PROJECT 2

Vinh Ton

2023-02-23

ABSTRACT

We have been given the task of exploring the limitations of ANOVA both with simulated data and with provided BRFSS data from the CDC in 2015. We have found that when constant variance and normality assumptions of ANOVA are violated, ANOVA becomes an increasingly unreliable test as magnitude of these assumption violations increase. Interestingly, sample size seems to play less of a role in this.

Then we conducted ANOVA and the Tukey-Kramer test on the BRFSS data. Further analysis revealed, however, that the selected samples violated to some degree both the constant variance and the normality assumptions. Thus, even though ANOVA and Tukey-Kramer supported the claim that the 32 samples did not have the same mean, its reliability is thrown into question by the fact that the necessary assumptions were violated.

INTRODUCTION

We have been introduced to the method of ANOVA and the Tukey-Kramer method of pair-wise comparisons. However, both of these methods rely on two major assumptions: normality, and constant variance, which both often break down in the real world and in the realm of big data. Thus, we are going to test how well both of these do in situations in which their assumptions are violated, using both simulated data and real world health data from the CDC's 2015 BRFSS, separated into 32 samples.

METHODOLOGY

PART I We know that under normality and equal variance assumptions, ANOVA is a good method for testing if multiple samples have the same mean or not. However, what if these assumptions do not hold?

We will test what happens when conducting ANOVA under each of these circumstances, as well as the result of what happens when there are hugely uneven sample sizes. We will do this by using simulated data. For simplicity's sake, we will only create 3 samples per ANOVA test. However, in each situation, we will conduct this 3-sample ANOVA test 100 times, and we will look at the p-values for each set of ANOVA tests as measure of how well ANOVA works under these varying circumstances.

WHAT IF: Unequal Variance?

In this situation, we will generate 3 rnorm samples, each of sample size 100, and with mean 0. However, Samples 1 and 2 will have the same variance of 1. Sample 3 will have a different variance.

In the first dataset, simulation A, the Sample 3 has a variance of 2, or double that of the other 2 samples with variance 1. We will generate this data and conduct the ANOVA test 10000 times, so that we can see how over a large number of tests with varying samples how the conclusion of the ANOVA test is affected.

However, what happens when the variance is significantly more uneven? We conduct the tests again with Sample 3 having a variance = 10, or ten times the variance of the other samples. This is what I will refer to as simulation B.

In the third simulation, we will repeat the process from Simulations A and B. Two of the samples will have constant variance of 1, but now Sample 3 will have a variance of 100. This will be referred to as simulation C.

Simulations A, B, and C were conducted under the conditions that the null hypothesis of ANOVA was true, e.g. all three samples had the same mean. But how well does ANOVA do when the null hypothesis is not true, and the constant variance assumption is violated? We will repeat each Simulation, but this time varying the means.

First, we repeat simulation A, in which two of the samples have the same variance and one doesn't. Sample 2, which has variance equal to Sample 2, will have a mean of 0.2. Sample 3, with unequal variance, will have a mean of 0. Then, we will repeat this process one more time, except Sample 1 and Sample 2 both have a mean of 0, and Sample 3 has a mean of 0.2.

We repeat this process ad nauseum for both Simulation B and Simulation C, to understand how well ANOVA does when the null hypothesis is false under difference degrees of unequal variance and in relation to which sample does not have the equal mean.

Normality Violation Now we are going to test how well ANOVA holds up when the normality assumption is violated. In this set of simulations, we will once again simulate a 3-Sample test 100 times. Samples 1 and 2 will be the standard normal distribution with sample size 10000. Sample 3, however, will be the t-distribution with varying degrees of freedom. This is because we know that as the degrees of freedom approaches infinity, the t-distribution becomes closer to the standard normal distribution. Thus we can see how well ANOVA holds up with varying approximations to normality. We will conduct this simulation three times, where Sample 3 has degrees of freedom 1, 100, 100000. Since a t-distribution always has a mean of 0, this set of tests will be conducted twice: once in which sample 1 and 2 both have a mean of 0, and one in which sample 2 has a mean of 0.2 and sample 1 has a mean of 0.

Then we will conduct a 3-sample test 10000 times in which Sample 3 is an exponential distribution. For all of its strengths, the primary flaw of the t-distribution is that it is symmetrical like the normal distribution. Thus we conduct a test with the exponential distribution to see how well ANOVA holds up in non-normal, and non-symmetrical environments. Again, the sample size for each sample will be 100, and the ANOVA test will be conducted 10000 times. Sample 1 and Sample 2 will both be distributed from the normal distribution with a mean 1 and variance 1. In the first set of simulations, Sample 3, the exponential distribution, will have a mean of 1, e.g. the exponential distribution will have lambda = 1. In the second set, Sample 3 will have a mean of 0.2, e.g. lambda = 5. Finally, in the third set, Sample 1 and Sample 3 will both have a mean of 1, with Sample 2 having a mean of 0.2.

Sample Size Disparity

Finally, how well does ANOVA hold when the sample sizes are large? We will once again conduct sets of 100 3-Sample ANOVA tests. This will be done six times: three times with a true null and varying sample sizes, and three times with a false null and varying sample sizes. Sample 1 and Sample 2 will have the same n = 100 However, Sample 3 will have varying samples sizes n = 10, 1000, 10000, conducted under the equal-means and un-equal means as was conducted previously.

Hierarchical Clustering We have also been asked to discover "Community structures" among K samples. For this simulation, we will create fifteen samples– 5 from the normal distribution with random means and variances, 5 from the t-distribution with random degrees of freedom, and 5 from the exponential distribution with random parameters. In discovering "community," we should expect the respective distributions to be closely related to one another. We should also expect the t-distribution samples to be closer to the normal distribution samples than the exponential distributions. This means we will choose our number of clusters K = 3– as we are sampling from 3 types of distributions.

PART II: KAGGLE DATA We have been asked to divide the data into 32 samples with respect to 5 binary variables. In light of my previous report, I have chosen the variables of: high cholesterol, high blood pressure, smoker status, sex, and difficulty walking or climbing stairs. These are the binary variables which piqued my interest previously, so I will explore them further with relation to BMI in this report as was requested.

Additionally, I will conduct ANOVA and the Tukey-Kramer test with these samples with respect to BMI, assuming that the normality and constant variance assumptions have already been met. However, then we will test whether these assumptions are actually true. For each sample, we will generate a random normal variable assuming that the sample's mean is the mean of the underlying population. The simulation sample size will be the same as the sample size. We will do this twice.

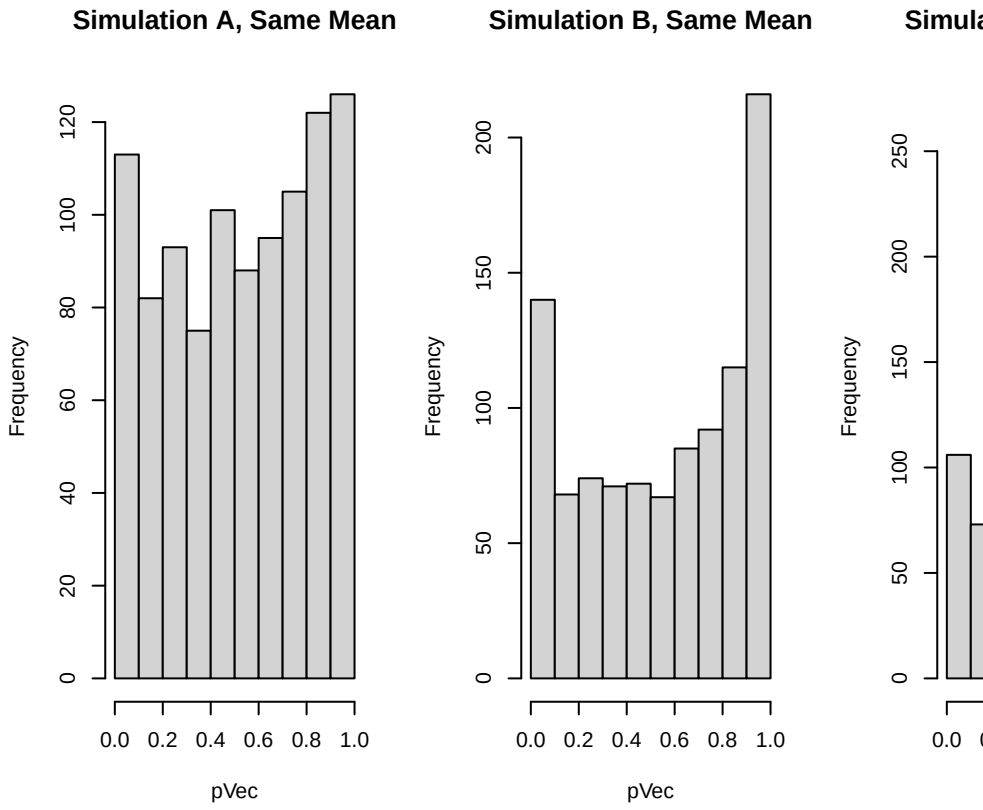
First, we will use the first sample's variance as the variance for every sample. Then we will graph each sample along with its respective normal random variable on the same histogram. Assuming that constant variance is met, we should expect the bar of the sample histogram to be similar in shape and proportion to each respective bar of the normal random variable.

In the second simulation, we will use each sample's variance as the variance of the simulation. This will test the normality assumption. If each sample is normal, we should expect the simulated data's histogram to be proportional to the actual data.

As our final step, we will use hierarchical clustering to discover community structures in the variance of each sample. With all of this information in hand, we will analyze how appropriate ANOVA and the Tukey-Kramer test are appropriate for this data.

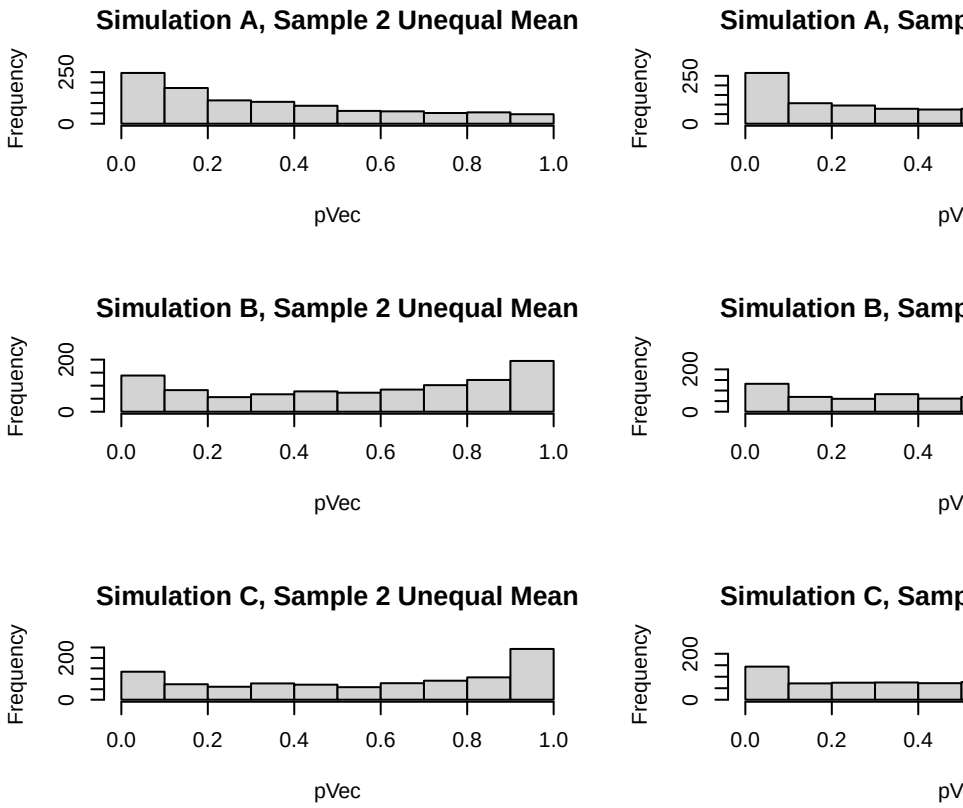
RESULTS

PART I: SIMULATION DATA



Unequal Variances, Equal Means

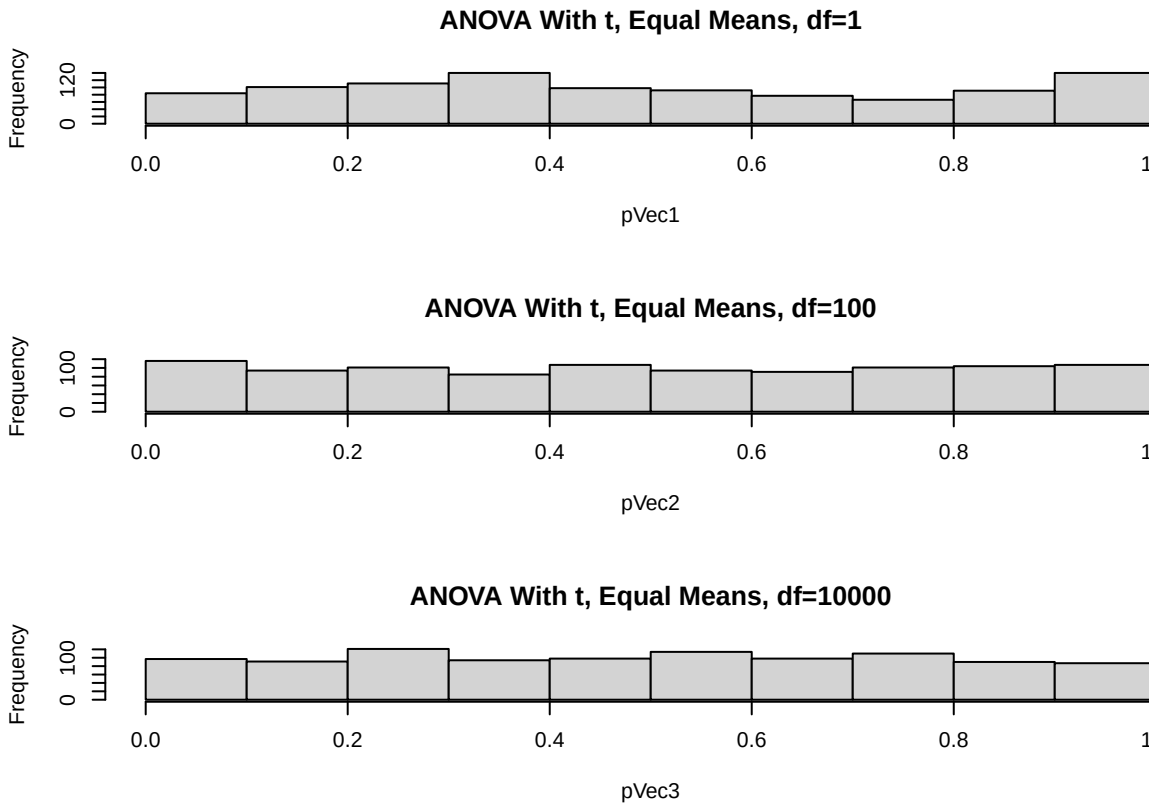
We can see that in our simulated tests, as the magnitiude of difference of variance increases, the distribution of the p-Values actually moves from a roughly uniform distribution to an increasingly bimodal distribution, with the two peaks at the extremes of the p-values, concentrated below a p-value 0.1 and above a p-value of 0.9. Additionally, with increasing magnitiudes of difference of variance, the p-value seems to cluster at much higher values, indicating a greater strength with which ANOVA retains the null. However, it also seems that in the situation with lowest difference of variance, there was also a greater frequency of incorrectly rejecting the null at an alpha of 0.10 compared to the others.



Unequal Variance, Unequal means

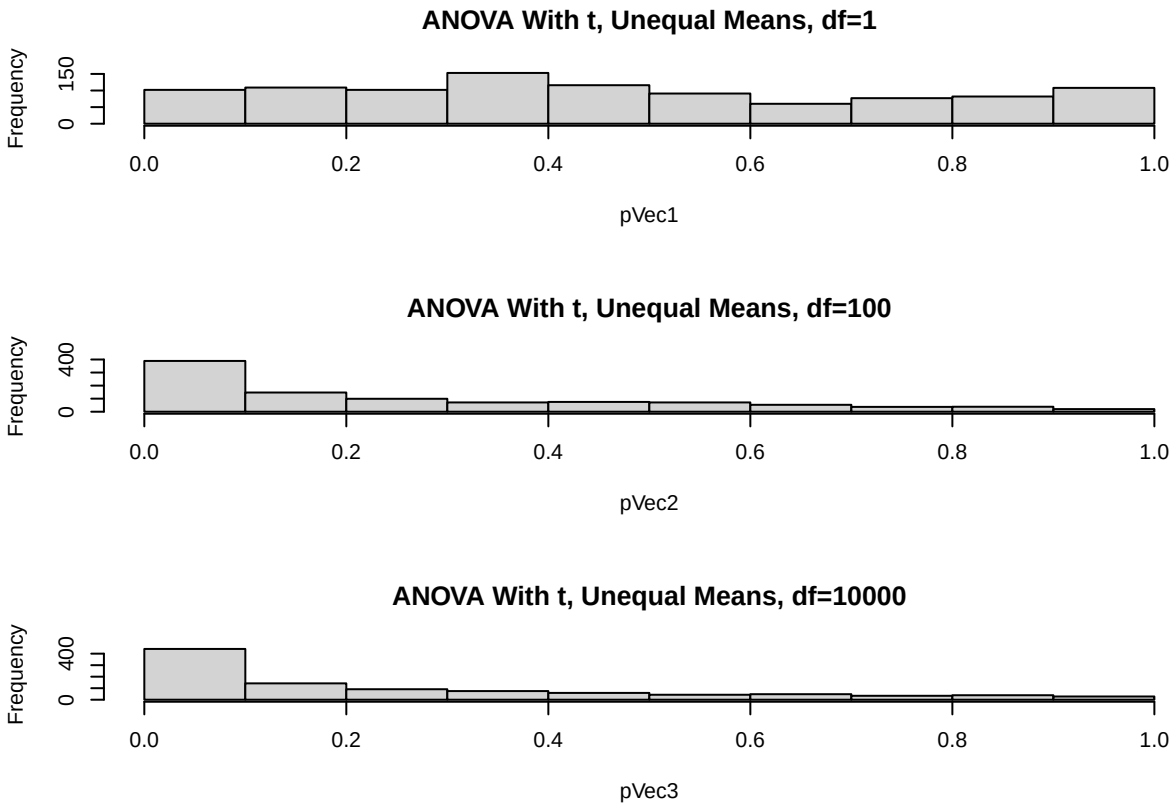
We can see here that in Simulation A, where the difference in variance is least pronounced, ANOVA correctly rejects the null hypothesis at alpha = 0.10 a significant portion of the time, as in both situations the distribution of p-values is distinctly right-skewed.

However, as the magnitude of difference of variance increases, the distribution of the p-values tends to even out more. Instead of being right-skewed, both Simulation B and C see bi-modal distributions at the lows and highs of the p-value. Most significantly, both Simulation B and C incorrectly fail to reject the null hypothesis more often than they reject, suggesting that ANOVA is increasingly inaccurate as difference in variance increases.

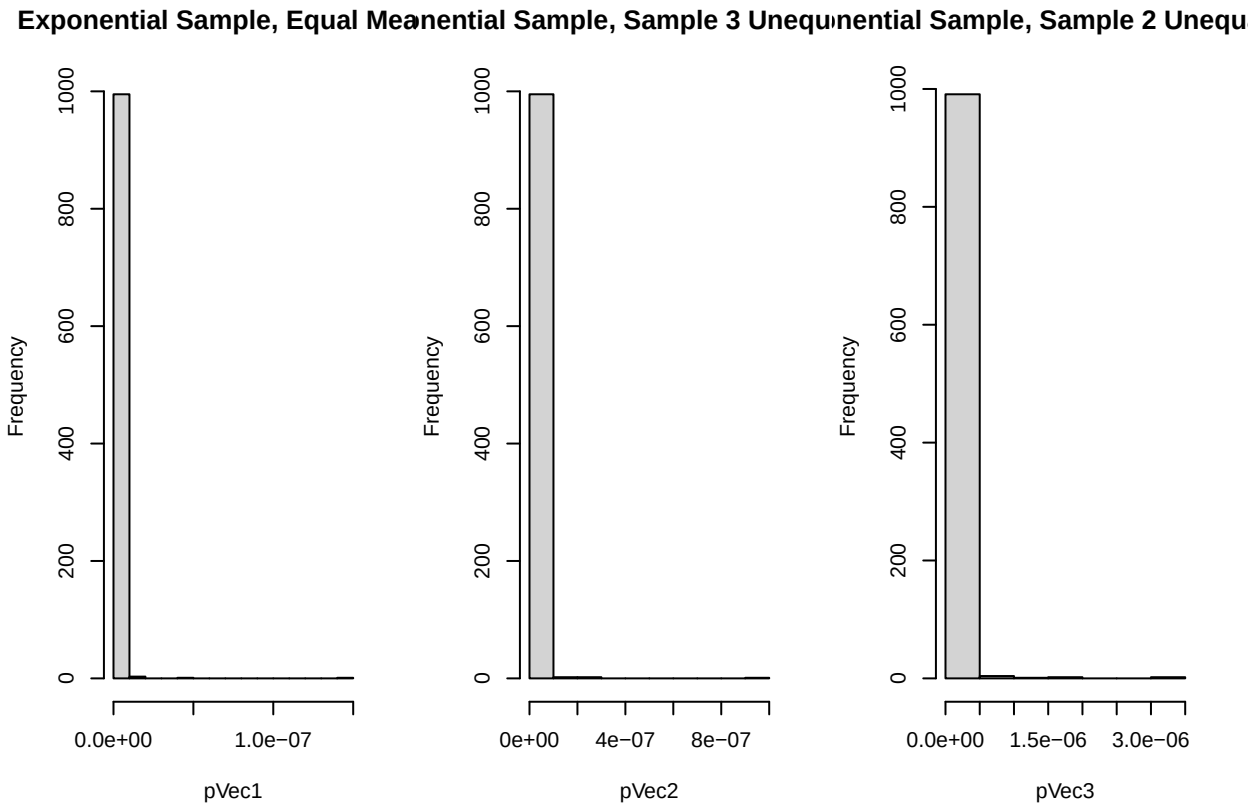


Normality Violation

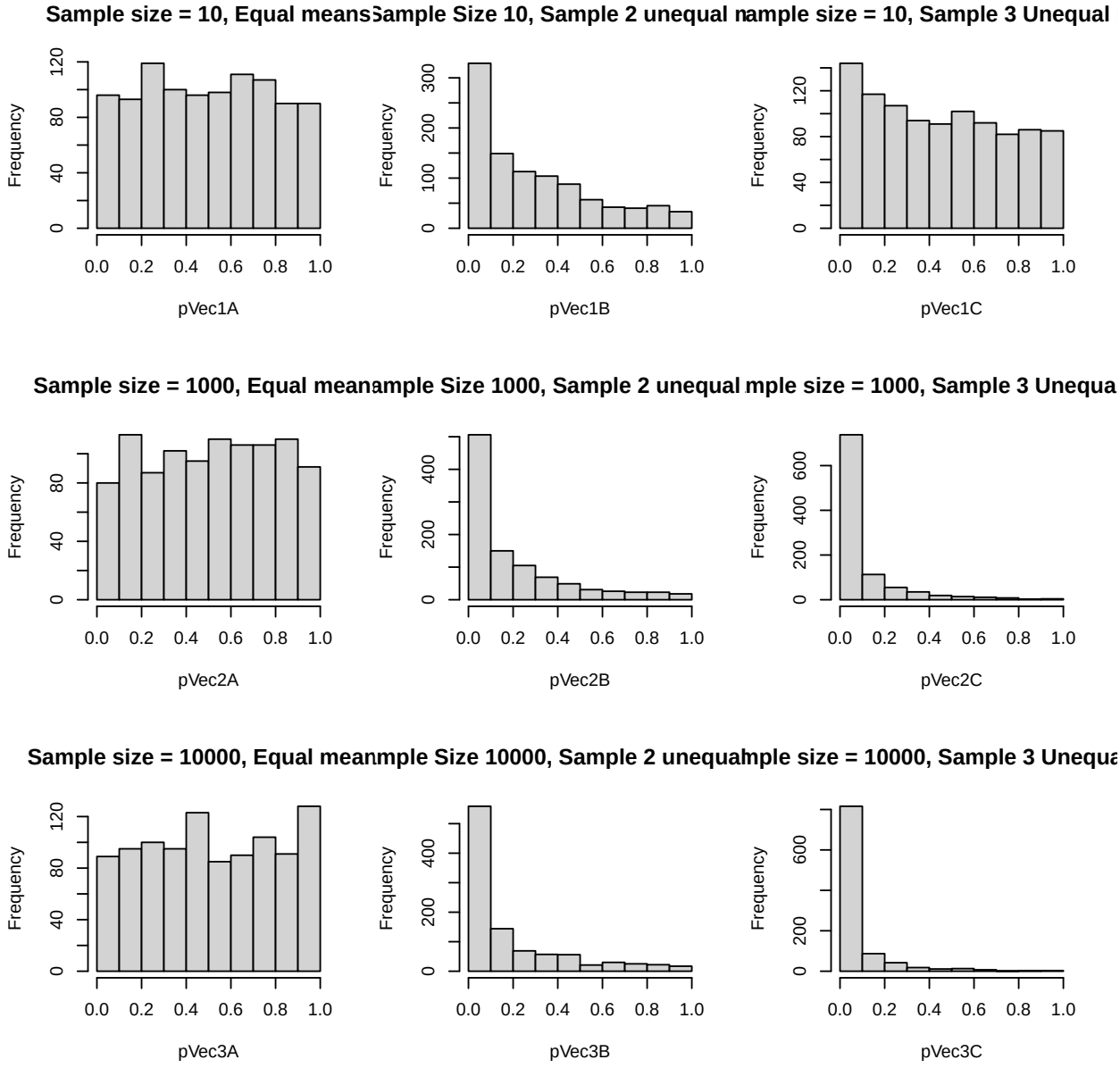
Here we have three histograms of p-values conducted from the ANOVA test in which 1 sample was sampled from a t-distribution, with degrees of freedom 1, 100, and 10000. Interestingly, the distributions seem relatively similarly and uniformly distributed, with the exception of the first simulation having a higher frequency of very high (>0.9) p-values, indicating greater strength to which ANOVA fails to reject.



We have similar histograms from the previous simulation, but now Sample 2 had an unequal mean. Interestingly, it appears that when using the t-distribution with increasing degrees of freedom, the p-values become right-skewed, while with df = 1, it follows a relatively uniform distribution. Interestingly at alpha = 0.10, the second and third simulations correctly reject the null a majority of the time.



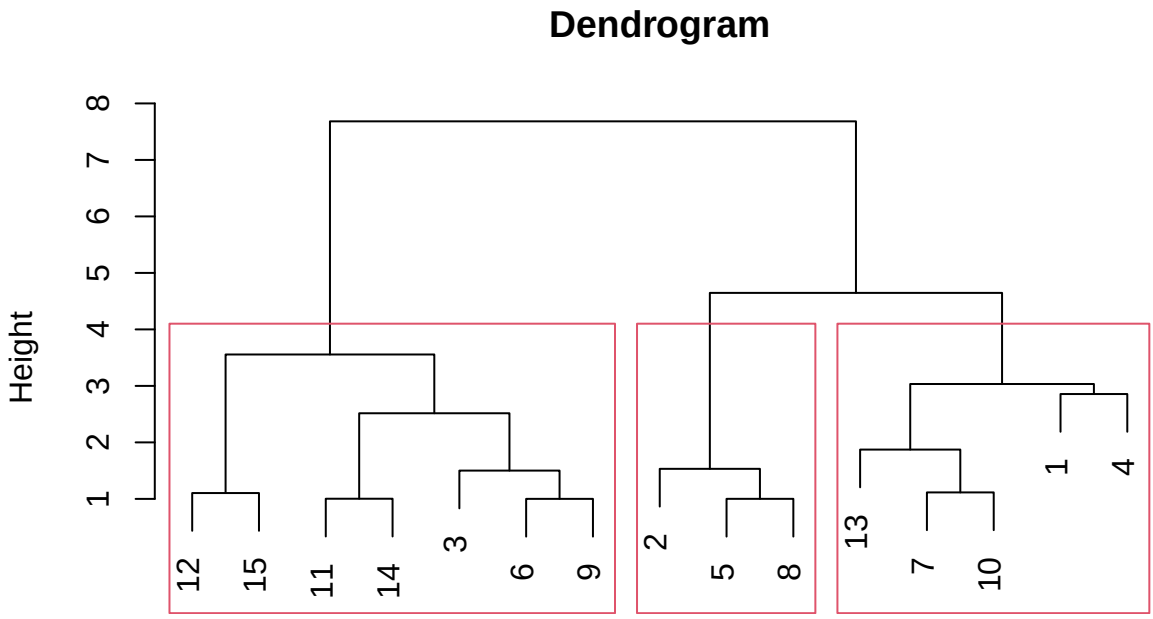
With an exponential distribution, we see that in every simulation, ANOVA rejects the null hypothesis completely at any common significance level, even in situations when the exponential sample has the same mean as the normal samples. In other words, when non-normality is violated, ANOVA is not reliable at all.



Here we can see that when the means are equivalent, then regardless of the difference in magnitude of the sample size, ANOVA manages to correctly retain the null hypothesis a majority of the time.

Interestingly, when the unequal sample size is very large relative to the others, at an alpha = 0.10, the simulations show that ANOVA correctly rejected the null a large majority of the time. The same can be said when the unequal sample size is very small, and the sample with an unequal mean is a sample with the equal sample size.

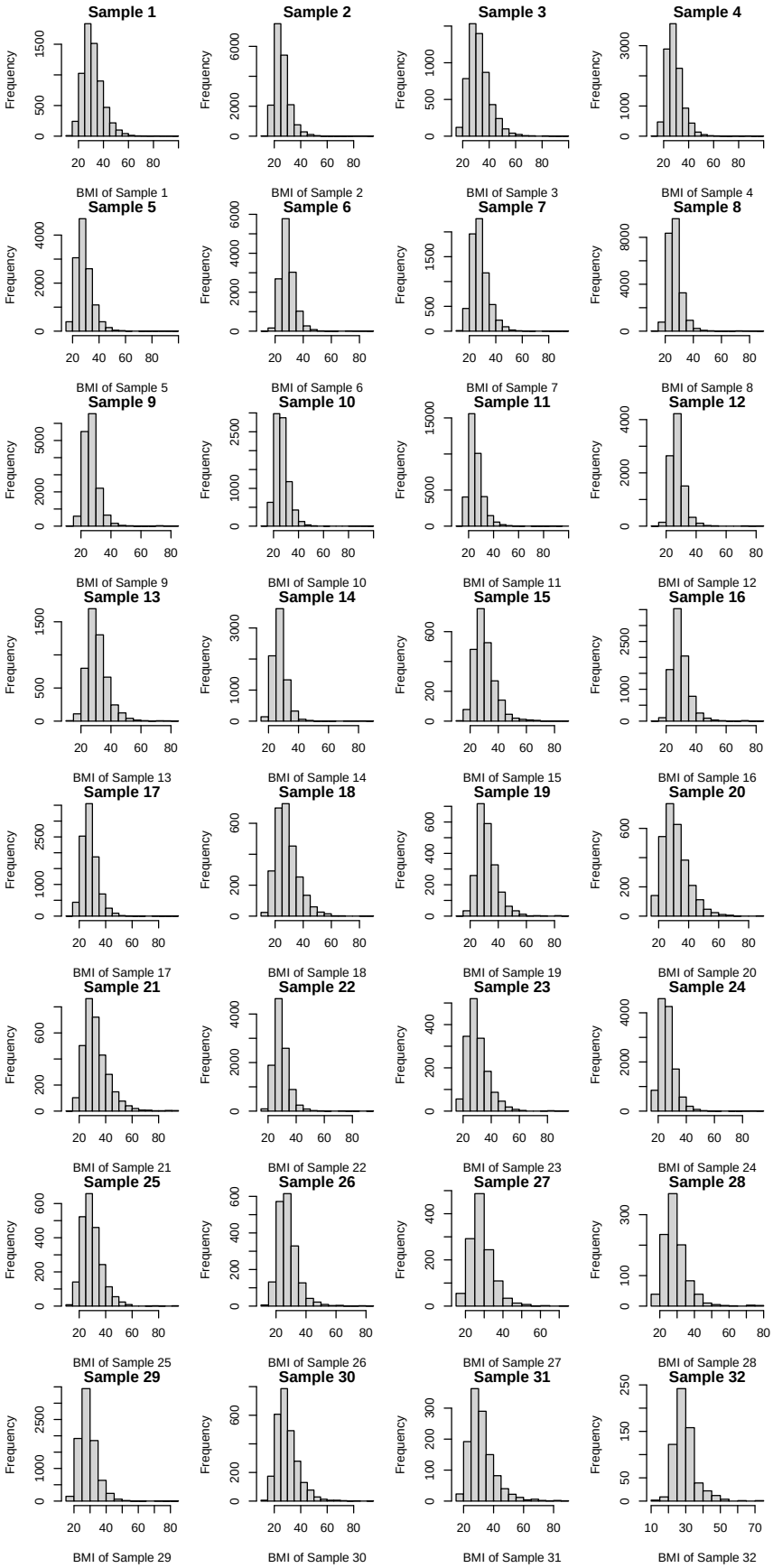
However, when the sample with a very small relative sample size has the unequal mean, the p-values take on a more uniform distribution, and in fact tend to incorrectly retain the null a large majority of the time at an alpha of 0.10. This makes sense, as inherently, samples with small sample sizes tend to have more variability and less consistency.



Clustering

Here we have our hierarchical clustering binary tree. Samples 1-5 are simulated from the normal distribution, Samples 6-10 are simulated from the t distribution, and Samples 11-15 are simulated from the exponential distribution, all with randomly chosen parameters. Here we can see the samples that are "closest" together in mean. For example, samples 2, 5, and 8 are more closely related to each other than to the other samples, and samples 1 and 4 are more closely related to each other than to samples 13, 7, and 10.

PART II: KAGGLE DATA

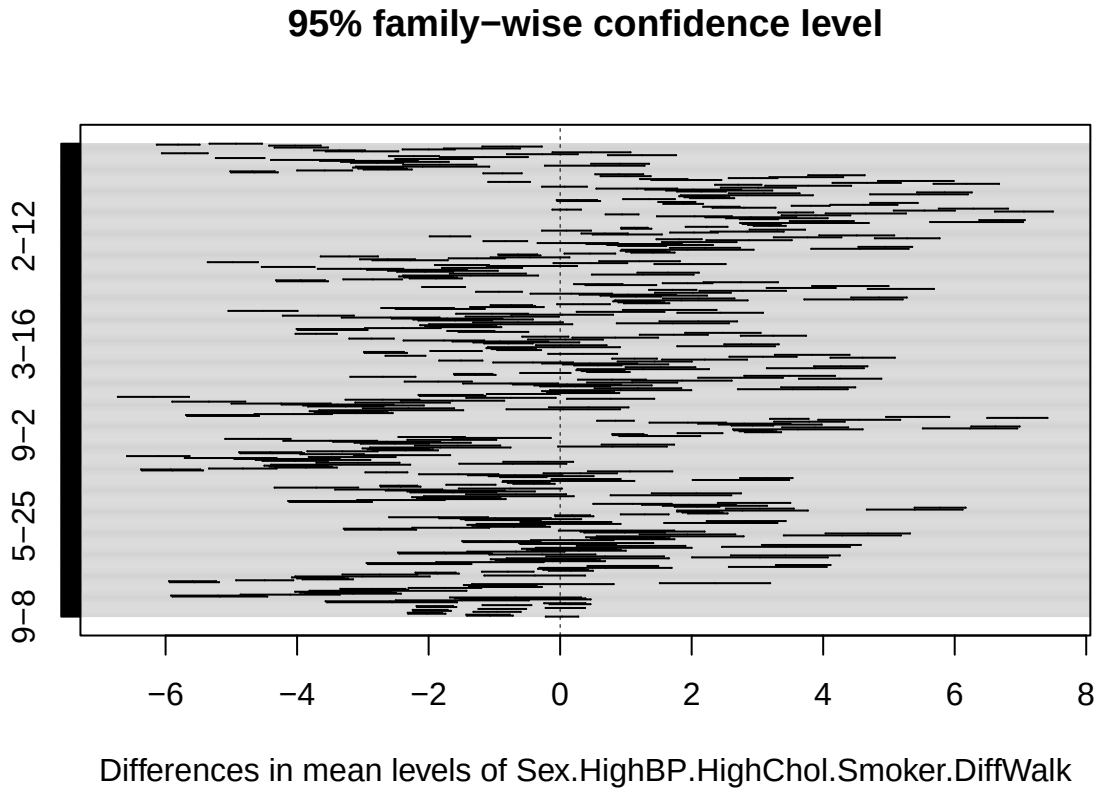


BMI Histograms

Here, we have 32 histograms, comparing the BMI of each sample chosen from the binary variables described in the methodology section. Our parameter of interest is whether each sample has the same mean.

ANOVA & Tukey-Kramer

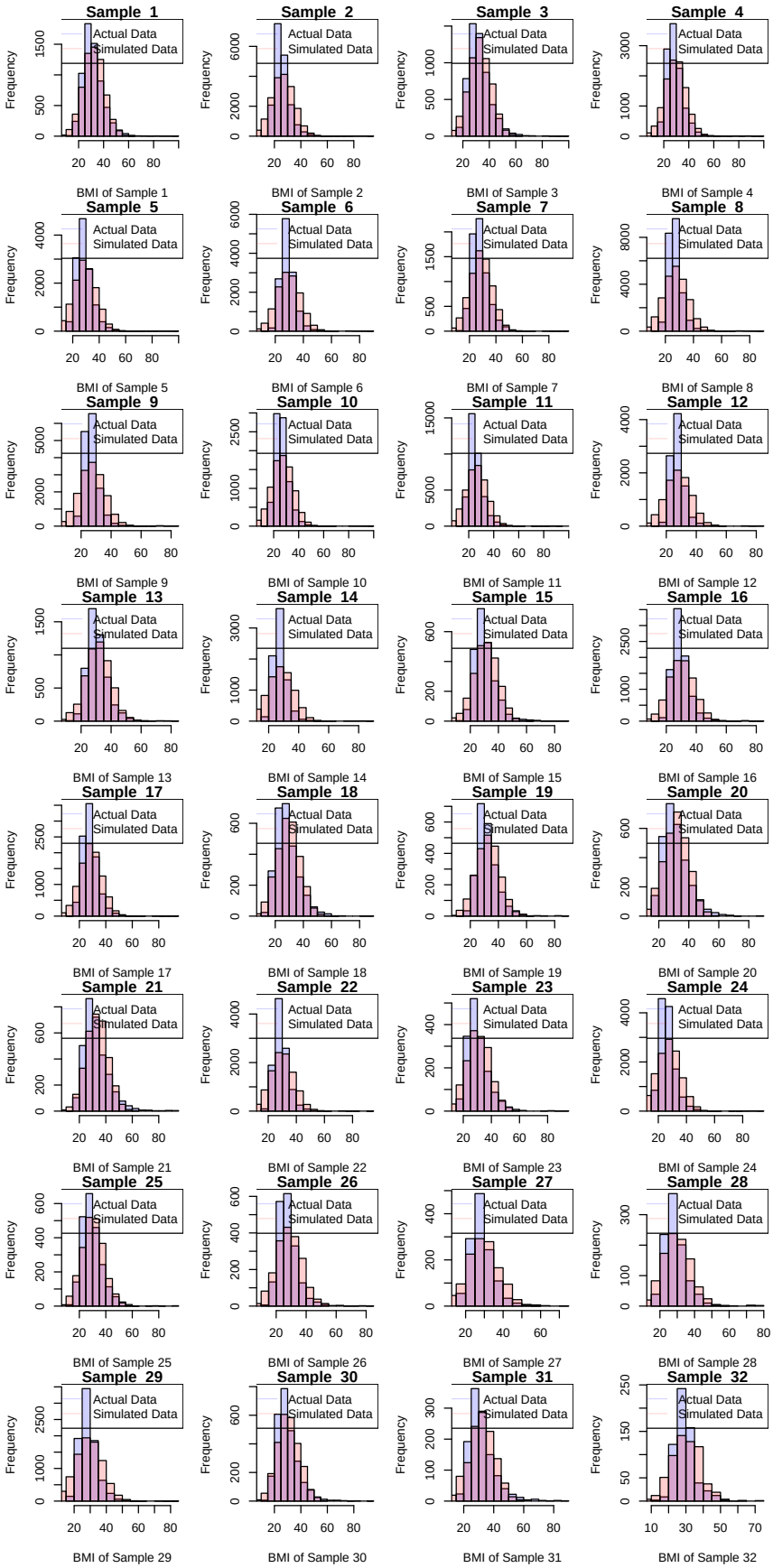
```
## Analysis of Variance Table
##
## Response: BMI
##                                        Df   Sum Sq Mean Sq F value    Pr(>F)
## Sex.HighBP.HighChol.Smoker.DiffWalk     31   861128 27778.3  689.54 < 2.2e-16
## Residuals                           253648 10218262    40.3
##
## Sex.HighBP.HighChol.Smoker.DiffWalk ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



We have conducted both ANOVA and the Tukey-Kramer test assuming that our normality and constant variation assumptions are true. In our ANOVA table, we see our p-value has a very, very, very small value– which means that ANOVA rejects the null hypothesis that all the samples have the same mean with respect to BMI.

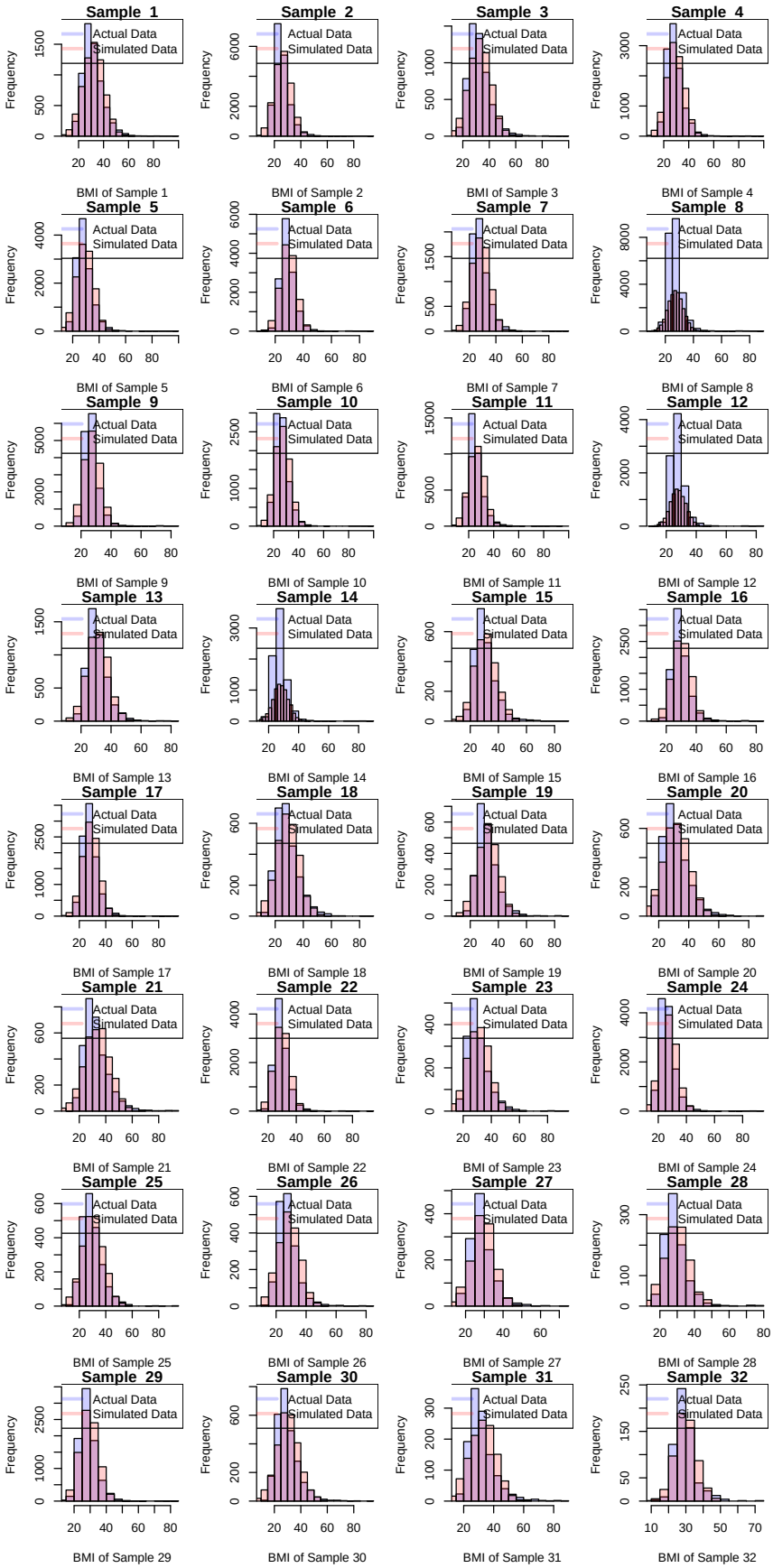
The Tukey-Kramer graph also shows all 32 pair-wise comparisons of the sample means. We can see that there are many pairs that do not include 0 in their confidence interval, meaning that at the 95% confidence level, there are many pairs that have a significant difference in mean.

However, do these assumptions truly hold?

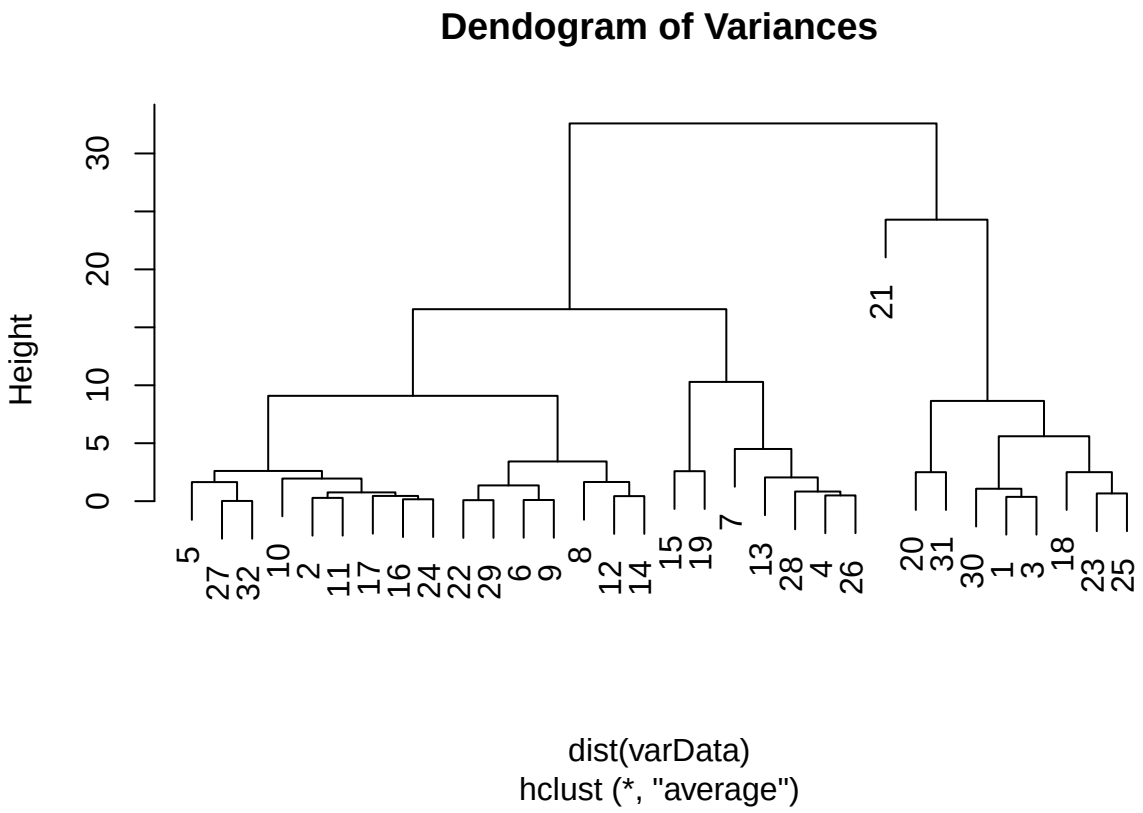


Constant Variance & Normality Testing

Here, the red represents the simulated data, with blue representing the actual data, and purple representing the overlap. If each sample had constant variance and normality, we'd expect most of each graph to be purple. However, we can see that in almost all of these samples, except Samples 1, 3, 18, 19, 20, 21, 25, 30, and 31, the actual data is much more concentrated in the center compared to the simulated normal distribution, indicating there is significantly less data in the tails than we would expect from a normal distribution. Thus we can say that the constant variance assumption is violated.


Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10
Sample 11
Sample 12
Sample 13
Sample 14
Sample 15
Sample 16
Sample 17
Sample 18
Sample 19
Sample 20
Sample 21
Sample 22
Sample 23
Sample 24
Sample 25
Sample 26
Sample 27
Sample 28
Sample 29
Sample 30
Sample 31
Sample 32
Actual Data
Simulated Data
Frequency
BMI of Sample 1
BMI of Sample 2
BMI of Sample 3
BMI of Sample 4
BMI of Sample 5
BMI of Sample 6
BMI of Sample 7
BMI of Sample 8
BMI of Sample 9
BMI of Sample 10
BMI of Sample 11
BMI of Sample 12
BMI of Sample 13
BMI of Sample 14
BMI of Sample 15
BMI of Sample 16
BMI of Sample 17
BMI of Sample 18
BMI of Sample 19
BMI of Sample 20
BMI of Sample 21
BMI of Sample 22
BMI of Sample 23
BMI of Sample 24
BMI of Sample 25
BMI of Sample 26
BMI of Sample 27
BMI of Sample 28
BMI of Sample 29
BMI of Sample 30
BMI of Sample 31
BMI of Sample 32

Here we see a number of samples in which the normality assumption is violated– namely samples 6, 12, and 22, in which the simulated data doesn't follow the actual data at all. Thus we can see that the normality assumption that is required for every sample in ANOVA and the Tukey-Kramer test is violated.



Here, we can see that there are number of samples whose variance is very far from the others. For example, sample 21 is far removed from either major cluster that we see. From the dendogram, we can also see that the cluster on the right-most end has a distance of over 30 units from the cluster on the left, suggesting that all of samples 21, 20, 31, 30, 1, 3, 18, 23, and 25 do not have a similar variance to all other samples.

CONCLUSION

From the simulations and tests carried out in Part I, we can conclude a few things in regards to ANOVA. For one, when either normality or constant variance assumptions are violated, ANOVA becomes increasingly unreliable as those assumptions are violated further. Interestingly, a significant difference in sample size did not seem to make much of a difference in regard to ANOVA tests, with the exception of samples with very small relative sample sizes.

Then we put this into practice in Part II. We conducted ANOVA on 32 samples broken from the BRFSS data, and found that it rejected the null hypothesis that every sample had the same mean. The Tukey-Kramer test supported this, as there were several samples with large differences in means. However, we also found through simulation that the 32 samples did not necessarily have constant variance or normality assumptions fulfilled, meaning that the results of these tests may not be the most reliable if we are to trust our simulations in Part I.

It seems that ANOVA and Tukey-Kramer, which both rely on normality and constant variance assumptions, do not seem to be very trustworthy in our practical example, as neither condition was successfully fulfilled. In future, it'd be interesting to conduct this sort of report again but using different variables in the BRFSS

data to see if those assumptions would hold in different circumstances, and perhaps generalizing those results to real world data.