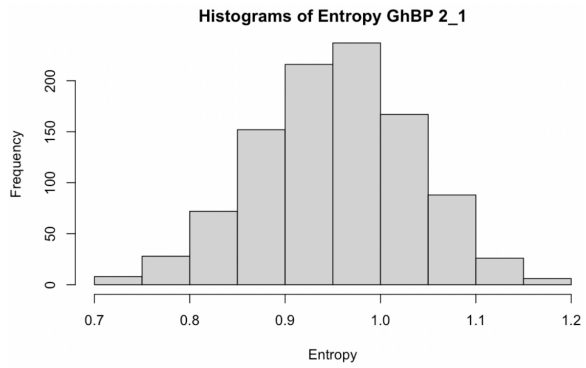




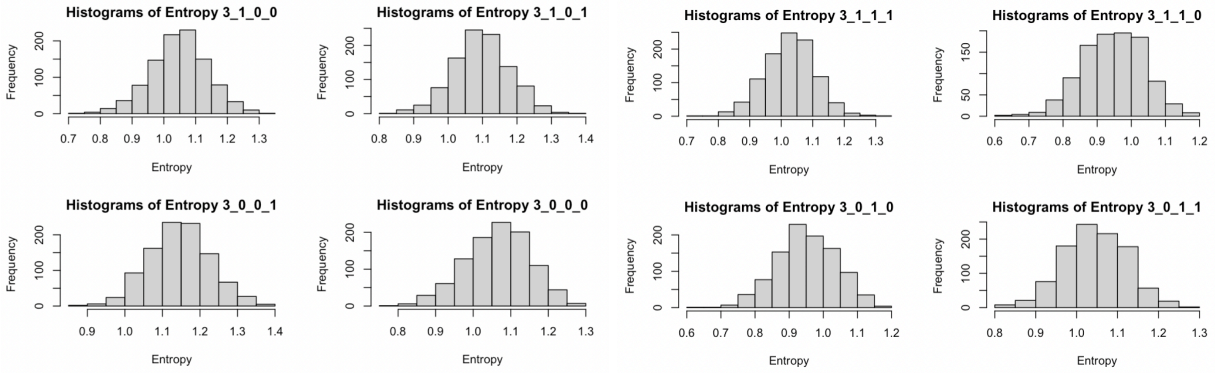

Histograms of Entropy GhAlc 2_0
Frequency
0
50
100
150
200
0.6
0.7
0.8
0.9
1.0
1.1
1.2
Entropy

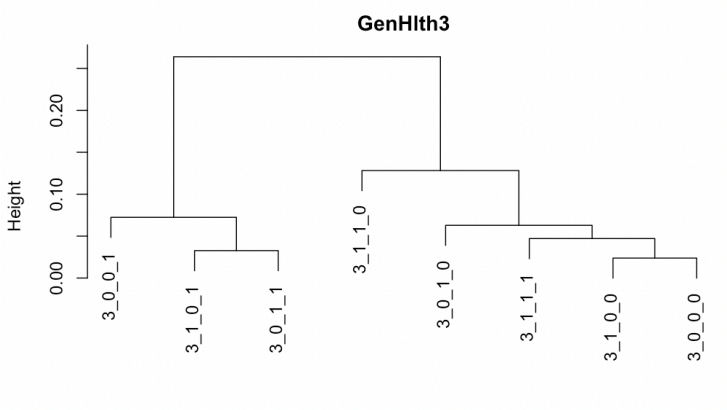

Histograms of Entropy GhAlc 2_1
Frequency
0
50
100
150
200
0.5
0.6
0.7
0.8
0.9
1.0
1.1
Entropy

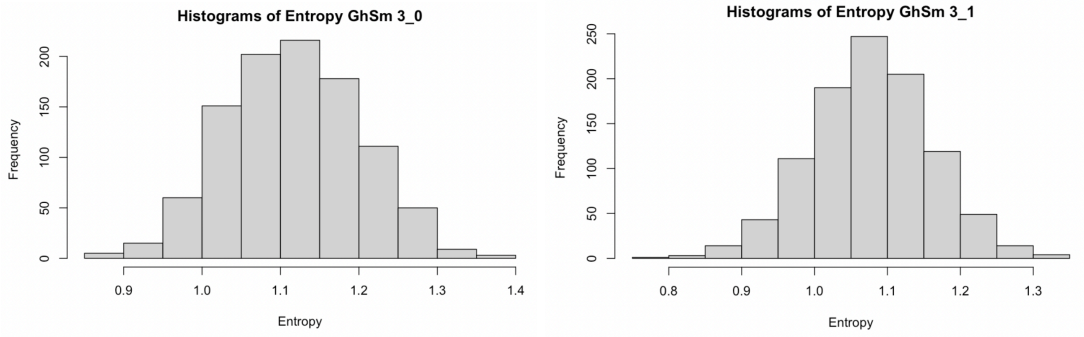

Histograms of Entropy GhBP 2_0
Frequency
0
50
100
150
200
0.6
0.7
0.8
0.9
1.0
1.1
1.2
Entropy

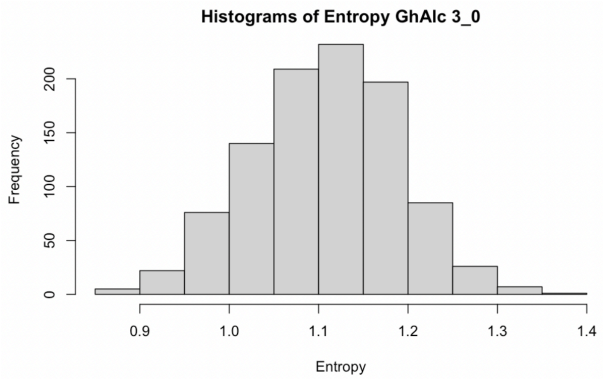

Histograms of Entropy GhBP 2_1
Frequency
0
50
100
150
200
0.7
0.8
0.9
1.0
1.1
1.2
Entropy

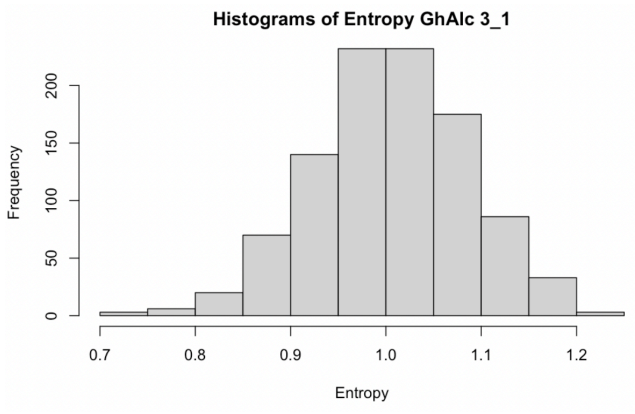
Charts for General Health lv. 3

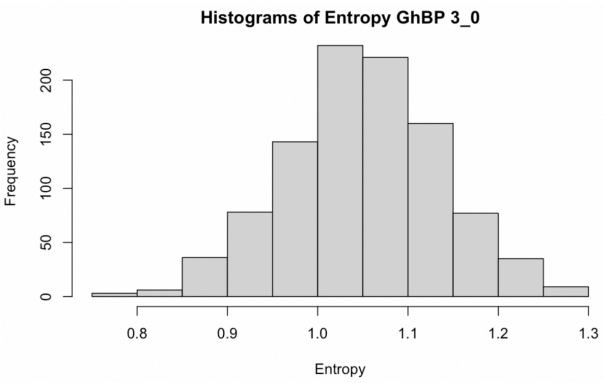


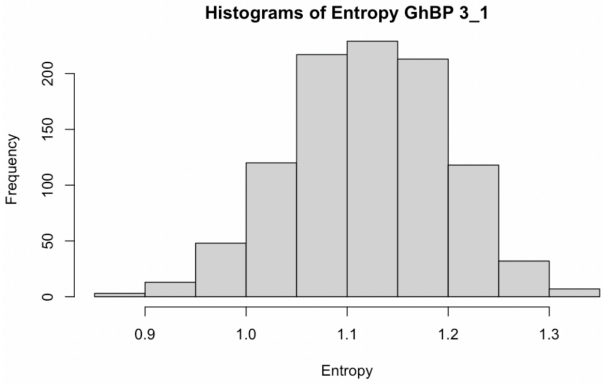




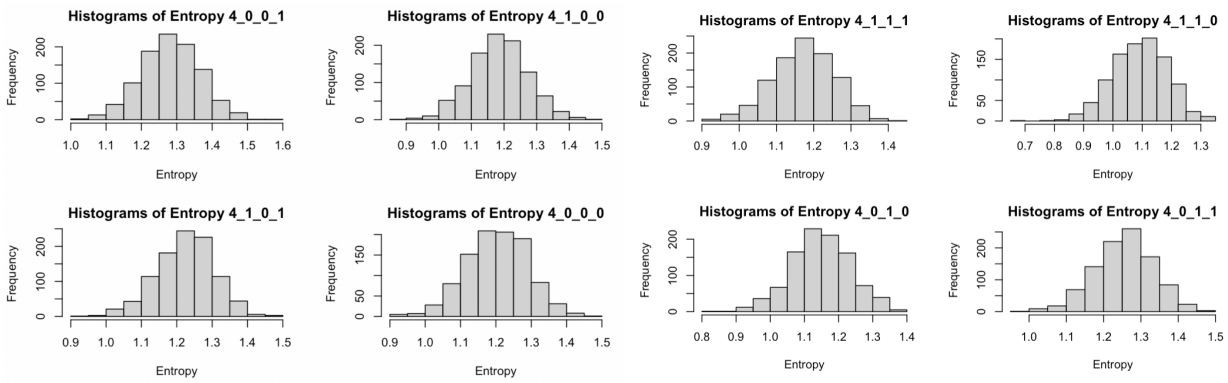

Histograms of Entropy GhAlc 3_0
Frequency
0
50
100
150
200
0.9
1.0
1.1
1.2
1.3
1.4
Entropy

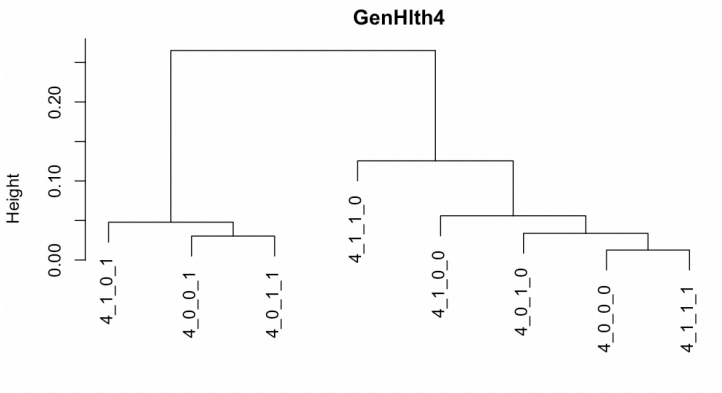

Histograms of Entropy GhAlc 3_1
Frequency
0
50
100
150
200
0.7
0.8
0.9
1.0
1.1
1.2
Entropy

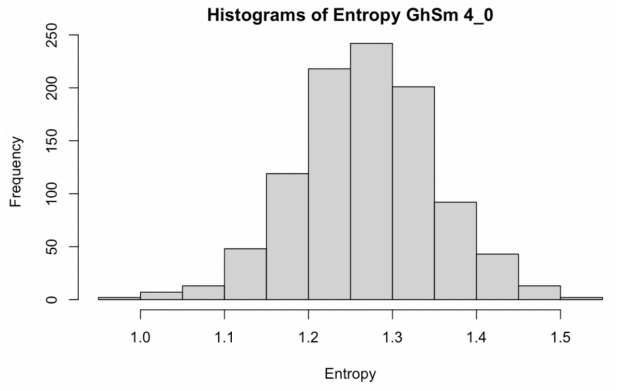

Histograms of Entropy GhBP 3_0
Frequency
0
50
100
150
200
0.8
0.9
1.0
1.1
1.2
1.3
Entropy

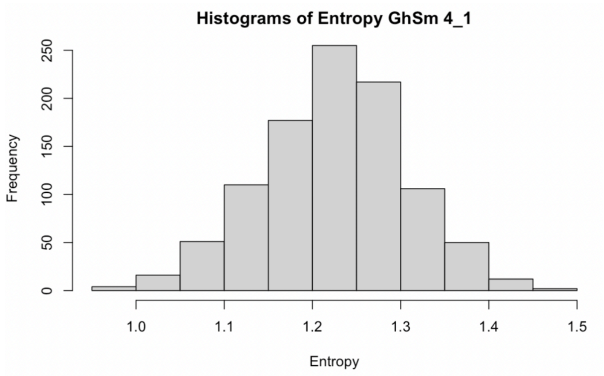

Histograms of Entropy GhBP 3_1
Frequency
0
50
100
150
200
0.9
1.0
1.1
1.2
1.3
Entropy

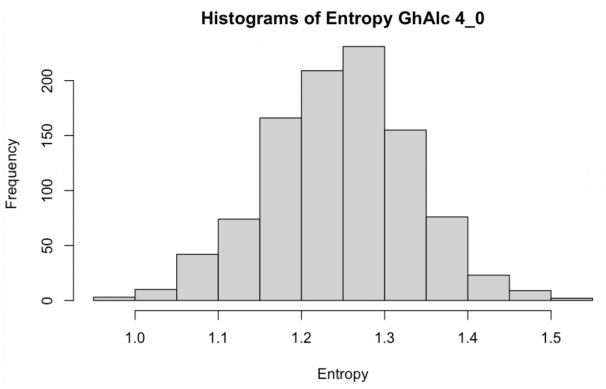
Charts for General Health lv. 4

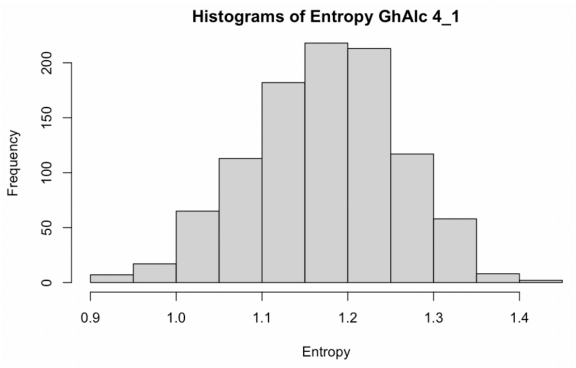


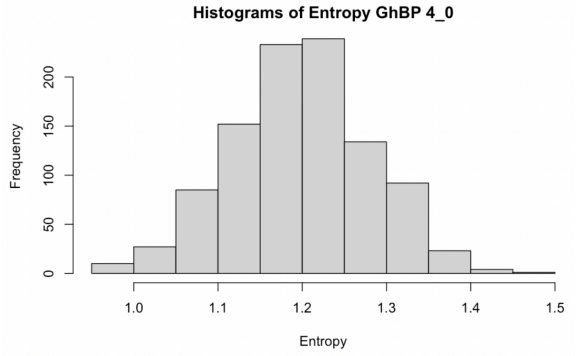


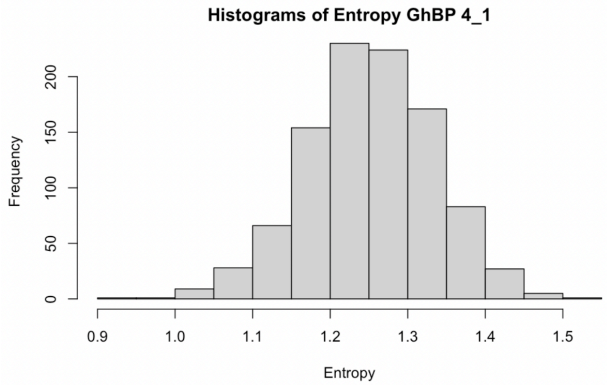




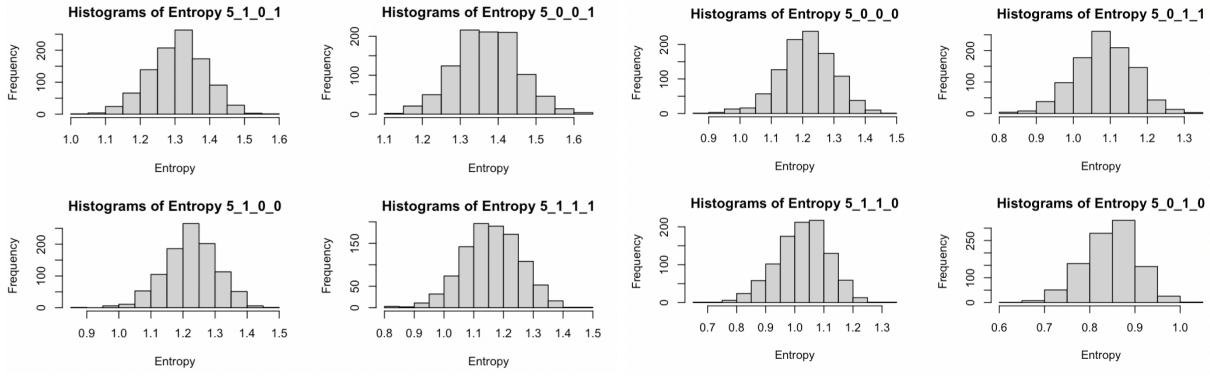

Histograms of Entropy GhAlc 4_0
Frequency
0
50
100
150
200
1.0
1.1
1.2
1.3
1.4
1.5
Entropy

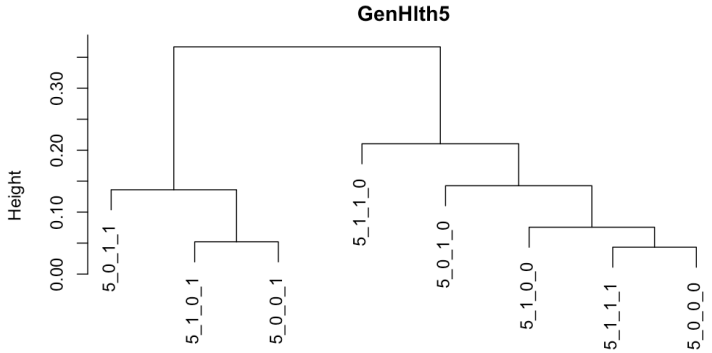

Histograms of Entropy GhAlc 4_1
Frequency
0
50
100
150
200
0.9
1.0
1.1
1.2
1.3
1.4
Entropy

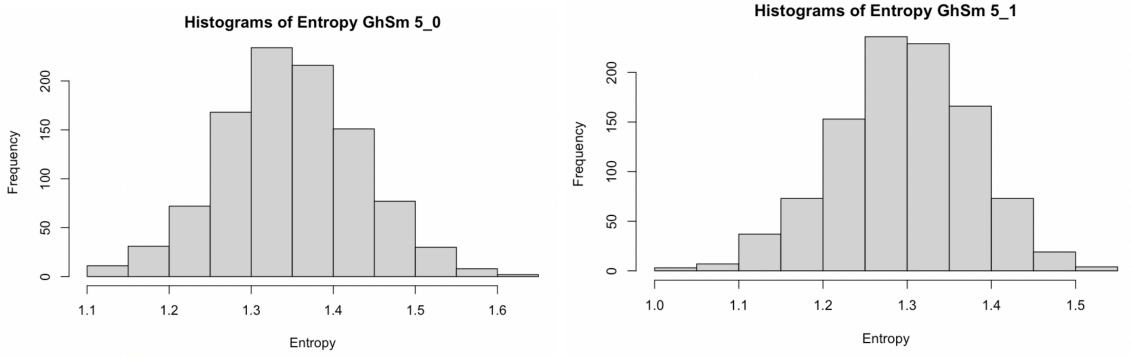

Histograms of Entropy GhBP 4_0
Frequency
0
50
100
150
200
1.0
1.1
1.2
1.3
1.4
1.5
Entropy

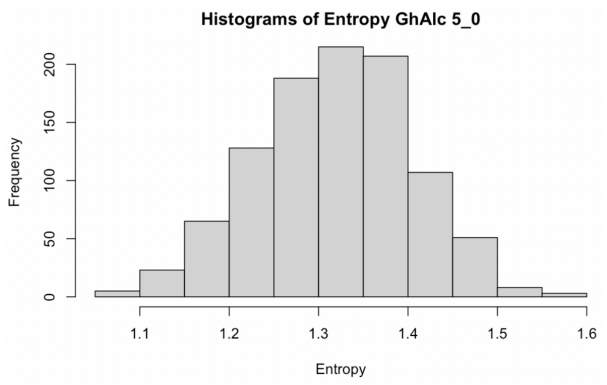

Histograms of Entropy GhBP 4_1
Frequency
0
50
100
150
200
0.9
1.0
1.1
1.2
1.3
1.4
1.5
Entropy

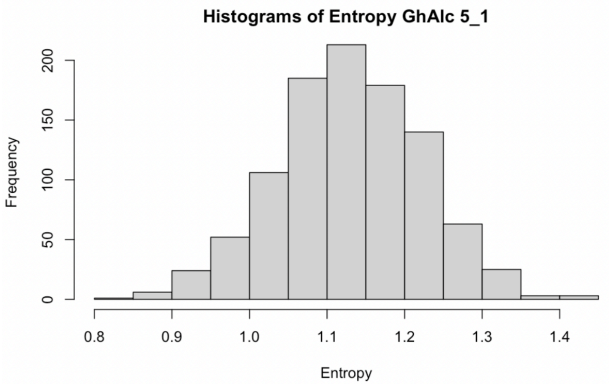
Charts for General Health lv. 5

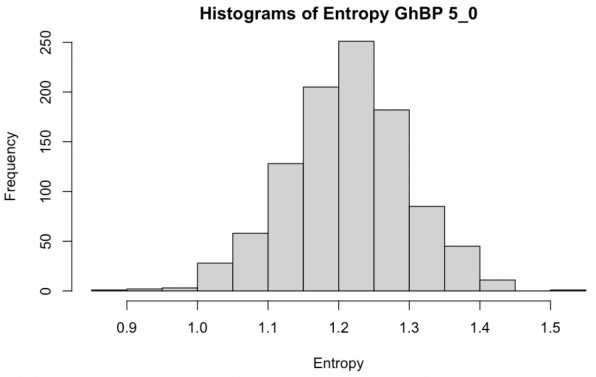


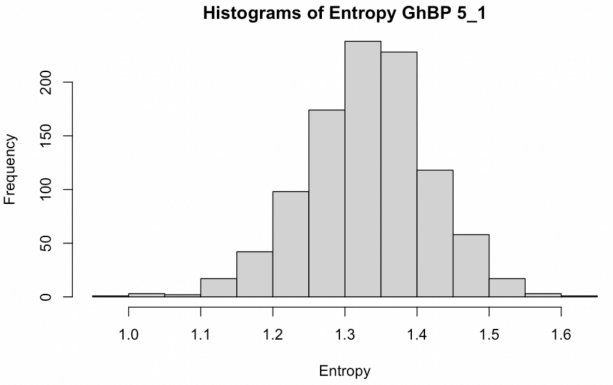




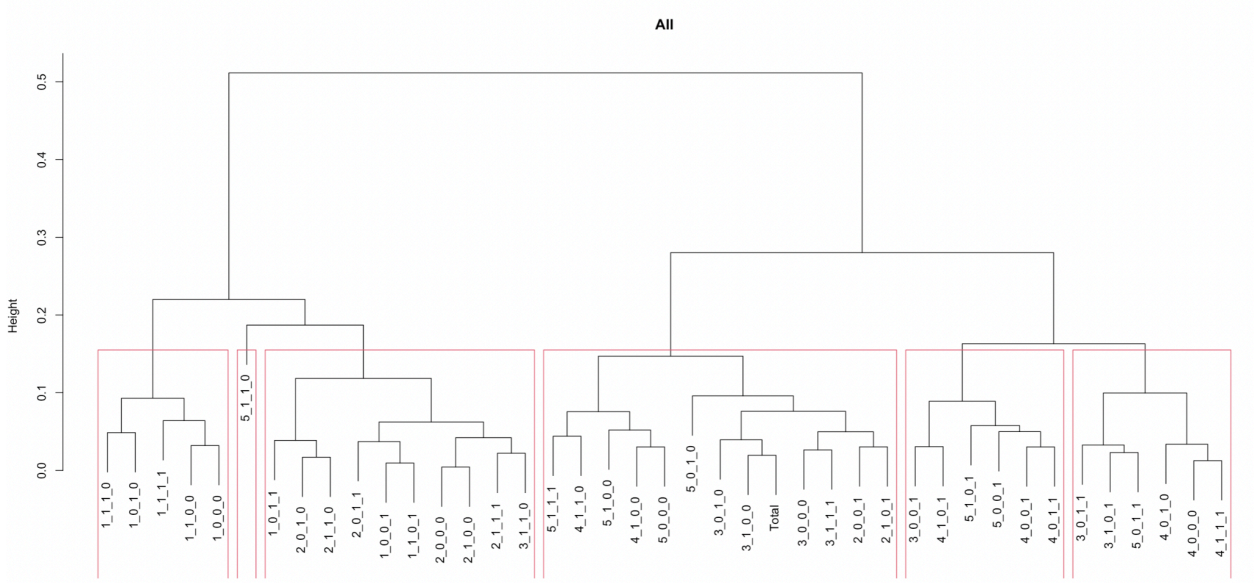

Histograms of Entropy GhAlc 5_0
Frequency
0
50
100
150
200
1.1
1.2
1.3
1.4
1.5
1.6
Entropy

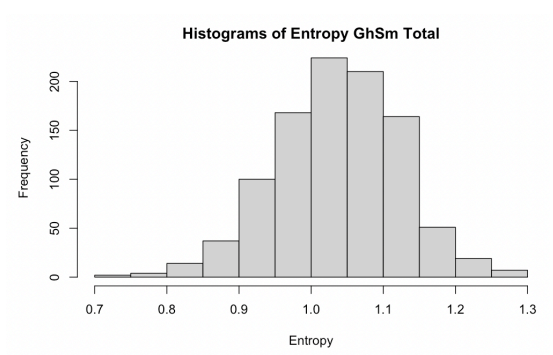

Histograms of Entropy GhAlc 5_1
Frequency
0
50
100
150
200
0.8
0.9
1.0
1.1
1.2
1.3
1.4
Entropy

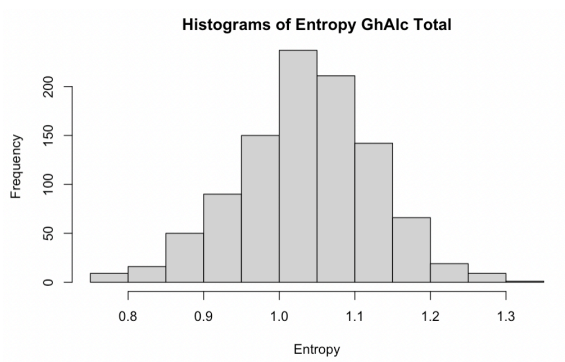

Histograms of Entropy GhBP 5_0
Frequency
0
50
100
150
200
250
0.9
1.0
1.1
1.2
1.3
1.4
1.5
Entropy

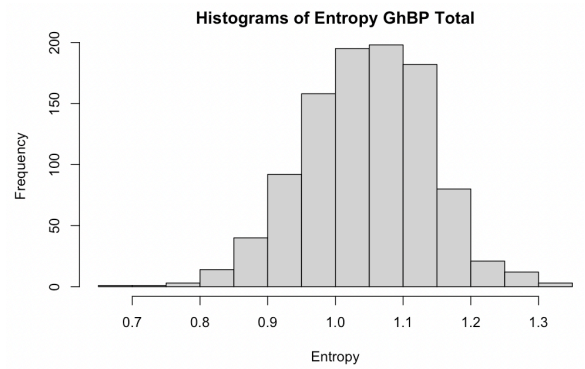

Histograms of Entropy GhBP 5_1
Frequency
0
50
100
150
200
1.0
1.1
1.2
1.3
1.4
1.5
1.6
Entropy

Charts for General Health









Conclusion

Regarding the Shannon entropy approach, the first thing to note is that entropy seems to increase along with general health. Among general health level 1, all entropy distributions peaked around 0.6-0.9. However, by general health level 5, most entropy distributions peaked around 1.2-1.4. This could likely imply that it becomes harder to distinguish interaction effects the worse off someone's general health is.

At general health level 1, the most noticeable groups are groups 1110 and 1111 (people with excellent general health who smoke and drink, along with people with excellent general health who smoke, drink, and have high blood pressure). While these interactions have entropy peaks around 0.6-0.65, their more general cases, along with other interactions around level 1 peak around entropies of 0.8. As interactions 1110 and 1111 both involve people who smoke and drink often, the evidence suggests that drinking and smoking may have a large interaction with people of excellent health. In our hierarchical tree, however, there doesn't seem to show any noticeable difference between interactions 1110 and 1111.

At general health level 2, interaction 2011 stands out from its more basic cases of people who just drink and people who just have high blood pressure, as well as its counterpart of people who drink, have high blood pressure, and smoke. While its more basic cases and related interactions have entropy peaks around 0.9, its peak lies at 0.8. Regarding the hierarchical tree, we note, however, that smoking does not seem to have a large interaction effect. Almost every pair, at its base level, consists of 2 interactions of the same characteristics except that one smokes and one doesn't. This can also be seen at interaction level 1 and level 3.

At general health level 3, no interactions stand out in particular regarding entropy. However, it is noted that interactions 3111 and 3011 are the only smoker vs non-smoker pair that are noticeably farther apart, in different clusters no matter how you slice it.

At general health level 4, most interactions have entropies peaking at around 1.2-1.3, except for interactions 4110 and 4010 (with entropies peaking at around 1.1 each). Here, we also start to notice that the pairs of smokers and non-smokers at level 4 don't tend to be as evident. In fact, we notice that all interactions with high blood pressure seem to be in the left cluster while all those without high blood pressure are in the other, with the exception of interaction 4111. This could likely allude to the idea that blood pressure is a great indicator for heart disease- although our entropy approach seems to say that alcohol consumption is a better indicator.

Finally, at general health level 5 -the poorest health- we note that interaction 5010 in particular has noticeably less entropy than other interactions. While other interactions have entropy peaks around 1.3-1.4, interaction 5010 has entropy ranging from 0.8-0.9. In fact, we find that people, in general, at health level 5 who drink tend to have smaller entropies as the more general case of general health level 5 to alcohol gives us an entropy peak around 1.1. This implies that alcohol, by itself, can be used as a helpful indicator of BMI. In the hierarchical clustering tree, however, this is less evident and we note that the similar interactions farthest apart are 5000 and 5011, again alluding to the notion that alcohol and blood pressure have a noticeable interaction.

Among the entropy and hierarchical clustering approach, we find that interactions involving alcohol tend to have lower entropy than similar interactions within the same health level. Smoking and high blood pressure variables, as interactions, don't seem to be as consistent

as they show promise at some health levels, but not others. Alcohol consumption, however, consistently appears as a good indicator, in both entropy and hierarchical clustering, for BMI.

When the results of the hierarchical clustering tree approach with its entropy counterpart, we notice that the two tell relatively different stories. One reason in particular is because the hierarchical tree is founded on some sort of distance matrix. While the entropy approach allows a user to compare distributions of entropies, via the sampled values, the hierarchical clustering method allows users to compare interactions based on some sort of measurement that might not fully represent the data. For instance, if there are 2 distributions -a unimodal, normal distribution centered at 0 vs a bimodal distribution with two peaks centered at -3 and 3- in such a case, if you were to use the mean as a distance measurement, the hierarchical tree method would tell you that the two distributions are close to the same. With entropy, however, you would be able to see that certain parts of the distribution, and the distribution itself, have lower entropy due to peaks. In such a manner, the entropy approach tends to be more accurate than hierarchical clustering, primarily because hierarchical clustering largely depends on the distance metric.

However, the hierarchical tree method is better suited for comparing values. When looking at multiple distributions via the entropy approach, it may become difficult to determine "how big is big." In specific, it's hard to tell how big a difference in interaction values might be unless you're able to compare differences in values of similar interactions. With hierarchical clustering trees, however, interactions are grouped together based on their "distance" apart from each other and thus, it's easy to see if one interaction is noticeably farther than the others. If you ever want to compare two hierarchical trees together, you could even merge their values. In this way, the hierarchical clustering method is significantly better at comparing entropies or

proportions between two interactions; however, this is on the assumption that a valid distance metric can be found. If not, the entropy approach tends to be more reliable.

As for the reliability of the approaches, there is a concern regarding the sampling from the multinomial distribution. In order to create a distribution of BMI values for each interaction, we've chosen to bootstrap the distributions, using probability vectors for each bin. From each bin, we sample the values where the probability of getting a value within each bin is equal to the bin's proportion in the original vector. This bootstrapping method has been widely accepted and used as a nonparametric way of creating a usable distribution, based on the original- without having to assume a gaussian or student's t distribution. Even when it comes to gaps and peaks, those areas would have higher proportions and the bootstrapping method would remain trustworthy. However, this bootstrapping method fails when the sample is not representative of the original population distribution, in which case, we cannot use it as an estimate. In this way, if there's been some sort of bias in our original sampling then the multinomial bootstrapping becomes more harmful than helpful. Because our dataset happens to be so large, it's unlikely that the sample may be biased because of a small sample size. So, our biggest concern lies in how the sample was initially obtained. If the sample was obtained with bias then not only would our tests and methods fail to help, they'd provide misinformation: something considerably worse. However, as long as we can ensure that our original sample is representative of the population there is no major concern. Since hierarchical clustering works based on distances between values, as seen in the 5 hierarchical clustering trees at each level, we can tell that hierarchical clustering works nevertheless since the distance between distributions or values scales to other values in the tree. However, since hierarchical clustering trees are founded upon the distance metric, we find that bootstrapping becomes more dangerous with the trees. Bootstrapping itself

tends to take parts of information from the data as we add our own random sampling; mixed with the simplification of data in order to measure distance, bootstrapping is more ill-advised in regards to hierarchical trees. We're still able to create trees and compare distances, as seen in the previous trees but the data becomes much more simplified than it would be with the entropy approach.

In conclusion, we find that a prime indicator for BMI, in any interaction, is whether or not a subject heavily drinks alcohol. Since the entropy tends to be lower compared to similar or broader cases, heavy alcohol consumption is a good indicator for BMI. Furthermore, we note that the entropy approach tends to be more accurate but the hierarchical clustering approach tends to do better at comparing values, so long as you can find an accurate distance metric. On top of that, bootstrapping works well with the entropy approach but becomes noticeably more dangerous with the hierarchical clustering approach, assuming the original sample is representative of the population of interest to begin with.