2012

THE EFFECT OF BASELINE CLUSTER STRATIFICATION ON THE POWER OF PRE-POST ANALYSIS

FENGJIAO HU
Virginia Commonwealth University

THE EFFECT OF BASELINE CLUSTER STRATIFICATION

ON THE POWER OF PRE-POST ANALYSIS

A Thesis submitted in partial fulfillment of the requirements for the degree of Master in Biostatistics at Virginia Commonwealth University.

by

FENGJIAO HU
M.S. in Georgia Southern University, 2010
B.S. in Huazhong Normal University, China, 2008

Director: ROBERT E. JOHNSON
AFFILIATE ASSOCIATE PROFESSOR, DEPT. OF BIOSTATISTICS

Advisor: WEN WAN
ASSISTANT PROFESSOR, DEPT. OF BIOSTATISTICS

Virginia Commonwealth University
Richmond, Virginia
July 18, 2012

Acknowledgement

I would like to thank Dr. Robert E. Johnson for his insightful guidance and constant encouragement. Without his ideas and help I could not have done this thesis. I would also thank Dr. Wen Wan, Dr. Donna McClish, and Dr. D'Arcy P. Mays for their suggestions and willingness to serve on my committee. I also would like to thank my parents whose encouragement motivated me to achieve my goals in education and in life. I applaud the faculty, staff, and fellow graduates of the Department of Biostatistics for their support.

Table of Contents

Page

List of TablesPage

List of Figures

Page

Abstract

THE EFFECT OF BASELINE CLUSTER STRATIFICATION
ON THE POWER OF PRE-POST ANALYSIS

By Fengjiao Hu, M.S.

A Thesis submitted in partial fulfillment of the requirements for the degree of Master in Biostatistics at Virginia Commonwealth University.

Virginia Commonwealth University, 2012

Major Director: Robert E. Johnson
Affiliate Associate Professor, Dept. of Biostatistics

Advisor: Wen Wan
Assistant Professor, Dept. of Biostatistics

The purpose of study is to check whether the power of detecting the effect of intervention versus control in a pre- and post-study can be increased by using a stratified randomized controlled design. A stratified randomized controlled design with two study arms and two time points, where strata are determined by clustering on baseline outcomes of the primary measure, is considered. A modified hierarchical clustering algorithm is developed which guarantees optimality as well as requiring each cluster to have at least one subject per study arm. The power is calculated based on simulated bivariate normal

distributed primary measures with mixture normal distributed baseline outcomes. The simulation shows that the power of this approach can be increased compared with using a completely randomized controlled study with no stratification. The difference of the power between with stratification and without stratification increases as the sample size increases or as the correlation of the pre- and post-measures decreases.

CHAPTER 1 Introduction

A completely randomized controlled study with two study arms and two time points (pre- and post-intervention) is considered for the power to detect differences between intervention versus control effects. The power may be increased by controlling the variance between subjects or units of randomization. Here a pre-post design can help control the within subject variance, since each subject uses its own as control (Park and Johnson (2005)). The variance may be further controlled by stratification into blocks prior to randomization, where the blocks are relatively homogeneous on the baseline primary measures (Park and Johnson (2005)). Although Park and Johnson (2006) pointed out that treating the baseline as a covariable provided maximal control of the variance, with or without pair-matching, we are interested in understanding the effect of baseline clustering when considering an analysis of pre-post differences.

In order to stratify subjects into relatively homogeneous blocks, partitions of the subjects clustered on the baseline values may be formed with the partition that minimizes the root mean squared error (RMSE) of the baseline values.

Since subjects within each block will be randomized to study arms, the subjects' baseline values must be clustered so that not only the minimal RMSE is achieved, but also each cluster must have at least the same number of subjects as arms of the study. Here we have two study arms, so each cluster must have at least two subjects to randomize.

Hierarchical clustering methods, such as *k-means*, are often used to identify clusters. The *k-means* procedure does not always find the optimal clustering for a set number of clusters since the procedure employs a random path. And these methods cannot easily constrain the result to have the minimum number of subjects for each cluster. Unconstrained, the solution may result in one or more clusters having less than the required number of subjects. Fusing neighboring clusters can resolve this, but may result in a non-optimal set of clusters (Park and Johnson (2004)).

In order to get an optimal set of clusters with constraints on number of subjects, a modified hierarchical clustering algorithm is developed here for identifying clusters of univariate baseline outcome data. It not only guarantees optimality, but also places the desired constraint on the minimum number of subjects in each cluster.

The power is calculated based on simulated bivariate pre-post measures where the measure at each time point is distributed as a mixture of normal distributions. The data are simulated based on the correlation between pre-post measurements, the proportion of baseline data that comes from the first distribution of the mixture normal distribution, the ratio of the two variances of the mixture normal distribution and the sample size. The power of testing the null hypothesis of no difference in intervention versus control effects using pre- and post-intervention design is compared using the cluster stratification versus no stratification.

CHAPTER 2 Background

Literature Review

Pair-wise matching on ordered baseline means can help reduce the variance of pre-post differences when the correlation between the pre-post measures is small, under the assumption of equal sample sizes (Park and Johnson (2005)). Park and Johnson (2006) pointed out that treating the baseline means as a covariable provided maximal control of the variance, with or without pair-matching. Pair-matching reduced the variance when using a posttest only analysis or pre-post differences. They conducted six design and analysis methods for controlling the variance of the differences in intervention versus control effects. Both with and without matching were considered, and for each case, three analysis designs were considered: (1) posttest only, (2) posttest analysis adjusting for baseline as a covariate, and (3) pre-post differences.

Park and Johnson (2004) conjectured that matching based on how the baseline values 'clustered' is better than pairing the baseline values after sorting, since pairing would cause dissimilar baseline values be forced to be in the same match group. Actually, clustering is a powerful tool in finding subsets or clusters which are homogeneous and/or well separated (Aloise, et al. (2009)). The most widely used method for clustering is the *k-means* method which partitions data into *k* clusters in which each observation belongs to the cluster with the nearest mean (Jain (2010)).

However, the problem is NP-hard (non-deterministic polynomial-time hard) in a general Euclidean space, even when the number of clusters k=2 (Aloise, et al. (2009), Dasgupta and Freund (2009)). There have been many attempts to provide an algorithm that solves this problem.

The term *k-means* was first used by MacQueen (1967). It described a simple and easy way to classify a given data set through a certain number of clusters (assume *k* clusters) fixed *a priori*. The *k-means* procedure consists of simply starting with *k* groups, each of which consists of a single random point, and thereafter adding each new point to the group whose mean the new point is nearest, which is determined by the distance calculated with Lebesgue measure. After a point is added to a group, the mean, or centroid, of that group is adjusted in order to take account of the new point. Thus at each stage the *k-means* are, in fact, the means of the clusters they represent (hence the term *k-means*). We may notice that the *k* centroids change their location step by step until no more changes are done. If the process converges, the final cluster seed will represent the centroids.

Another method called "nearest centroid sorting" (Anderberg (1973)) was provided after the *k-means* algorithm. The procedure is an improved *k-means* algorithm, it has the same process but the new point is added to the group when it achieves the minimum sum of squared Euclidean distances from each entity to the centroid of the cluster to which it belongs. The centroid of that group is adjusted after the new point is added.

Dasgupta and Freund (2009) also provided a definition of the *k-means* problem. Assume the input is a set of *n* vectors $S = \{x_1, \ldots, x_n\}, x_i \in R^D$. The output is a set of *k*

vectors $R = \{\mu_1, \ldots, \mu_k\}$, $\mu_i \in R^D$, where k is much smaller than n. The set R is called a codebook. R is a good codebook for S if for most $x \in S$ there is a representative $r \in R$ such that the Euclidean distance between x and r is small. The average quantization error of R with respect to S is

$$Q(R,S) = E\left[\min_{1 \le j \le k} \|X - \mu_j\|^2\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n} \min_{1 \le j \le k} \|x_i - \mu_j\|^2$$

where $\|\cdot\|$ denotes Euclidean norm and the expectation is over X drawn uniformly at random from S. The goal is to construct a codebook R with a small average quantization error. The k-optimal set of centers is defined to be the codebook R of size k for which $Q(R,S)$ is minimized; the task of finding such a codebook is sometimes called the *k-means* problem.

The standard iterative *k-means* algorithm (Lloyd (1982)) is a widely used heuristic solution. The algorithm iteratively calculates the within-cluster sum of squared distances, modifies group membership of each point to reduce the within-cluster sum of squared distance, and computes new cluster centers until local convergence is achieved.

Although it can be proved that the procedure will always terminate, the *k-means* algorithm is not, in general, able to find the optimal solution. The solution is not far from optimal (Dasgupta and Freund (2009)); that is, the set of k clusters that achieves minimum RMSE. The algorithm is also significantly sensitive to the initial randomly selected cluster centers, which causes the algorithm to be neither optimal nor repeatable. The repeatability

of the algorithm is very important since the subject should fall into the same group to receive the intervention regardless of how many times the clustering is applied. Often one restarts the procedure a number of times to mitigate the problem of optimality. It has been recommended to use an *all partition* method where all possible partitions meeting the constraints are considered in order to find the optimal partition; however, this is feasible only for small *N*, as the number of cluster partitions to be considered increases almost exponentially with *N* (Park and Johnson (2004)). This may lead to a substantial increase in runtime.

The importance of initial seed selection was also demonstrated by Milligan (1980). He compared fifteen clustering algorithms and found out that the *k-means* algorithm produced excellent repeatability, or recovery of cluster structure, when the starting seeds were obtained from the group average method or when valid *a priori* information existed. Despite this advantage, all *k-means* algorithms produced recovery values which were significantly worse than others in the error-free condition when random starting seeds were used. So the starting partition must be close to the final solution if the *k-means* algorithm is to be expected to give good recovery.

Since the standard iterative *k-means* algorithm does not guarantee optimality and repeatability, Wang and Song (2011) developed a dynamic programming "CKmeans" algorithm to find optimal one-dimensional clustering. The algorithm first arranges the baseline measures in a non-descending order and then calculates the minimum Euclidean sums of squares of within-cluster distances from each subject to its corresponding cluster mean iteratively by adding one observation at a time. Simulation is used to prove that their

method not only guarantees optimality but also has a fast runtime, especially as the number of clusters is big. This algorithm cannot guarantee each cluster has the same number of subjects as number of arms.

Summary

Matching based on how the baseline values 'cluster' can help reduce the variance of pre-post differences. In general, *k-means* methods are used to cluster subjects at the baseline, but these methods do not guarantee optimality and repeatability. The improved "CKmeans" algorithm leads to an optimal and repeatable unconstrained solution, but it cannot guarantee at least two subjects in each cluster. In this thesis, we develop a modification to "CKmeans" that achieves an optimal solution when applied to univariate baseline outcome data and guarantees each cluster has at least two subjects to randomize into two study arms. Power is calculated for a stratified randomized controlled design with strata defined as clusters generated by this improved clustering algorithm. The calculated power is compared to the power achieved using a completely randomized controlled study with no stratification.

CHAPTER 3 Methods

Clustering on baseline

In order to cluster on baseline with minimum RMSE, and with the restriction that every cluster has at least two subjects, we provide a new algorithm which is an improvement of "CKmeans.1d.dp" algorithm (Wang and Song (2011)). Since minimizing Euclidean sums of squares (ESS) is equivalent to minimizing RMSE when dealing with univariate data, we calculate RMSE by getting the square root of the ratio of *withinss* and degrees of freedom, where *withinss* represents the minimum Euclidean sums of squares of within-cluster distances from each subject to its corresponding cluster mean, and the degrees of freedom is the difference between total subjects and number of clusters.

Let $x_1,...,x_n$ be the non-descending sorted baseline values of n subjects. We seek the cluster partition that minimizes RMSE. First of all, we consider clustering i subjects into m clusters in general with minimum *withinss*. We record the corresponding minimum *withinss* in entry $D[i,m]$ of an $n+1$ by $\left\lfloor \frac{n}{2} \right\rfloor + 1$ matrix D, since n subjects can be clustered into at most $\left\lfloor \frac{n}{2} \right\rfloor$ clusters based on the constraint that every cluster has at least two subjects, where $\left\lfloor \frac{n}{2} \right\rfloor$ means the integer part of $\frac{n}{2}$. Since the last row of the matrix D means clustering all the subjects. Thus the minimum value of the last row in matrix D

corresponds to the cluster partition with the smallest *withinss* value, the solution to the original problem. Let j be the index of the smallest number in cluster m in an optimal solution to $D[i,m]$, and j must be $2m-1 \le j \le i-1$. Here j cannot less than $2m-1$ because it must have at least two subjects in each cluster, that is at least $2m-2$ totally for the first $m-1$ clusters. It is evident that $D[j-1,m-1]$ must be the optimal *withinss* for the first $j-1$ points in $m-1$ clusters, for otherwise one would have a better solution to $D[i,m]$. This establishes the optimal substructure for dynamic programming and leads to the recurrence equation

$$D[i,m] = \min_{2m-1 \le j \le i-1} \left\{ D[j-1,m-1] + d\left(x_j,...,x_i\right) \right\},\ 1 \le i \le n, 1 \le m \le k$$

where $d\left(x_j,...,x_i\right)$ is the sum of squared distances from $x_j,...,x_i$ to their mean. The matrix is initialized as $D[i,m]=0$ when $m=0$ and $i=0$; $D[i,m]=\infty$ when $m=0$ xor $i=0$. Using the above recurrence, we can obtain $D[n,m]$ the minimum *withinss* if all *n* numbers are clustered into *m* groups, with minimum $RMSE[n,m] = \sqrt{\frac{D[n,m]}{n-m}}$.

In order to make the program more efficient, it is suggested in Wang and Song (2011) that one compute $d\left(x_j,...,x_i\right)$ in the recurrence; $d\left(x_j,...,x_i\right)$ can be computed progressively based on $d\left(x_{j+1},...,x_i\right)$. Using a general index from 1 to *i*, we iteratively compute

$$d\left(x_1,\ldots,x_i\right) = d\left(x_1,\ldots,x_{i-1}\right) + \frac{i-1}{i}\left(x_i-\mu_{i-1}\right)^2, \text{ with } \mu_i = \frac{x_i+(i-1)\mu_{i-1}}{i},$$

where μ_i is the mean of the first *i* elements. To find a clustering of data with minimum *withinss* of $D[n,k]$, an auxiliary *n* by *k* matrix *B* is defined to record the index of the smallest number in cluster *m*

$$B[i,m] = \underset{2m-1 \le j \le i-1}{\arg\min} \left\{ D[j-1, m-1] + d\left(x_j, \ldots, x_i\right) \right\},\ 1 \le i \le n, 1 \le m \le k$$

Then we backtrack from $B[n,k]$ to obtain the starting and ending indices for all clusters and generate an optimal solution to the *k-means* problem.

Randomization

We randomize the baseline measures within each cluster to intervention and control groups. If the number of measures is even in the cluster, then half of them will be assigned to intervention and half to control. Since some of the clusters have an odd number of measures (if any cluster has an odd number of measures, then there must be an even number of such clusters). If the number of measures is odd, then the first odd number cluster will be assigned one more subject to intervention group and the second odd number cluster will be assigned one more subject to control group. If more clusters have odd numbers, then this process is repeated.

The randomization procedure is using a probability-tree method to do simple random sampling for each cluster. Let *k* be the number of subjects that should be assigned into intervention, and *N* be the total subjects in the cluster. Generate a random variable from uniform distribution $U(0,1)$, and then compare this random variable to the ratio $\frac{k}{N}$.

If random variable is not greater than $\frac{k}{N}$, then we assign this subject to intervention group, or else assign this subject to control group. The algorithm continues using $N = N - 1$ and $k = k - 1$ if the subject is assigned into intervention group, or $N = N - 1$ and $k = k$ if the subject is assigned into the control group.

Modeling the treatment effect

After we assign the subjects into intervention and control groups, a pre-set treatment effect is added to the post measures if the subject is in the intervention group. The differences between pre-post measures are calculated for each subject and a linear model, given by $d_i = \beta_0 + \beta_1 t_i + \sum_1^k \beta_{2j} x_j + \varepsilon_i$, is used to model the effect of the i^{th} subject in the intervention or control group. The difference between the intervention and control is $d_i^T - d_i^C = \beta_1 + (\varepsilon_i^T - \varepsilon_i^C)$, where d_i^T and d_i^C are the effects under the intervention and control group, respectively, $t_i = 1$ means the i^{th} subject is assigned into intervention group, and $t_i = 0$ means control group. Also $x_j = 1$ means the i^{th} subject is assigned into the j^{th} cluster and k is the total number of clusters used to achieve minimum sum of squared Euclidean distances from each entity to the centroid of the cluster to which it belongs. We assume the residuals $\varepsilon_i \sim N(0, \sigma^2)$ as is common. Although we know here the residuals are not normally distributed, since the baseline data are not normal, but mixture normal distributed.

Simulation

The bivariate normally distributed pre-post measures of *n* subjects are simulated. For each subject, pre- and post-measures are $(X_i, Y_i) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho), i = 1, 2, \ldots, n$, with equal means $\mu_1 = \mu_2$, equal variances $\sigma_1^2 = \sigma_2^2$, and ρ is correlation between pre-measure X_i and post-measure Y_i. The baseline measures follow a mixture normal distribution $X_1, \ldots, X_n \sim h(x_i) = pf(x_i) + (1-p)g(x_i)$ where $f(x_i) \sim N(\eta_1, \theta_1^2)$ and $g(x_i) \sim N(\eta_2, \theta_2^2)$ are independent. The fixed number $p \in [0,1]$ is the proportion of pre- and post-measures that follow the distribution $f(x_i) \sim N(\eta_1, \theta_1^2)$. When $p = 0$ or 1, the distribution of baseline measures can be changed into a unimodal normal distribution.

In order to simplify, set $\theta_2^2 = 1$, and let $\tau^2 = \theta_1^2$ (or $\tau^2 = \theta_1^2 / \theta_2^2$) and $\Delta = \eta_1 - \eta_2$. By setting the mean of baseline measures $E(X_i) = 0$, we have $p\eta_1 + (1-p)\eta_2 = 0$. Then we can calculate $\eta_1 = (1-p)\Delta$ and $\eta_2 = -p\Delta$. Let *m* be the number of subjects that have pre-measures from the distribution $f(x_i) \sim N(\eta_1, \theta_1^2)$; then we can assume

$m \sim Binomial(n, p)$ and

$X_i \sim N((1-p)\Delta, \tau^2)$, with $(X_i, Y_i) \sim N((1-p)\Delta, (1-p)\Delta, \tau^2, \tau^2, \rho)$ for $i = 1, 2, \ldots, m$;

$X_i \sim N(-p\Delta, 1)$ with $(X_i, Y_i) \sim N(-p\Delta, -p\Delta, 1, 1, \rho)$ for $i = m+1, m+2, \ldots, n$.

The number of simulations is dependent on the probability *pr* of the result of a random process; here the result is the significant evidence that the effects in interventions and controls are different, that is presented as *p*-values less than 0.05. Since the result of a random process follows a Bernoulli distribution with probability *pr*, then the variance of

the result occuring is $pr(1-pr)$. We usually use $pr=0.5$ as default if we have no knowledge of the value of *pr*, but if we know the null hypothesis that there is no difference between intervention and control is not true, then we set up $pr=0.05$ since we know the *p*-value should be less than 0.05.

And also we could define $\theta=0.95$, where it means we wish to estimate *pr* with at least 95% confidence. We need to find a standard normal variable *z* such that $P(-z \le Z \le z)=0.95$ that is $z=\Phi^{-1}\left(\frac{1+\theta}{2}\right)=1.96$, where Φ^{-1} is a quantile function. Then the number of simulations should be no smaller than $N = pr(1-pr)\left(\frac{z}{E}\right)^2$, where $E=0.005$ for the maximum error in estimate.

In all, 7299 simulations were conducted under the null hypothesis that there was no treatment difference between the intervention and control groups, while 38415 simulations were conducted under the alternative hypothesis that treatment effects were different for intervention and control groups so that estimate will be within 0.005 of the true value within at least 95% confidence.

The following describes the set of parameters used in the simulations. The proportion is defined as the first normal distributed data weighted in the two mixed normal distribution, proportion sets to 0.1, 0.3, 0.5, 0.7, and 0.9. The variance ratio between the first normal and the second normal distribution is denoted as τ. In addition, we always set the variance for the second normal distribution as 1. The ranges of τ values are 1, 2 and 3.

The means difference between the first and the second normal distributions is denoted as δ. The δ values are 0, 1, 3, and 6. The correlations between baseline and post-baseline measures are set to low ($\rho = 0.2$) and high ($\rho = 0.8$) respectively. The sample size is evenly allocated to the intervention and control groups and the total sample size of entire sample is chosen as 20, 40 and 60. The treatment effect range is from 0.1 to 0.5 increasing by 0.1.

Evaluation

A linear model is used to model the effect of the i^{th} subject in the intervention or control group, $d_i = \beta_0 + \beta_1 t_i + \sum_1^k \beta_{2j} x_j + \varepsilon_i$, with the difference between the intervention and control is $d_i^T - d_i^C = \beta_1 + (\varepsilon_i^T - \varepsilon_i^C)$. Considering cluster as a fixed effect, to test the null hypothesis H_0 : there is no treatment effect, we test $H_0 : \beta_1 = 0$. For every simulation, the data are used to test the null hypothesis and a *p*-value is calculated. The number of *p*-values less than 0.05 divided by total number of simulations is the estimated probability that the *p*-value less than 0.05; that is, the power for a set of parameters when treatment effect is greater than 0, and otherwise it is type I error when no treatment effect.

CHAPTER 4 Results

All power calculations with or without clustering were simulated based on the parameters mentioned in the methods section, but the only results presented here have weighted proportions p for the first distribution of 0.3 and 0.5; the range of variance ratio τ values were 1 and 3; the means differences δ between the first and the second normal distributions were 0, 3, and 6; the correlations ρ between baseline and post-baseline measures were set to low (=0.2) and high (=0.8); the sample sizes n were 60, 40, and 20, and the treatment effect d range was 0, 0.3 and 0.5. And the presented results here can represent all the power calculation we conducted.

Table 1 presents the power with or without clustering for different values of ρ, δ and d when fixing other parameters p=0.3, τ^2=1 and n=40. From the table, the power with or without clustering are very close to 0.05 when there is no treatment effect (d=0), and the bold numbers show that the power increases with clustering, and the rate of increase is higher for $\rho = 0.2$ compared to $\rho = 0.8$.

delta	rho	d=0 cluster	d=0 no-cluster	d=0.3 cluster	d=0.3 no-cluster	d=0.5 cluster	d=0.5 no-cluster
0	0.2	0.051	0.048	**0.146**	0.114	**0.327**	0.231
3	0.2	0.054	0.055	**0.131**	0.114	**0.275**	0.230
6	0.2	0.046	0.049	**0.148**	0.111	**0.318**	0.235
0	0.8	0.050	0.050	**0.319**	0.309	**0.697**	0.682
3	0.8	0.045	0.052	0.309	0.313	0.677	0.683
6	0.8	0.051	0.053	**0.322**	0.308	**0.698**	0.681

Table1: Power for p=0.3, τ^2=1 and n=40

In order to present the effects of the parameters on the power, plots of power versus d for different values of δ, and power versus δ for different d are plotted and presented in Appendix A. In Figure 1, the three plots on the left side are plots of power versus treatment effect d where values of δ are 0, 3 and 6 from top to bottom; and the three plots on the right side are plots of power versus δ where values of d are 0, 0.2 and 0.5.

As seen in Figure 1 and other plots in the Appendix, the power with or without considering clustering prior to randomization increases as the treatment effect d increases, and also power with clustering prior to randomization was higher than the one without clustering no matter the change of the mean differences δ between two mixture normal distributions and also the change of the treatment effect d. Except that when the $d=0$, that is under the null hypothesis, there is no treatment effect, here the power actually is type I error, and it maintains at 0.05 no matter with or without clustering. That means we can control type I error by this method. When the two distributions have same variances, the difference between the power with cluster stratification versus without is smaller—for fixed treatment effect d—when the mean difference is moderate, $\delta = 3$, but larger for zero or larger mean differences beyond moderate, δ is 0 or 6.

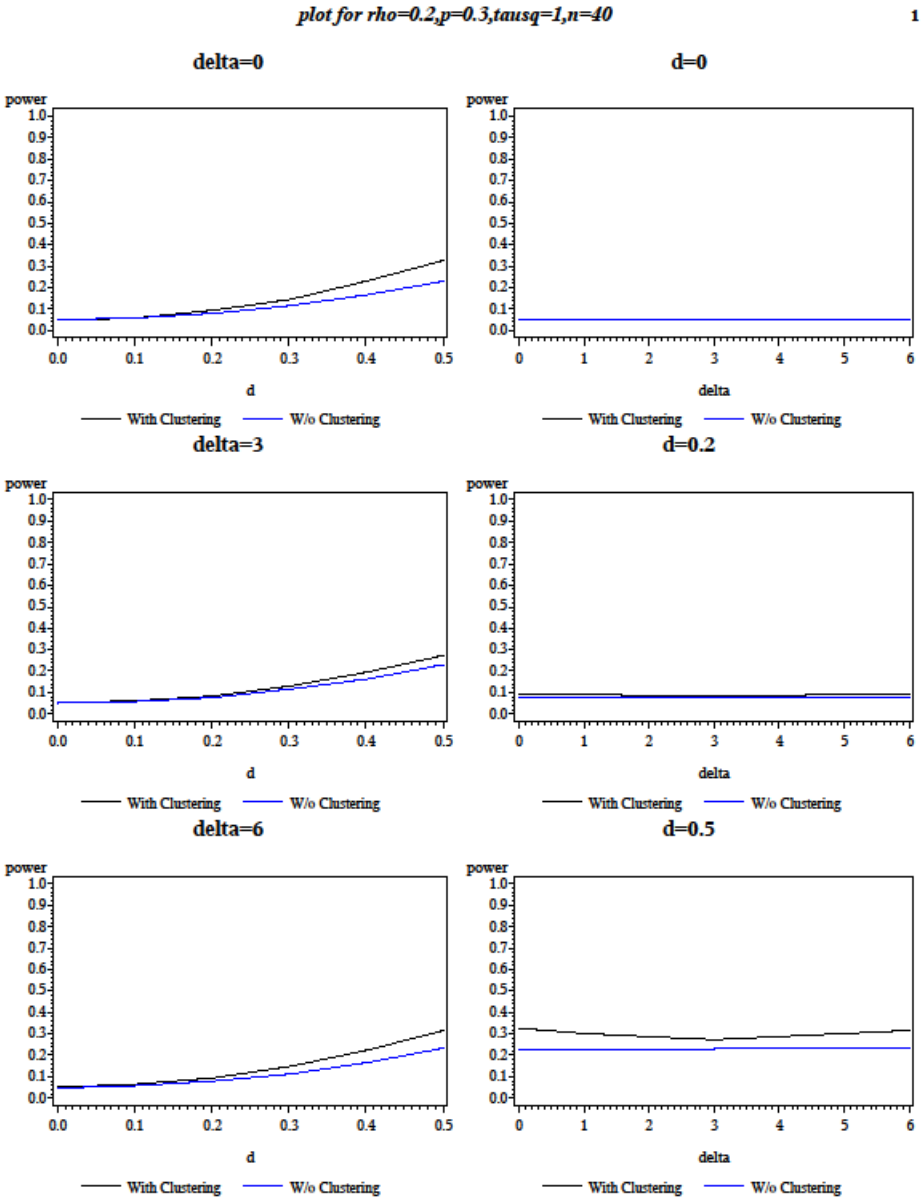


Figure 1: Power versus d or δ (ρ =0.2, p =0.3, τ^2 =1 and n =40)

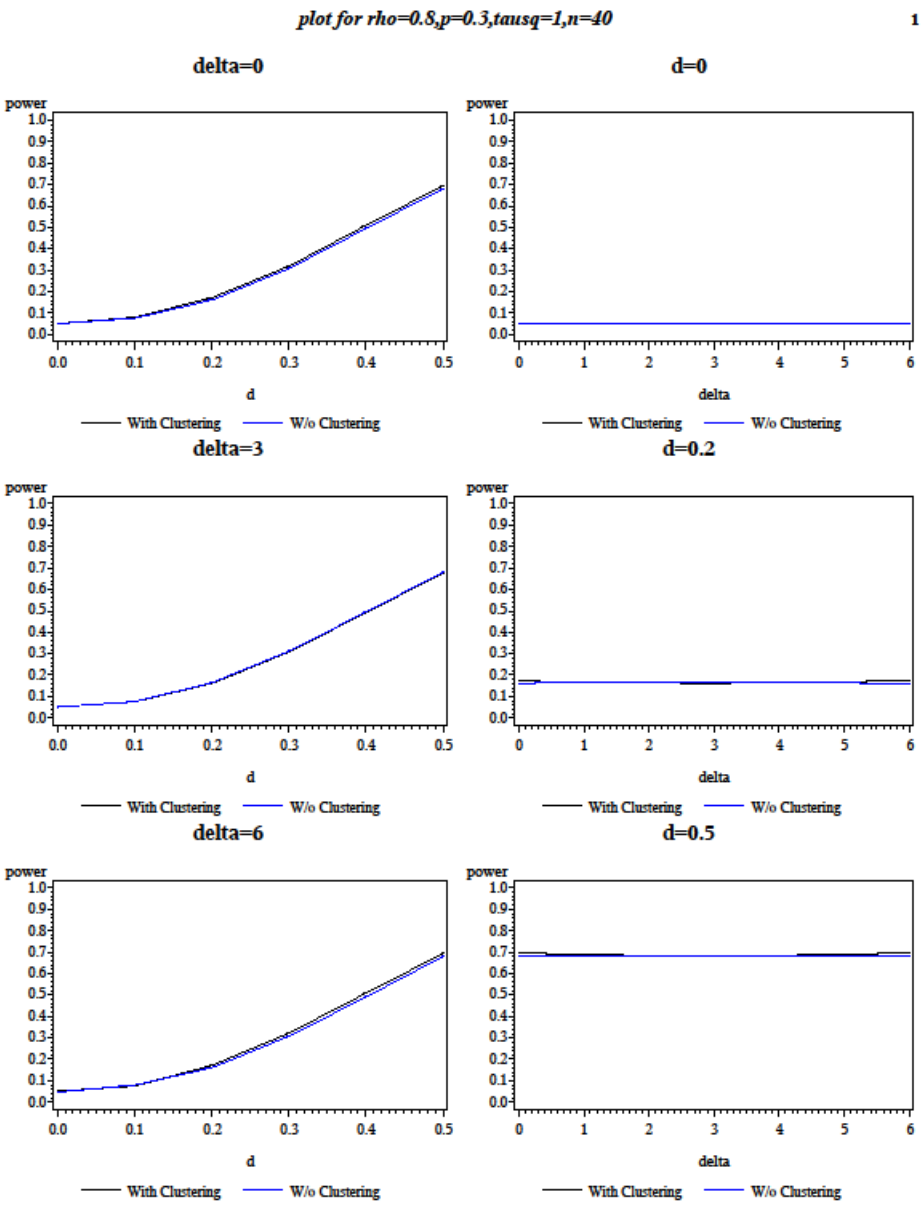


Figure 2: Power versus d or δ (ρ =0.8, p =0.3, τ^2 =1 and n =40)

In Figure 2 with $\rho = 0.8$, differences between the power with or without stratification were too small to be noticed. This is because as the correlation ρ of the mixture distribution increases, the overall variance decreases, so that there is less room to decrease

variance to increase power. As we know, clustering will increase the degrees of freedom for analysis, but it still does not hurt the power. Also we can conclude that as ρ increases, the difference between the power with versus without clustering decreases. As we change the proportion p of the first distribution in the mixture normal distribution, and keep the rest of the parameters for the sample, the power will not change much according to the plots in the Appendix A.

Figure 1, 3 and 4 have the same parameters and only differ by the sample size n (n=40 for Figure 1, n=20 for Figure 3 and n=60 for Figure 4). The power increases as the sample size increases. However when the sample sizes increase from 20 to 40, the differences of the power between with or without clustering increase, but when the sample sizes increase from 40 to 60, it is very difficult to notice the increase.

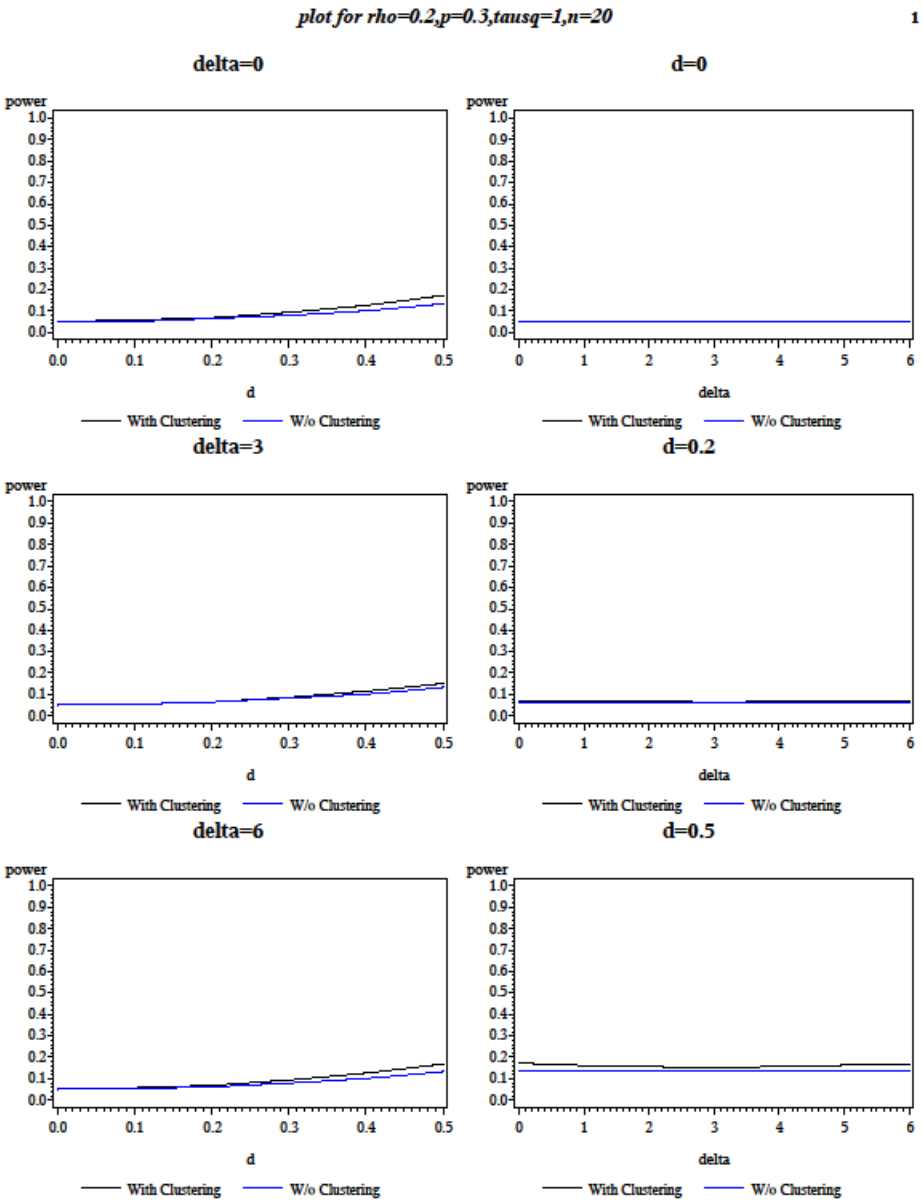


Figure 3: Power versus d or δ (ρ =0.2, p =0.3, τ^2 =1 and n =20)

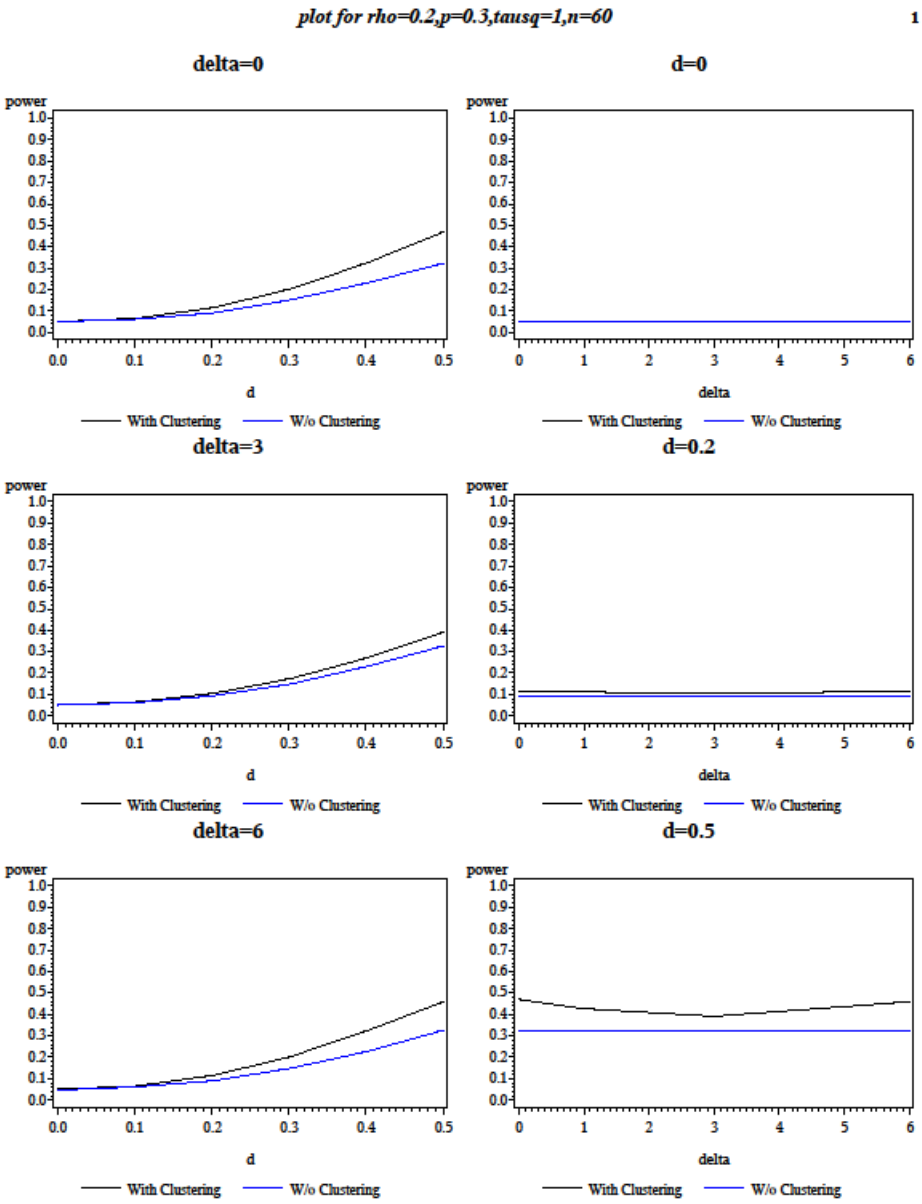


Figure 4: Power versus d or δ (ρ =0.2, p =0.3, τ^2 =1 and n =60)

CHAPTER 5 Discussion

According to all the results, no matter with or without baseline clustering, the estimated type I errors are close to 0.05. This means our method achieves the nominal type I error probability, even with non-normal data.

The power of detecting the treatment effect increases with clustering on baseline prior to randomization. As expected, the power increases as the treatment effect and sample size increase. When the sample size is large, there is little room for improvement in power. For all the results we simulated, the largest increase of power is 15%, corresponding to a sample size of 60, treatment effect of 0.5, proportion of the first distribution equal to 0.3, and the two distributions having equal means and variances with pre-post correlation being 0.2.

However, the improvement in power with baseline clustering is small, even negligible, when the correlation between pre-post measurements is high. For example, when the correlation is 0.8 and the two distributions have the equal variance, it is very difficult to notice the increase of the power related to baseline clustering. The power could be decreased by stratification since clustering takes up some degrees of freedom in modeling the treatment effect. In this simulation study, only increases in power, though some negligible, were observed.

When the two distributions have equal variances, the improvement in power is also smaller when the mean difference of the two distributions is $\delta = 3$ as compared to $\delta = 0$ or 6. The two distributions have the same mode when $\delta = 0$ but the modes will separate as

the mean difference increases. For moderate separation the overall distribution is less concentrated around its center; that is, the data do not tend to cluster and clustering on baseline corresponds to less improvement in power. As the mean difference reaches 6 or more, the two distributions are almost completely separated and the data tend to cluster around one or the other mode. In this case the baseline measures are naturally clustered and we see that baseline clustering improves the power.

This simulation study shows that baseline clustering improves the power, though less so if the pre-post measures are highly correlated or if the data do not naturally cluster. However, clustering on baseline data before randomization is recommended since it will improve the power of a pre-post test.

CHAPTER 6 Summary and Future Studies

The power of detecting the treatment difference is increased by clustering on baseline measures prior to randomization.

In this thesis, the clustering method used can only applied to one dimensional data. An algorithm that reaches optimality, is repeatable, and meets constraints on cluster size remains to be developed for multidimensional data.

We treated cluster membership as a fixed effect. Since randomization occurs within cluster, treating the cluster as a random effect should not alter our results. However, from sample to sample the clustering will be different. An additional element of variance is induced by this method of clustering. Park and Johnson (2006) studied the variance for pair-matching. Future work could do the same for cluster-matching.

In our method, we fixed the variance of the second distribution in the mixture normal distribution for the baseline measurements to be 1, set the variance of the first distribution as the ratio of the two distributions τ^2, and we used different values of τ^2 for simulation. Actually the total variance for the paired differences are an increasing function of τ^2. The observed change in power is affected by the total variance as well as—possibly—the change of the ratio of the two variances. These two effects on variance are confounded in this study. In order to avoid this limitation, we can fix the total variance, but still study the ratio of the two variances.

Literature Cited

Aloise D, Deshpande A, Hansen P, and Popat P. (2009) NP-hardness of Euclidean Sums-of-Squares Clustering. *Machine Learning*, 75(2): 245-248.

Anderberg, M. R. (1973), *Cluster Analysis for Applications*, New York: Academic Press.

Dasgupta S and Freund Y (2009), Random Projection Trees for Vector Quantization. *IEEE Transactions on Information Theory*, 55(7):3229-3242.

Jain, A.K. (2010), Data Clustering: 50 Years Beyond K-means. *Pattern Recognition Letters*, 31(8): 651-666.

Lloyd, S.P. (1982), Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129-137.

MacQueen, J.B. (1967), Some Methods for Classification and Analysis of Multivariate Observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 281-297

Milligan, G. W. (1980), An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms, *Psychometrika*, 45, 325–342.

Park, M and Johnson, RE (2004), Methods for Matching Clusters on Baseline Outcome Measures Prior to Randomization. *2004 Proceedings of the American Statistical Association*, 423-426, Alexandria, VA: American Statistical Association.

Park, M and Johnson, RE (2005), Stratification on Baseline Measure: The Variance Effect. *2005 Proceedings of the American Statistical Association*, Vol. 4, Alexandria, VA: American Statistical Association.

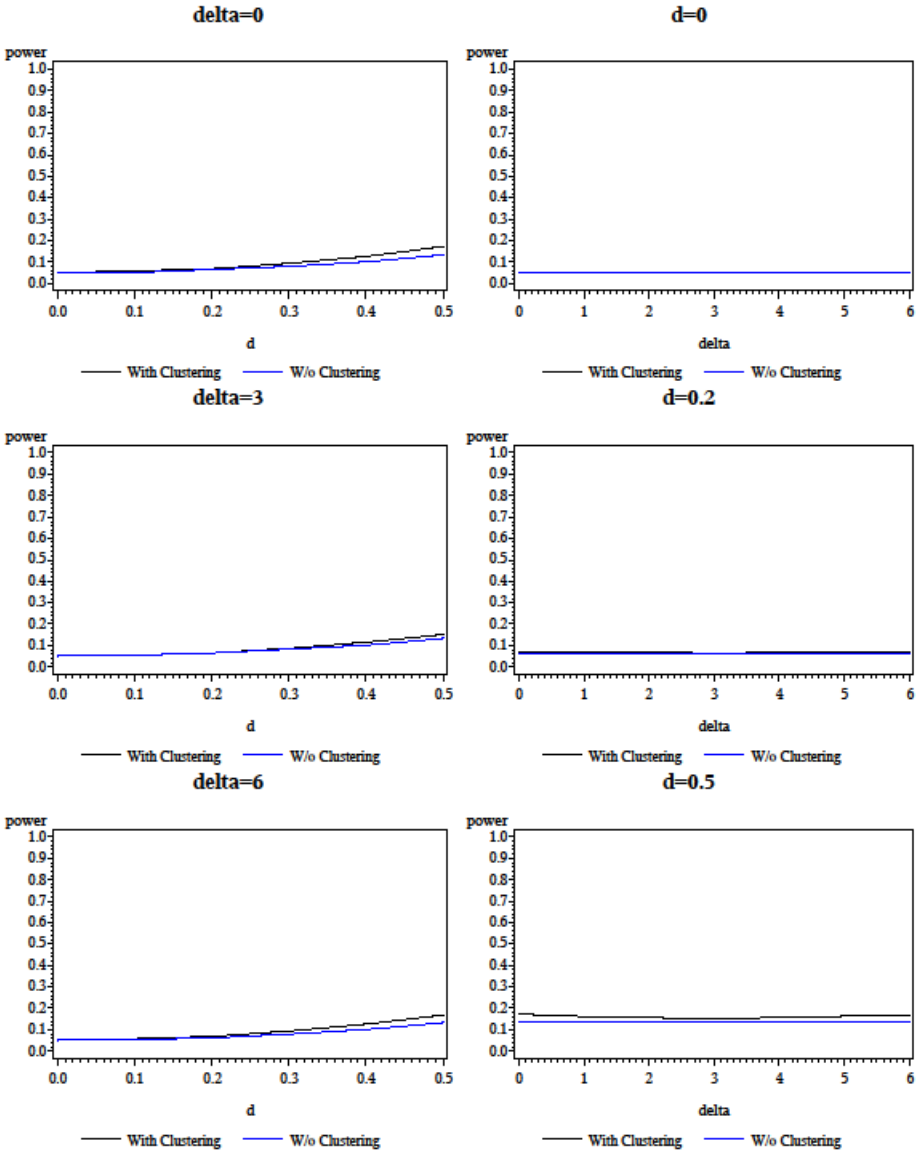
Park, M and Johnson, RE (2006), Design and Analysis Methods for Cluster Randomized Trials with Pair-Matching On Baseline Outcome: Reduction of Treatment Effect Variance. *2006 Proceedings of the American Statistical Association*, Alexandria, VA: American Statistical Association.

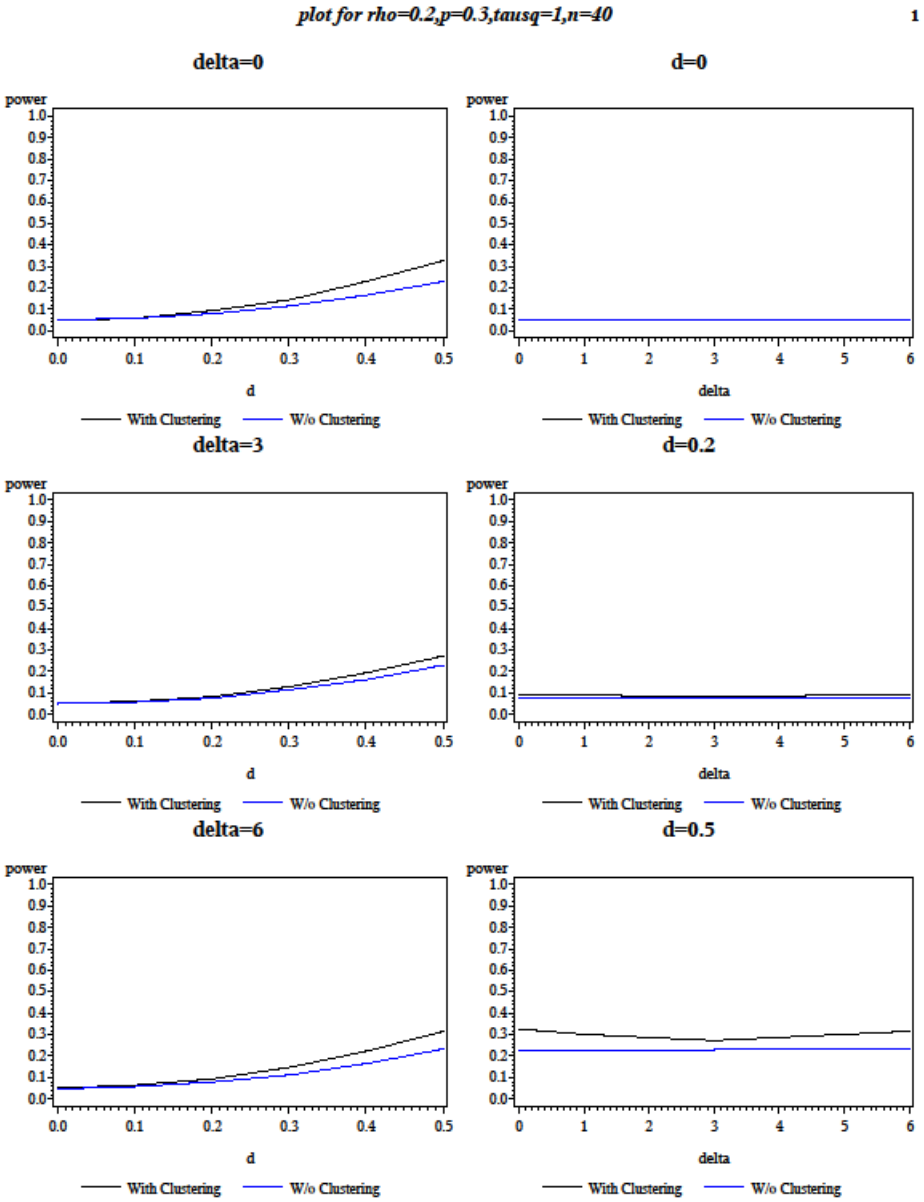
Wang, H and Song, M (2011), CKmeans.1d.dp: Optimal *k-means* Clustering in One Dimension by Dynamic Programming. *The R Journal*, Vol. 3/2, December 2011.

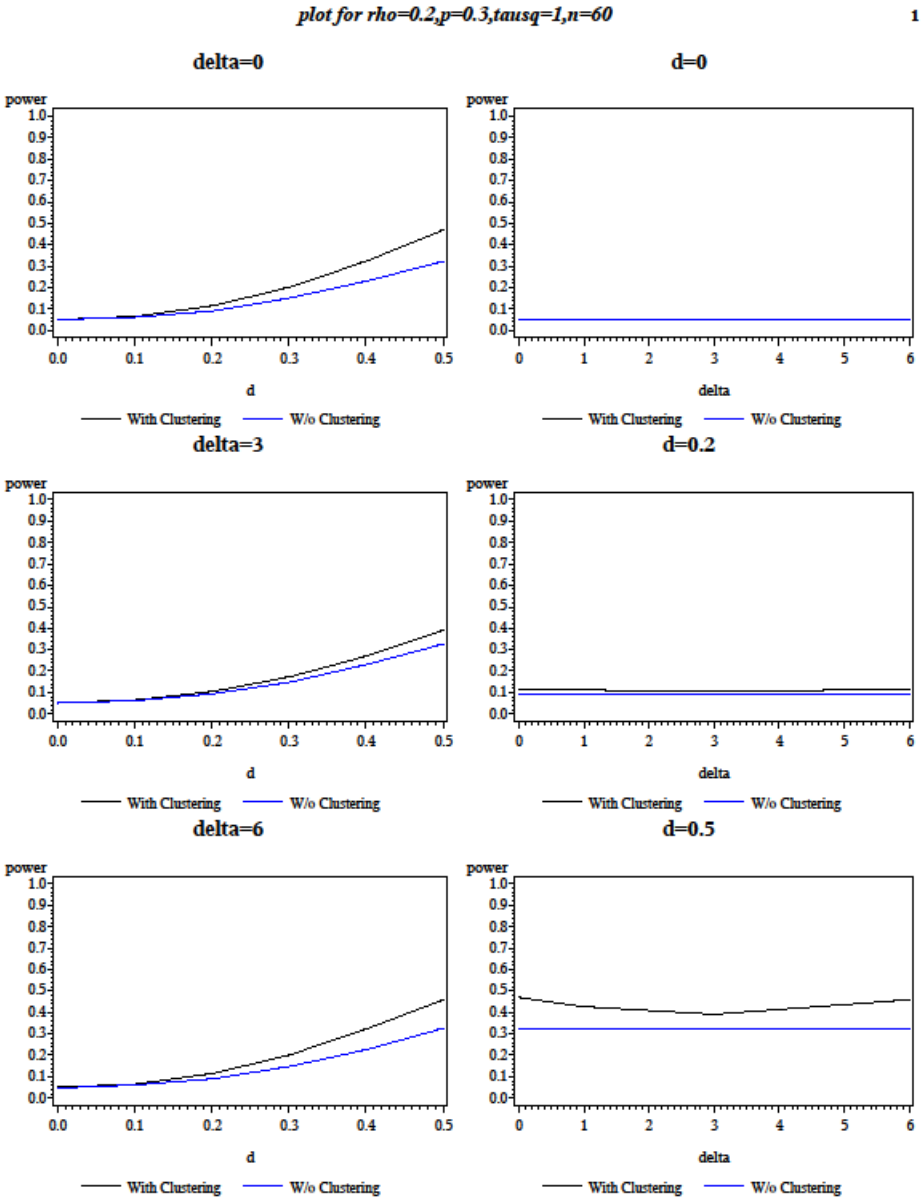
APPENDIX A

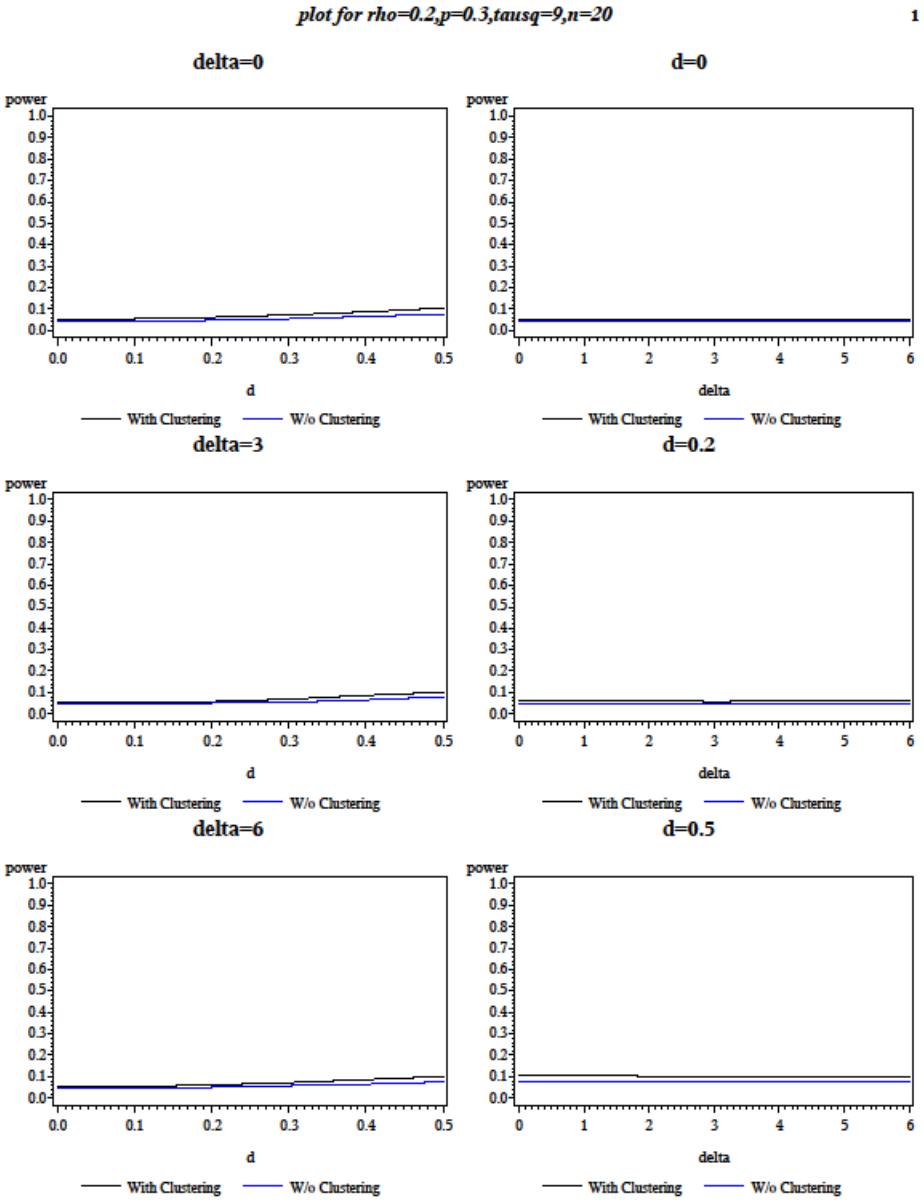
Power versus d or δ

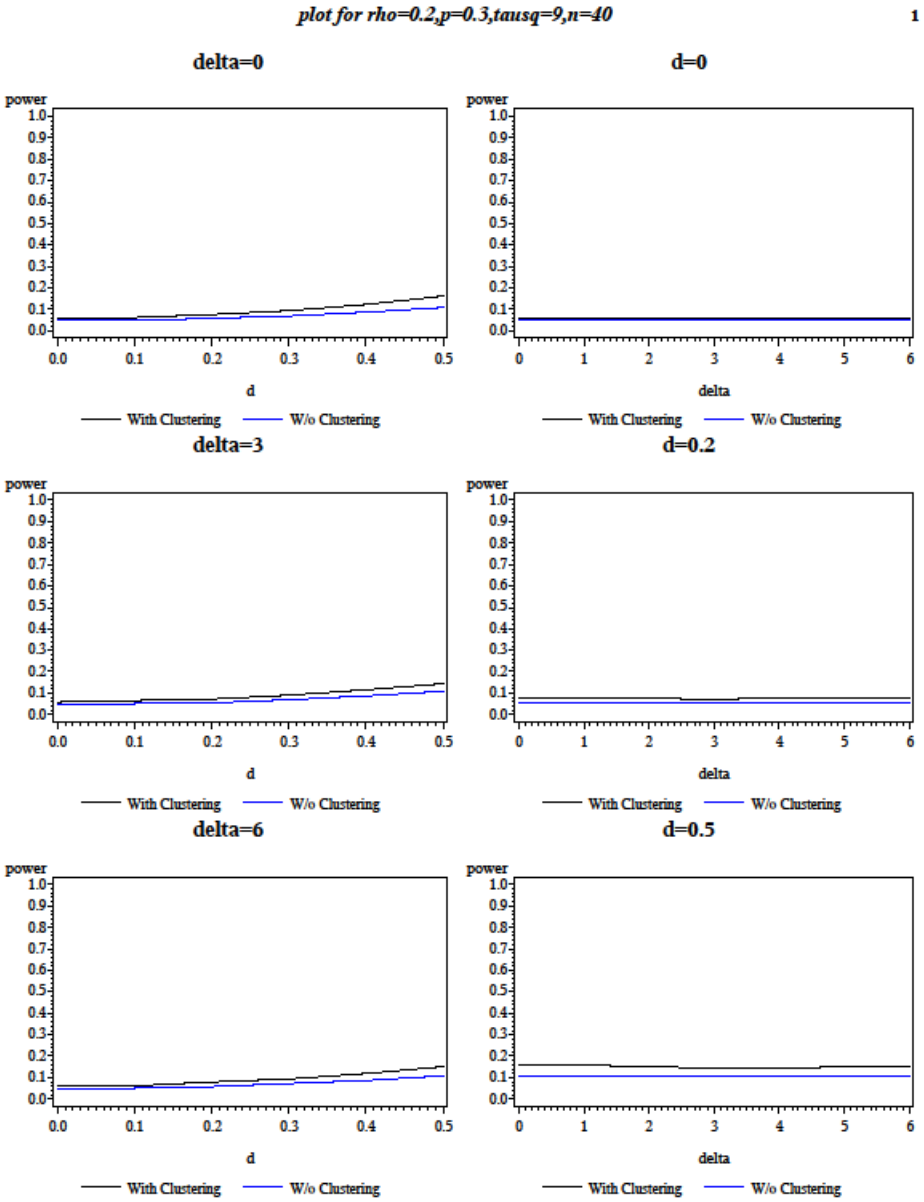
plot for rho=0.2,p=0.3,tausq=1,n=20 1

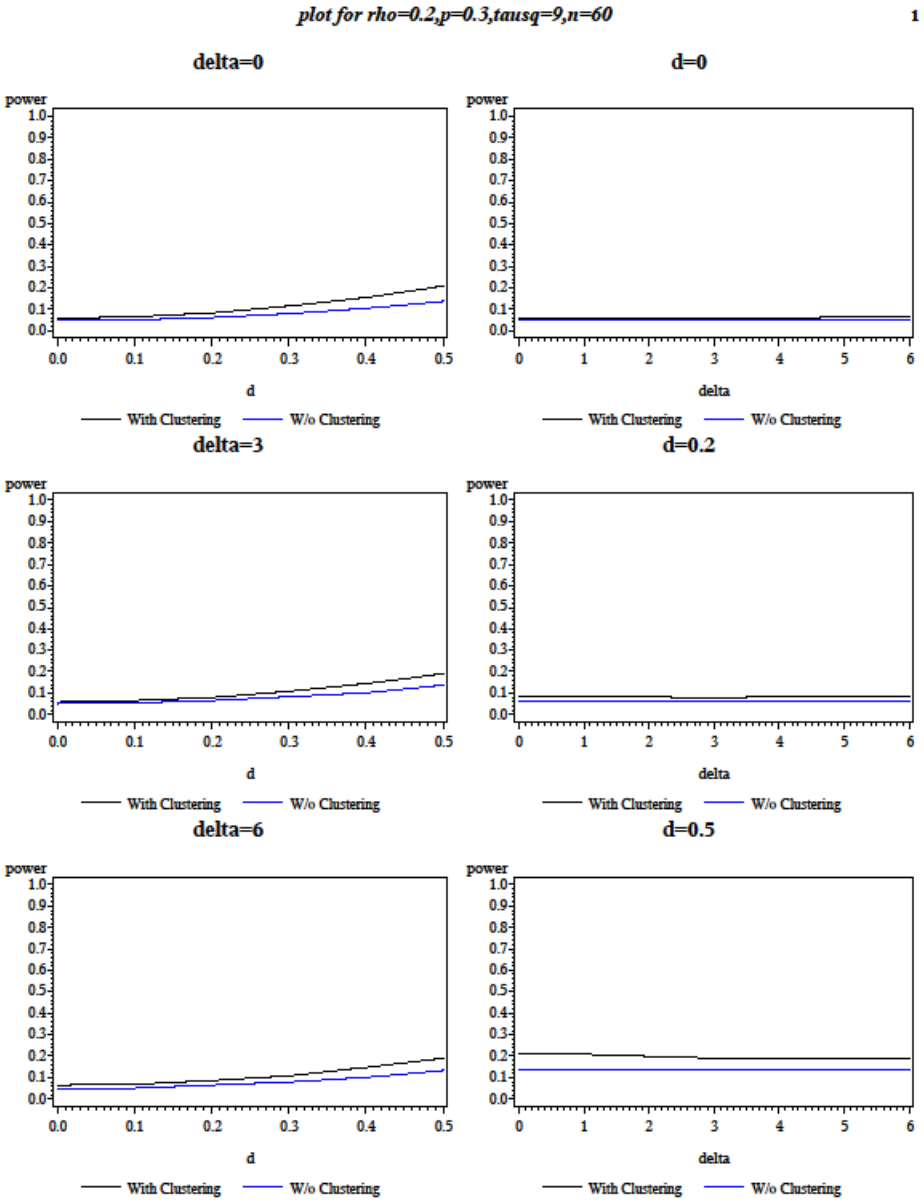


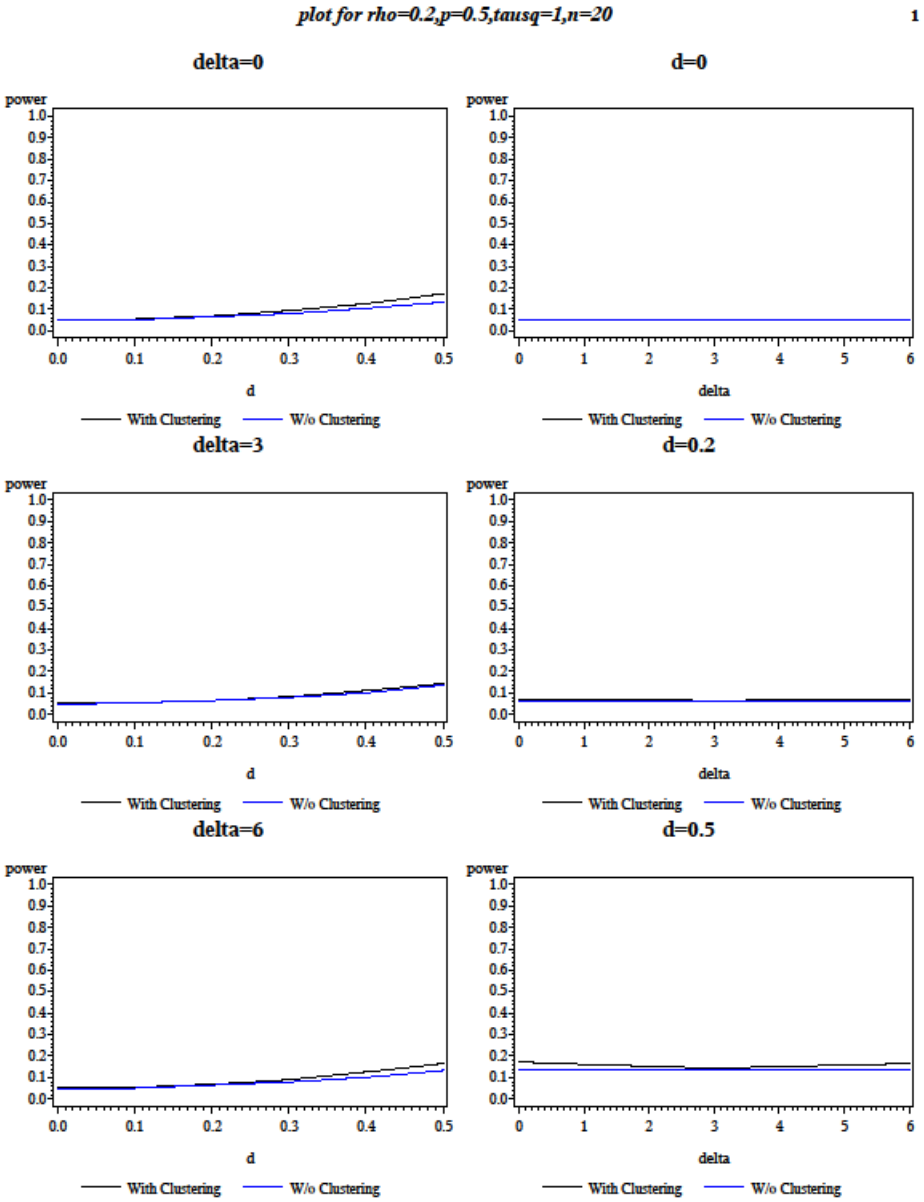


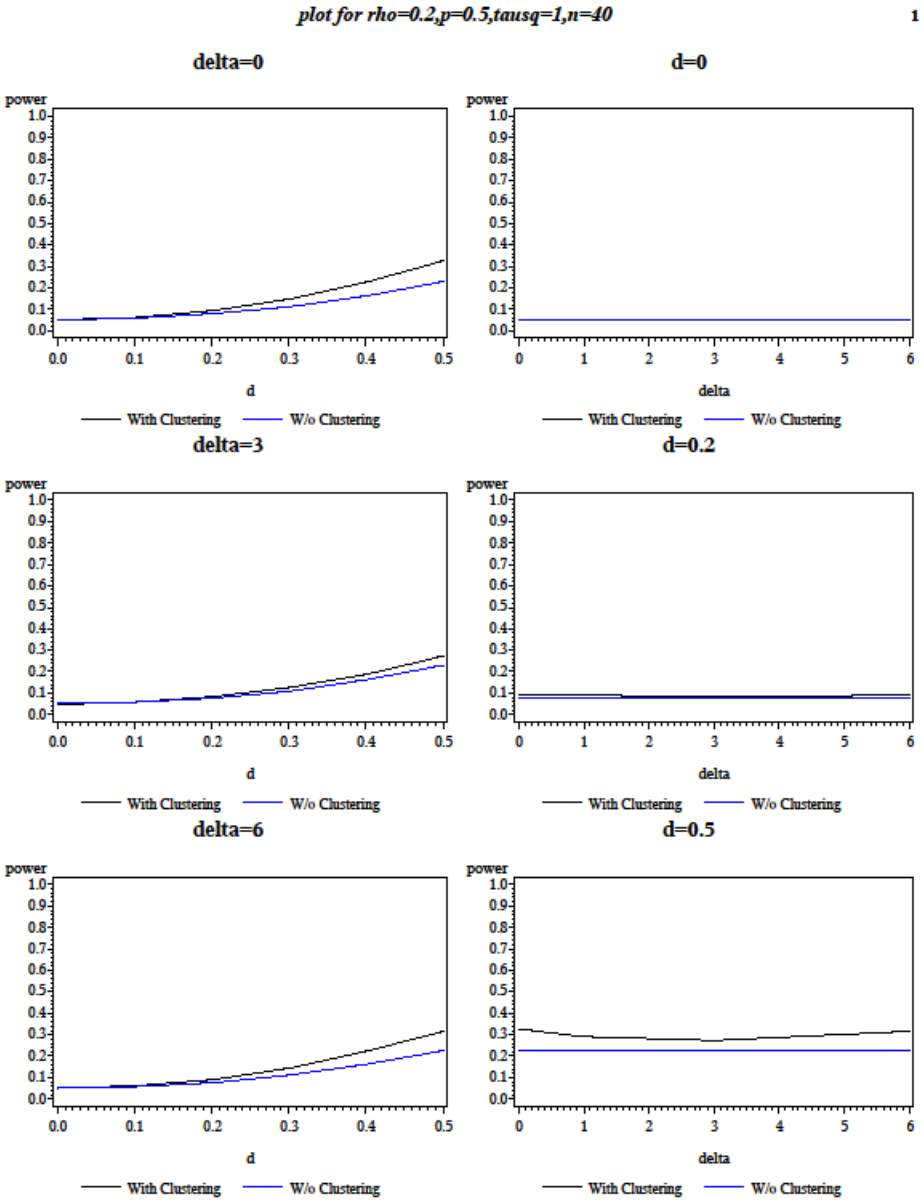


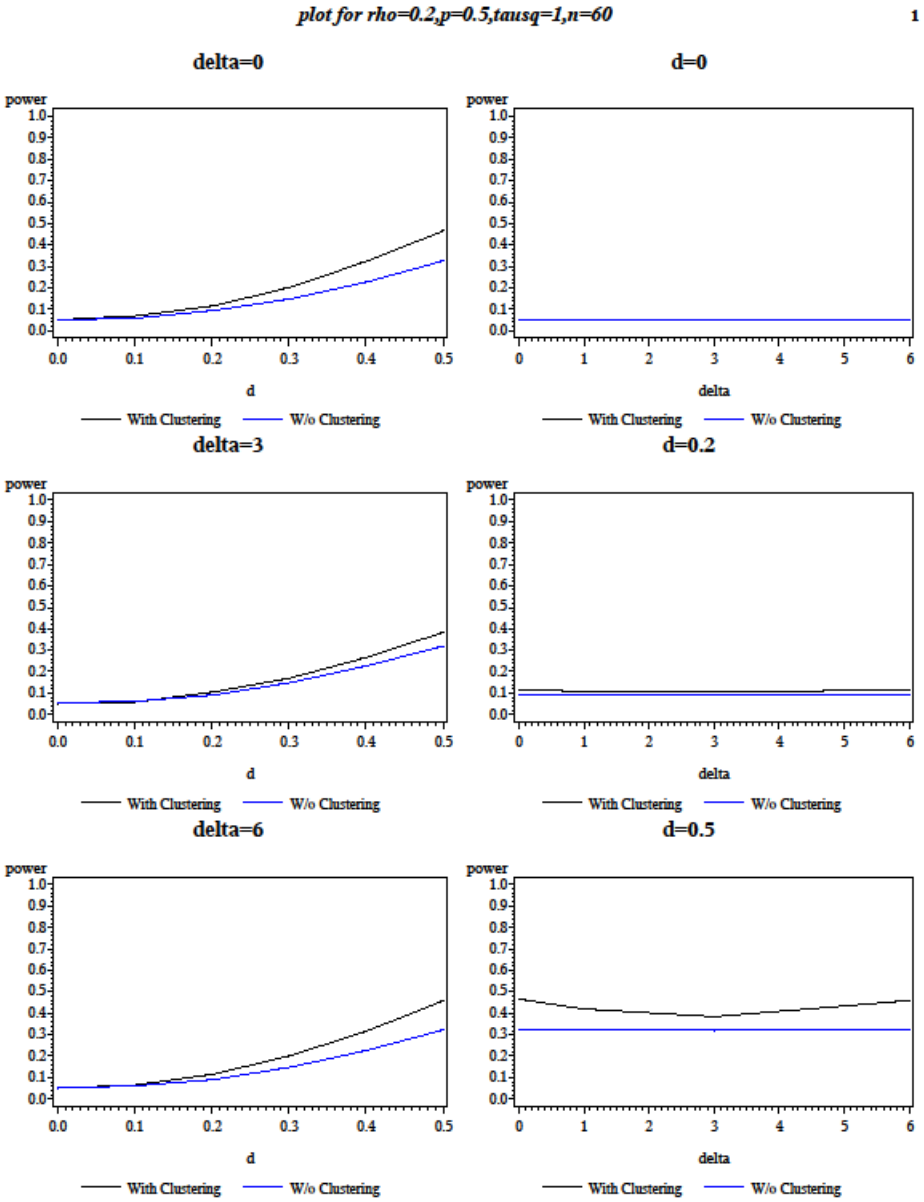


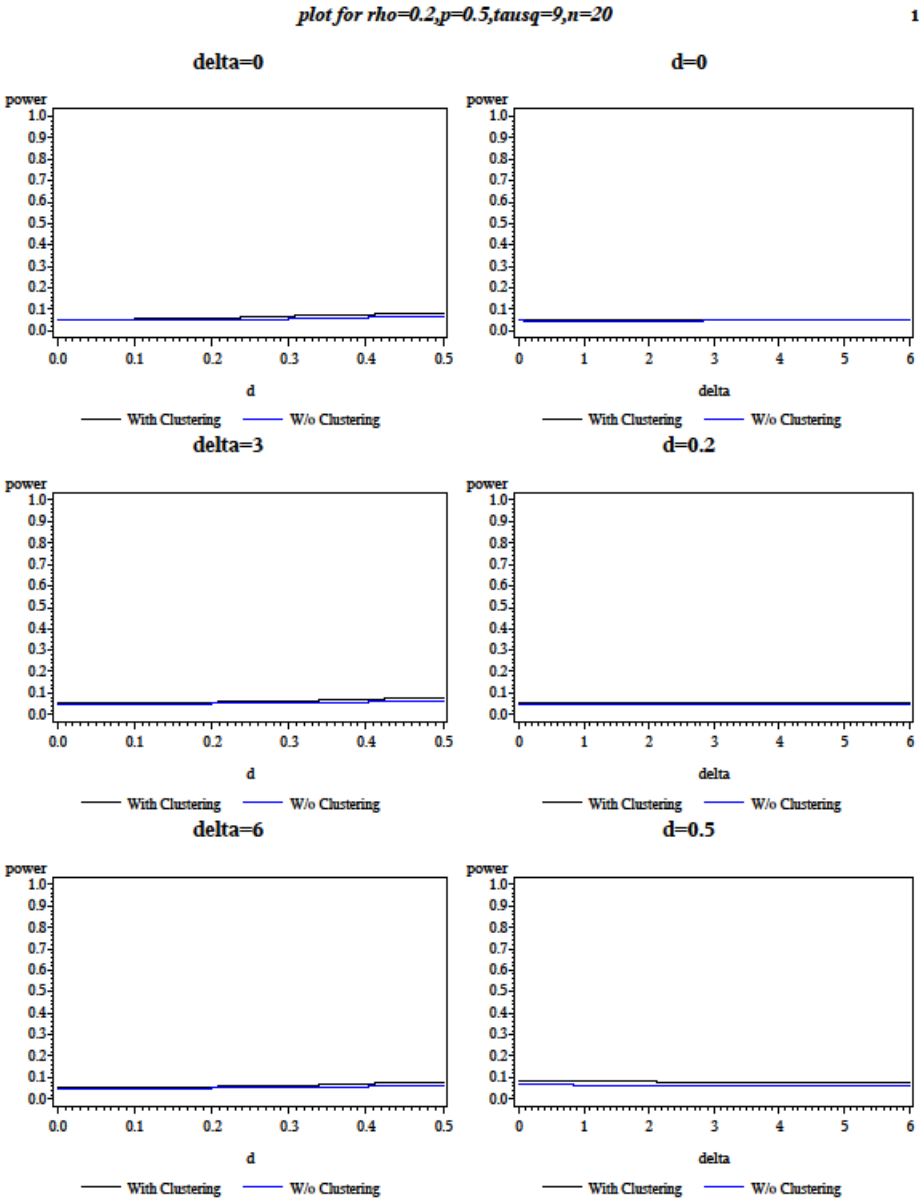


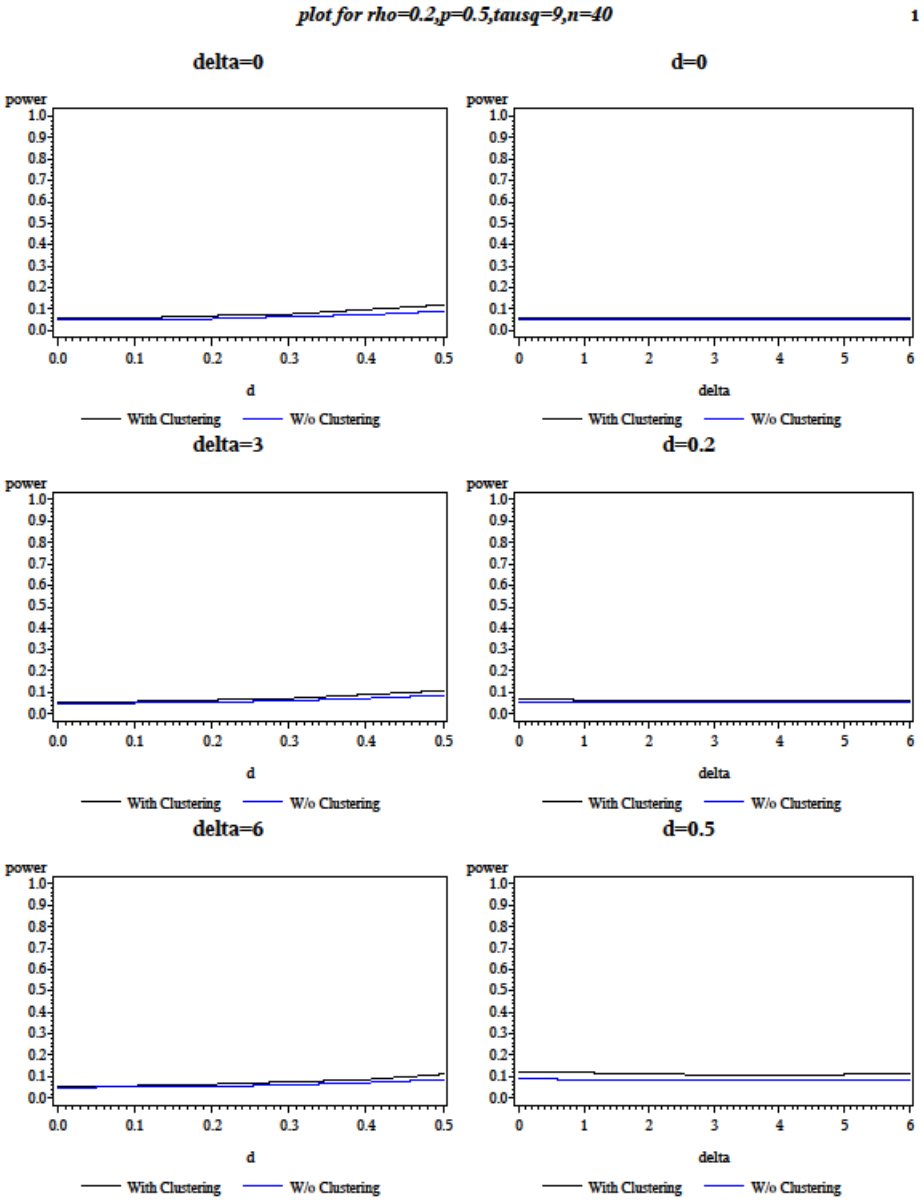


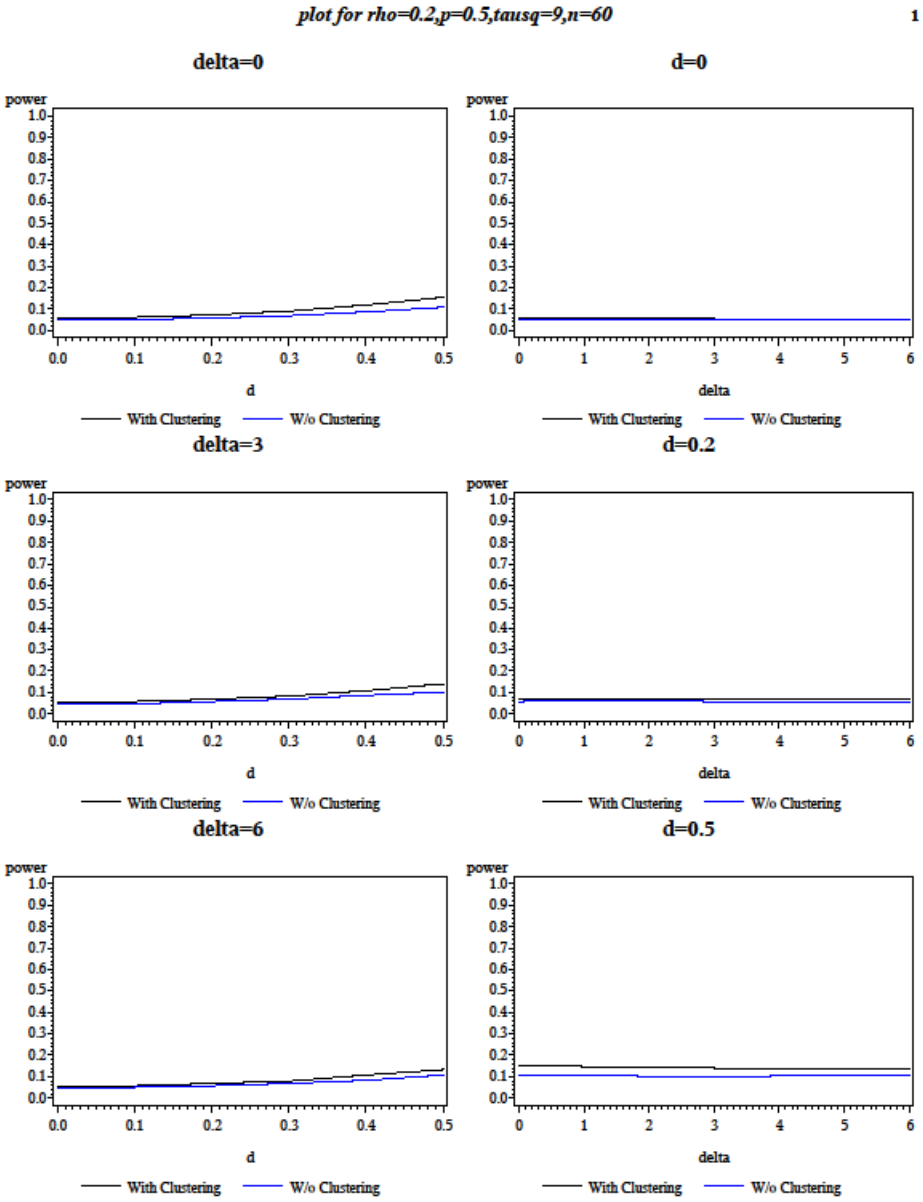


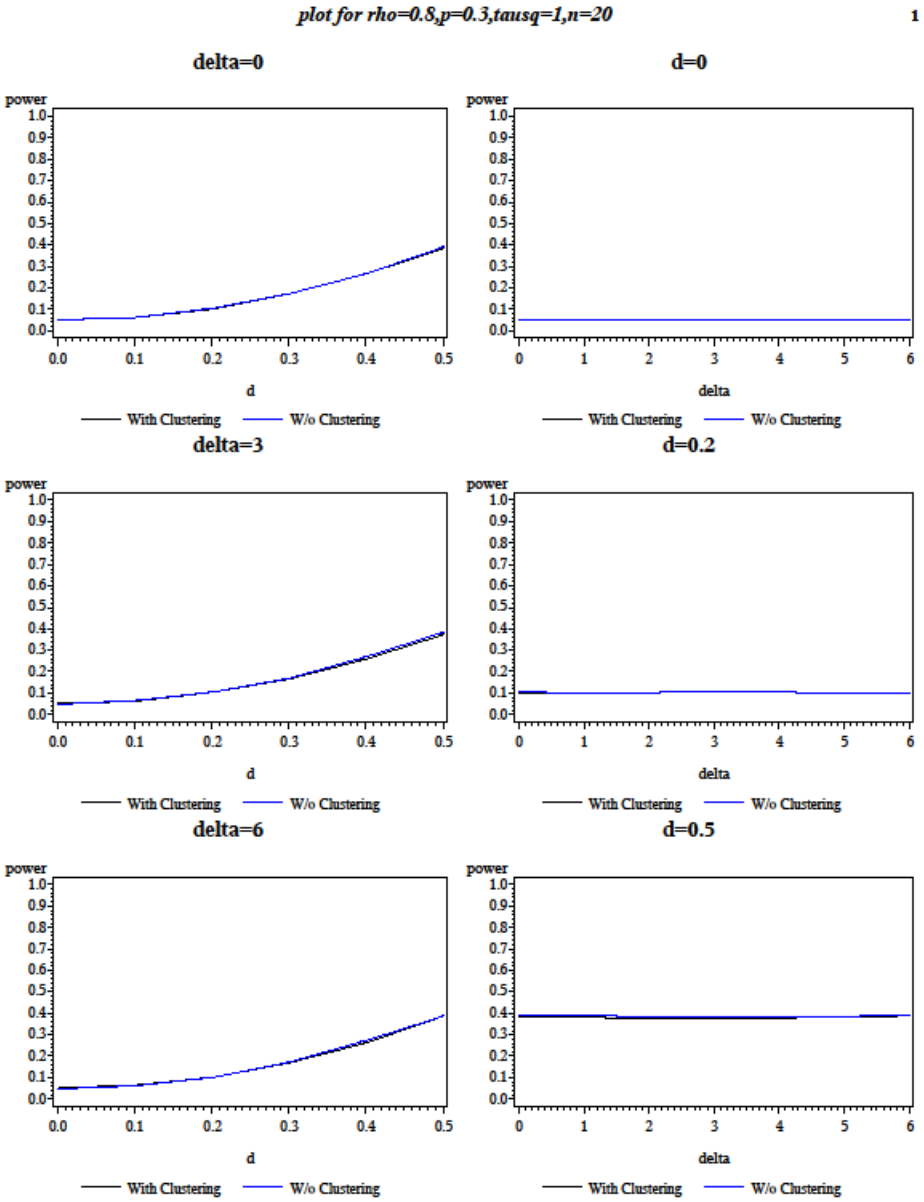


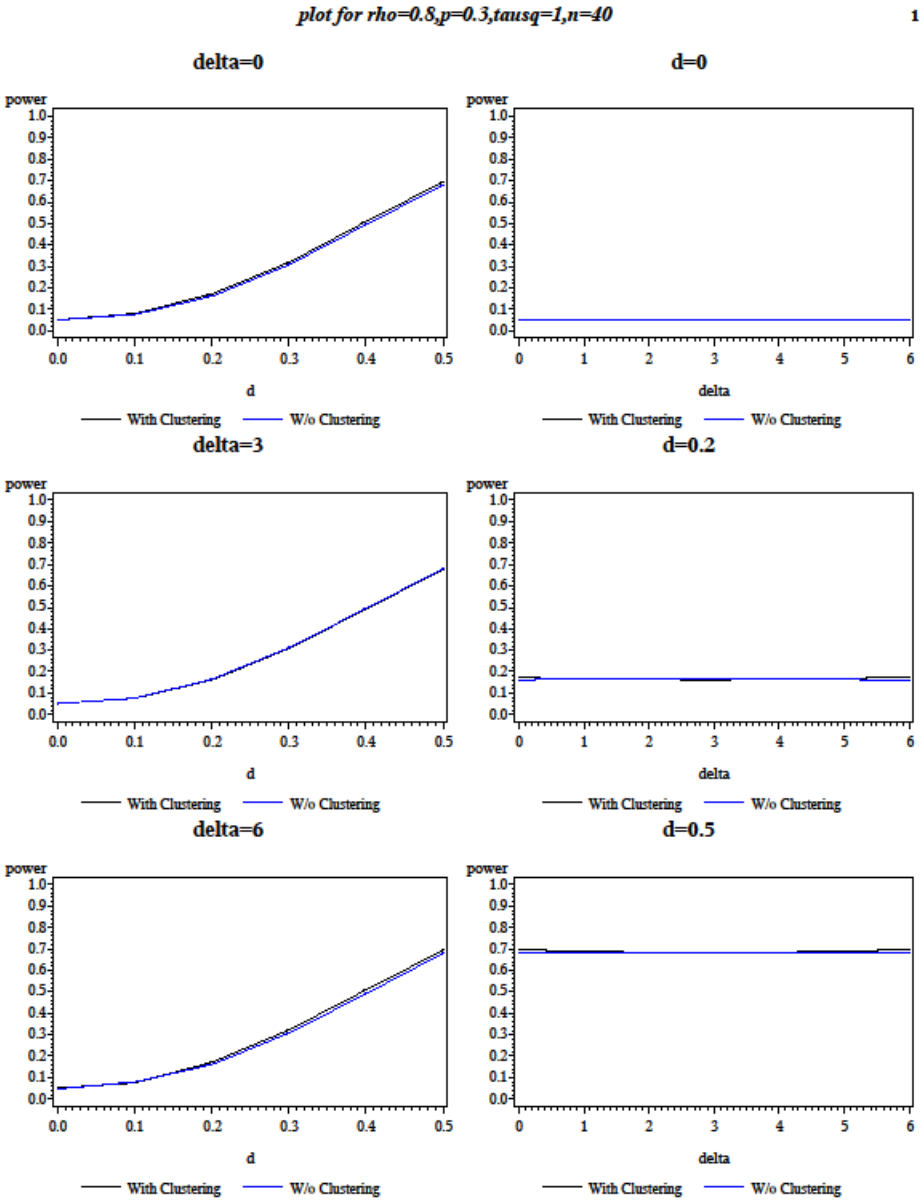


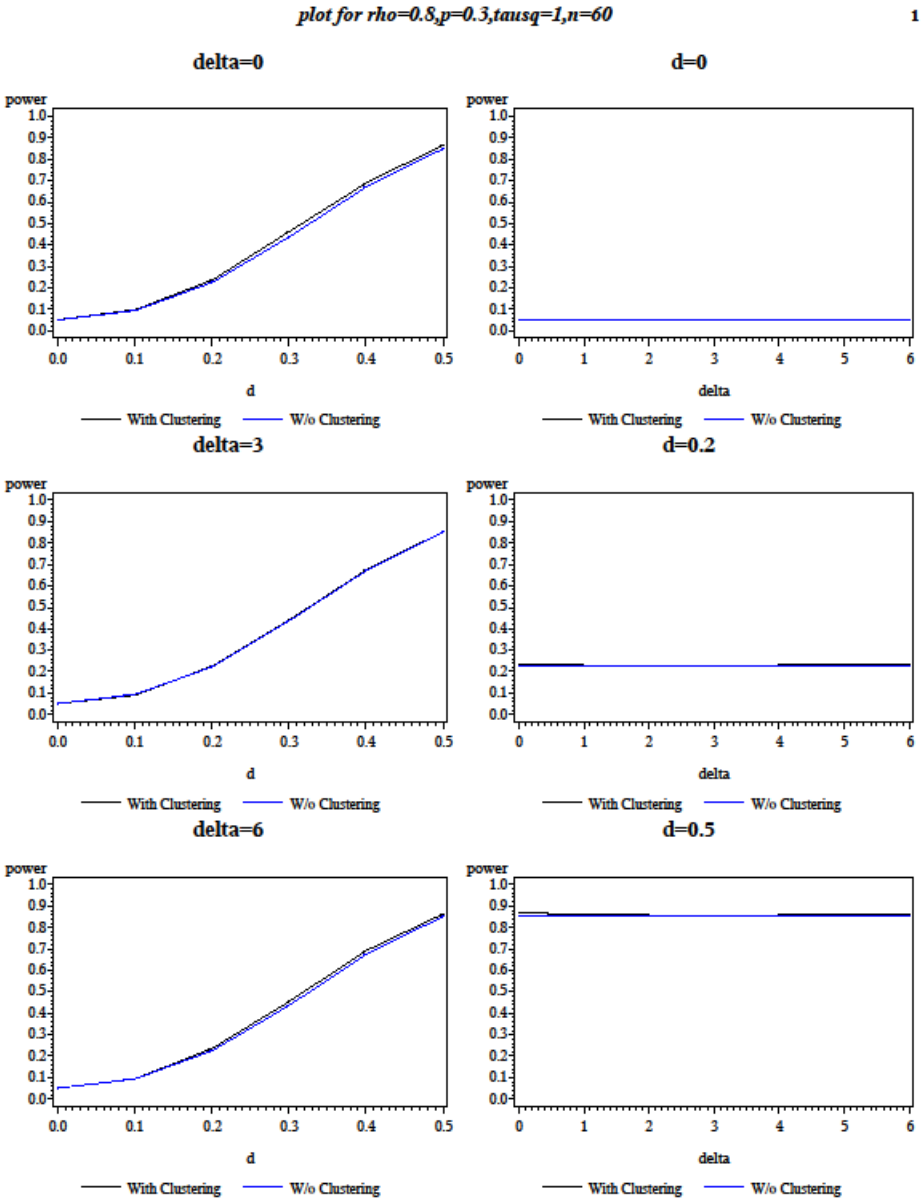


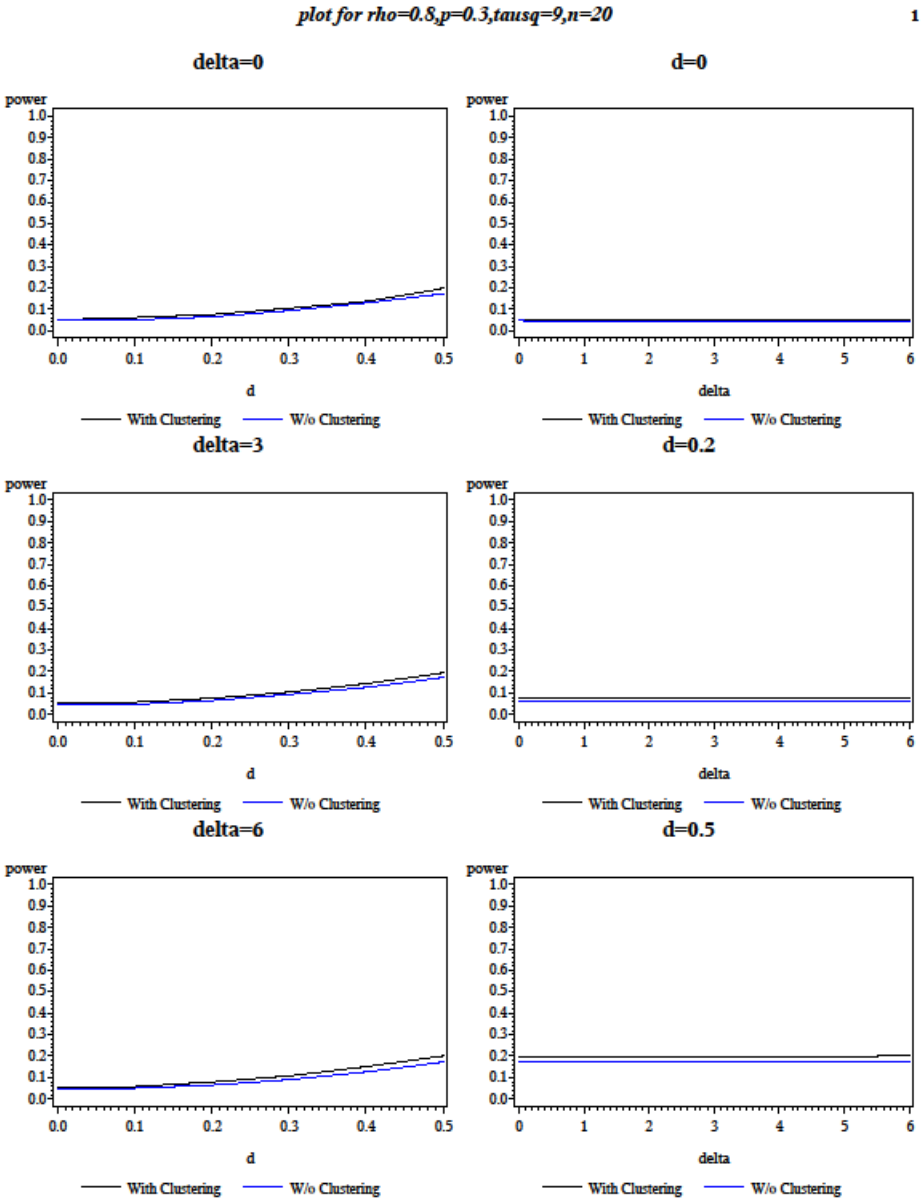


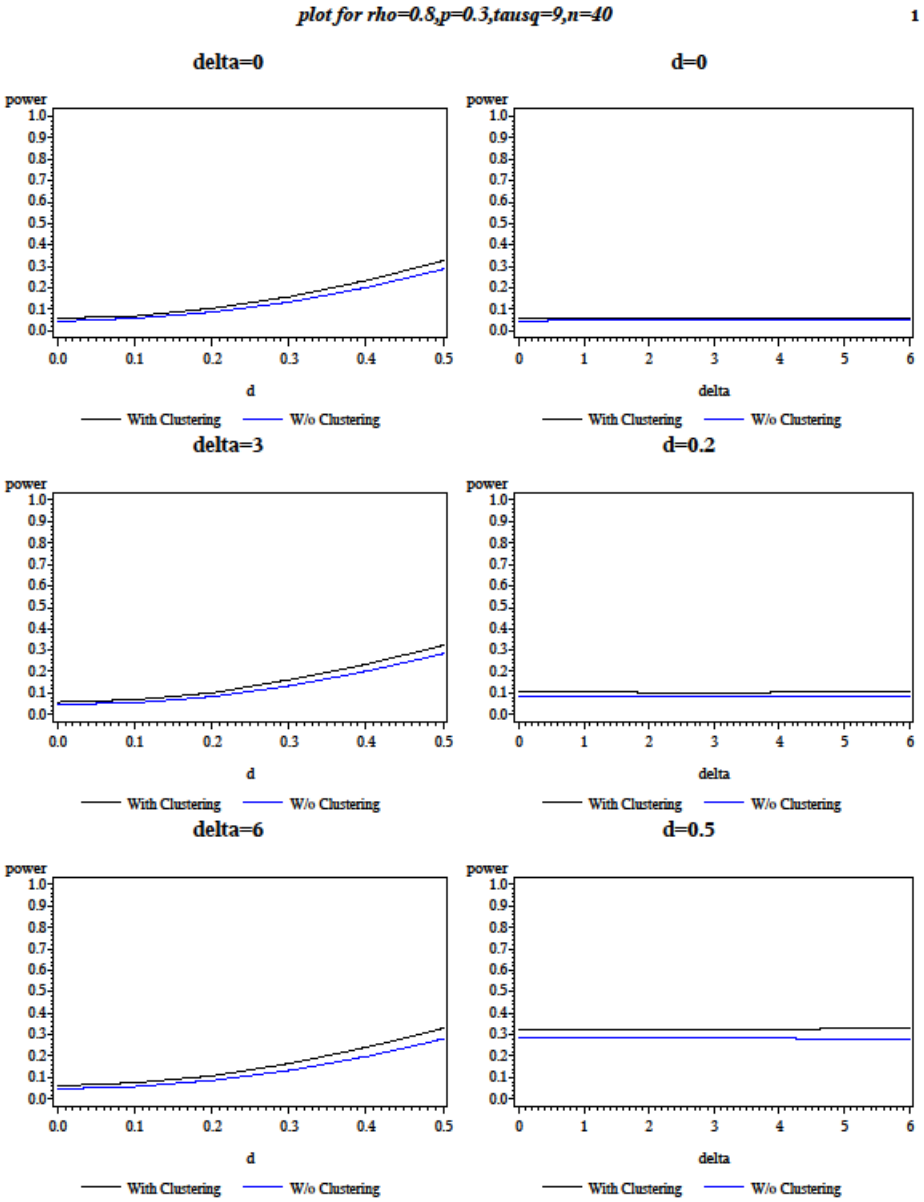


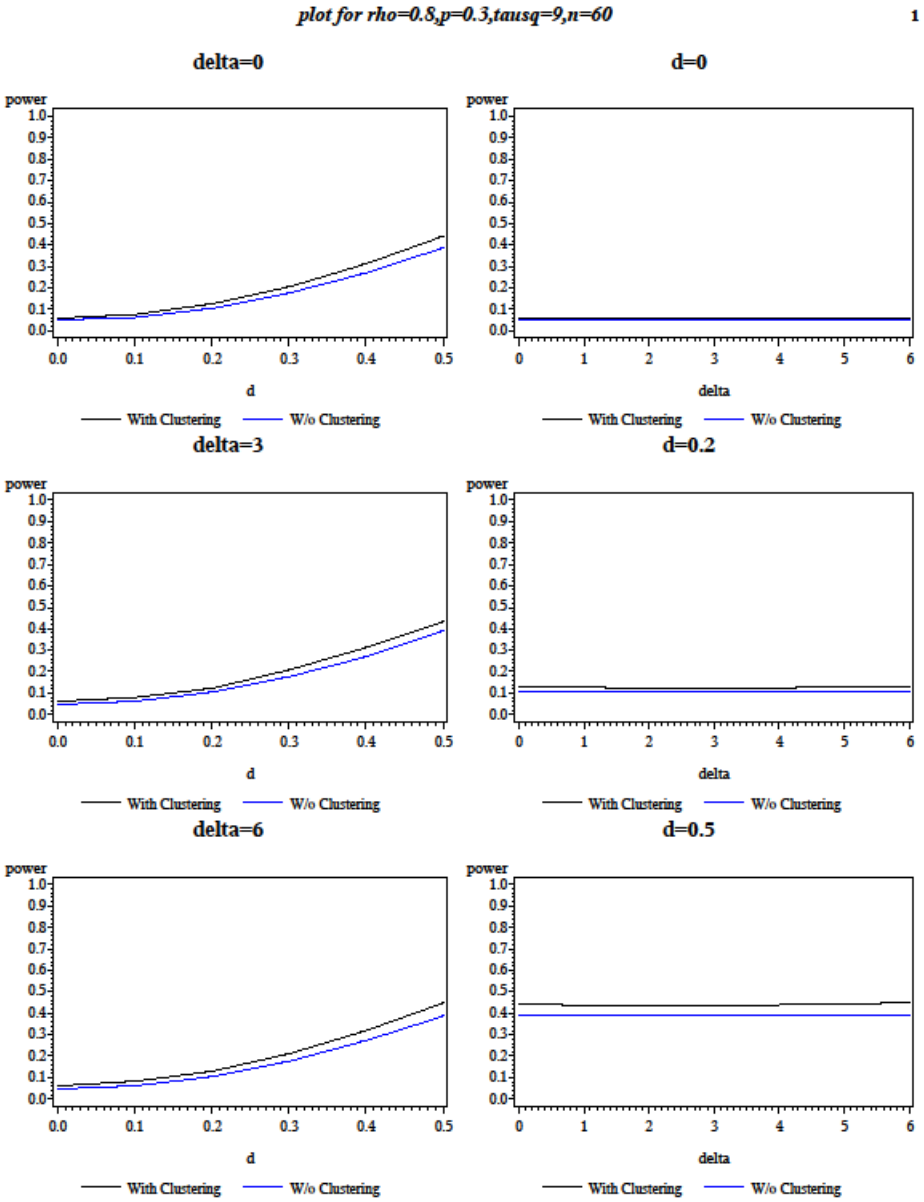


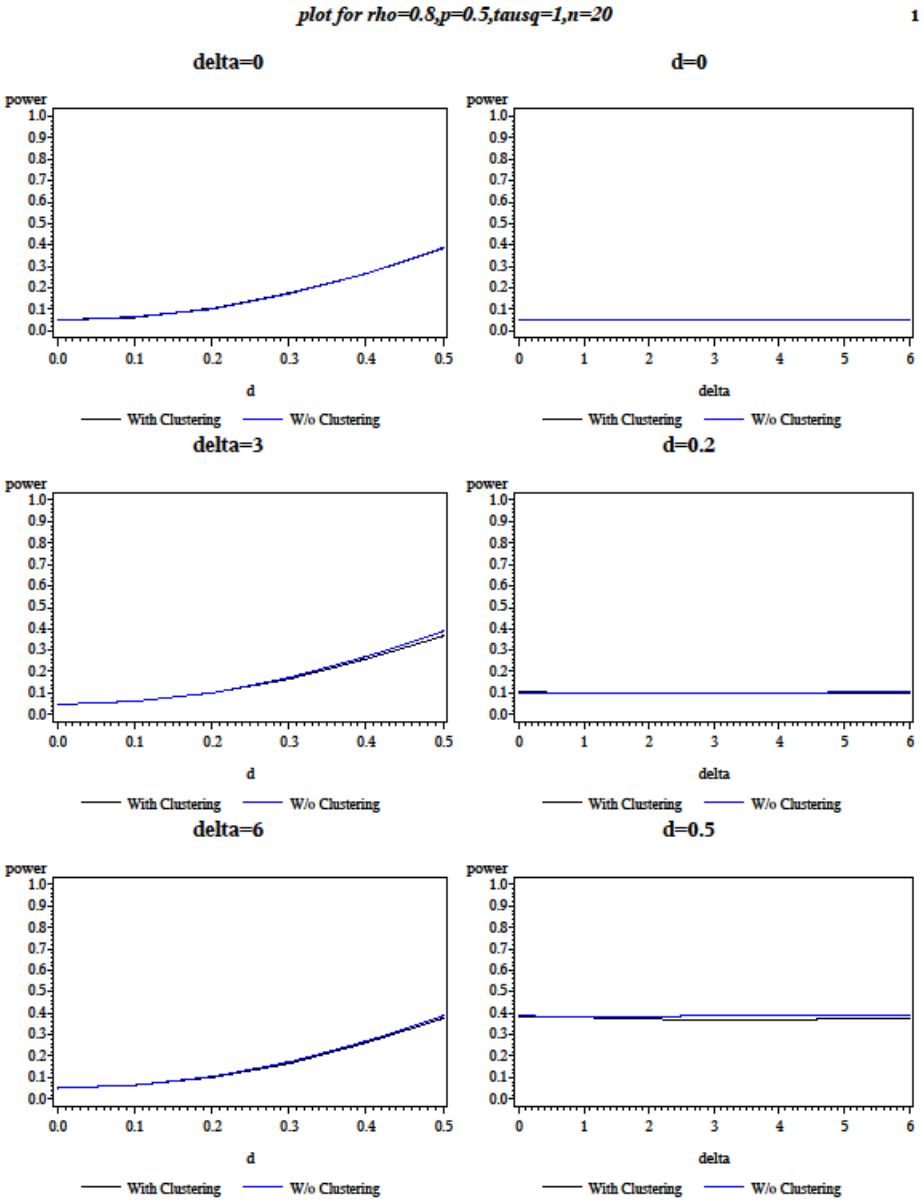


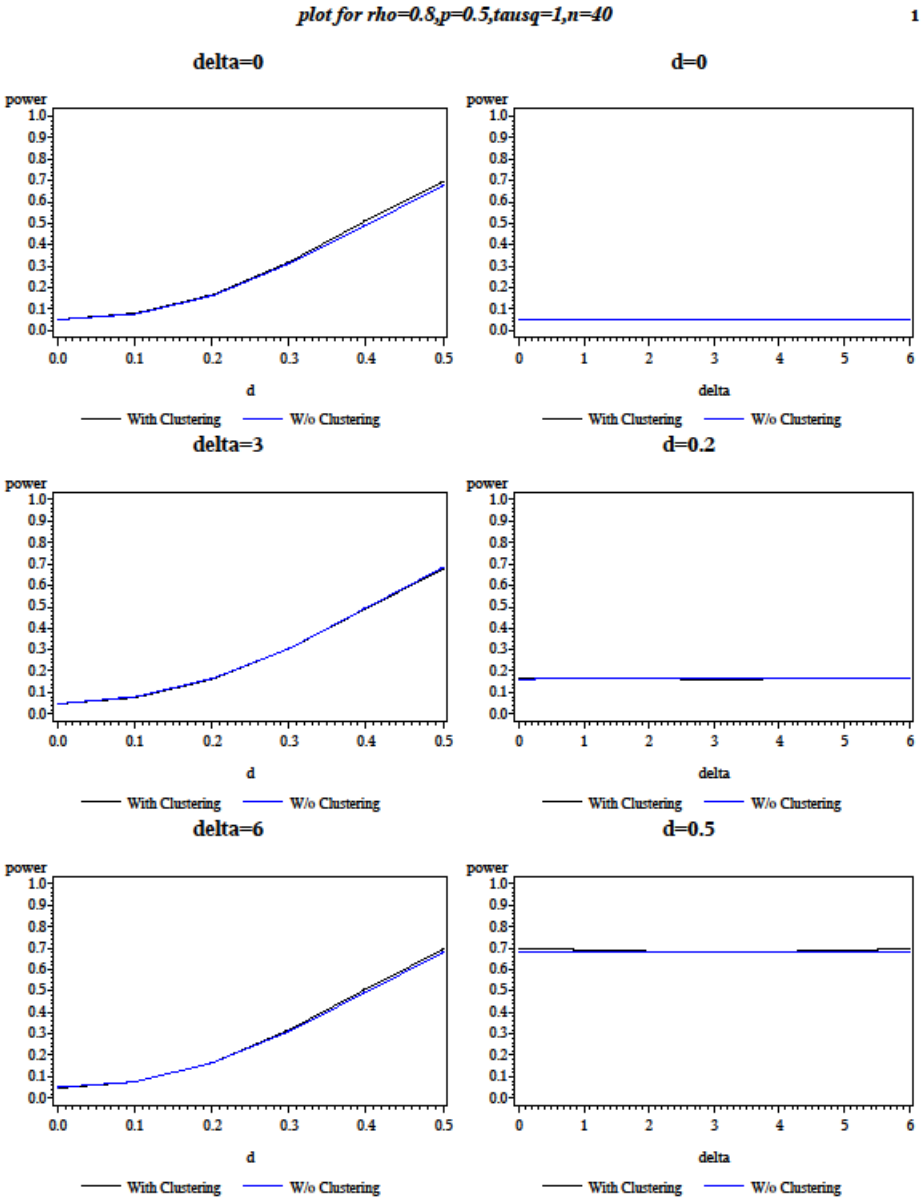


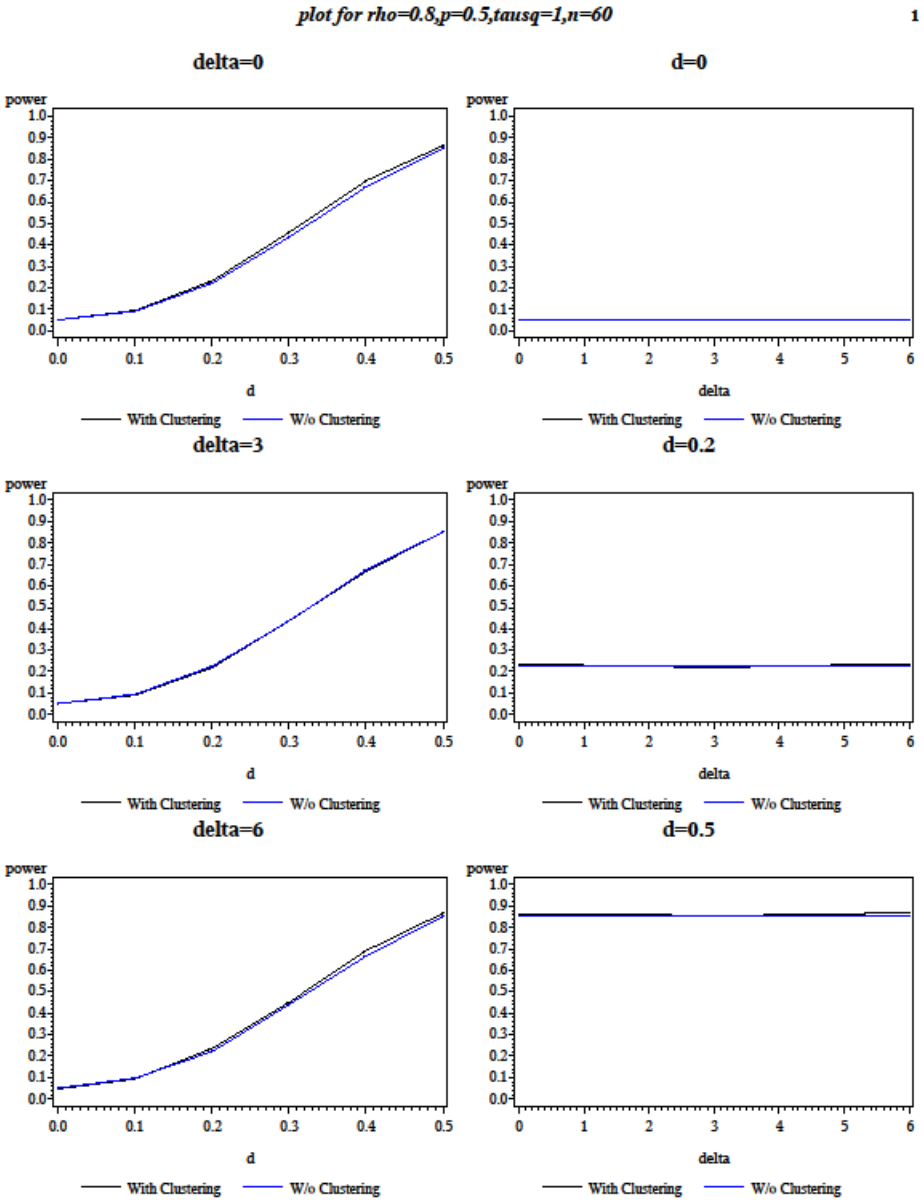


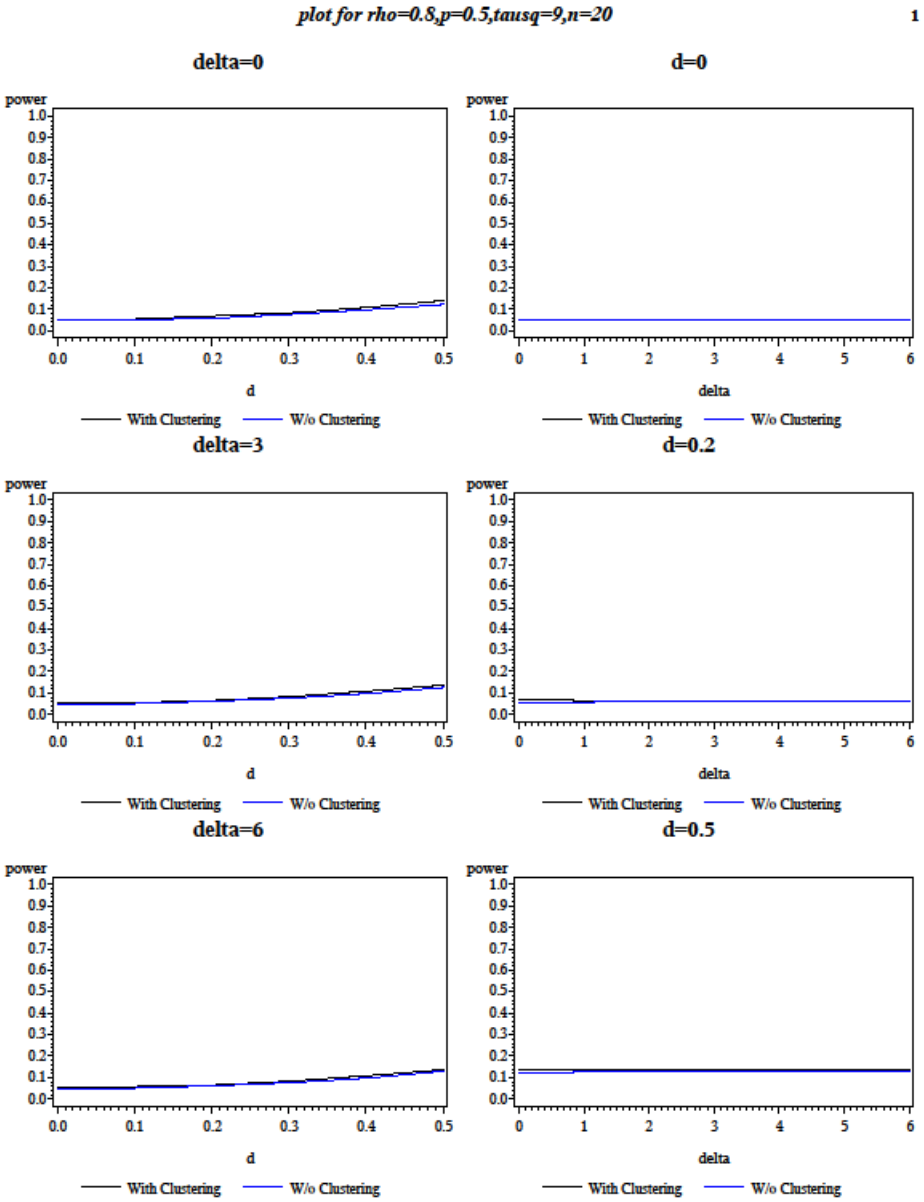


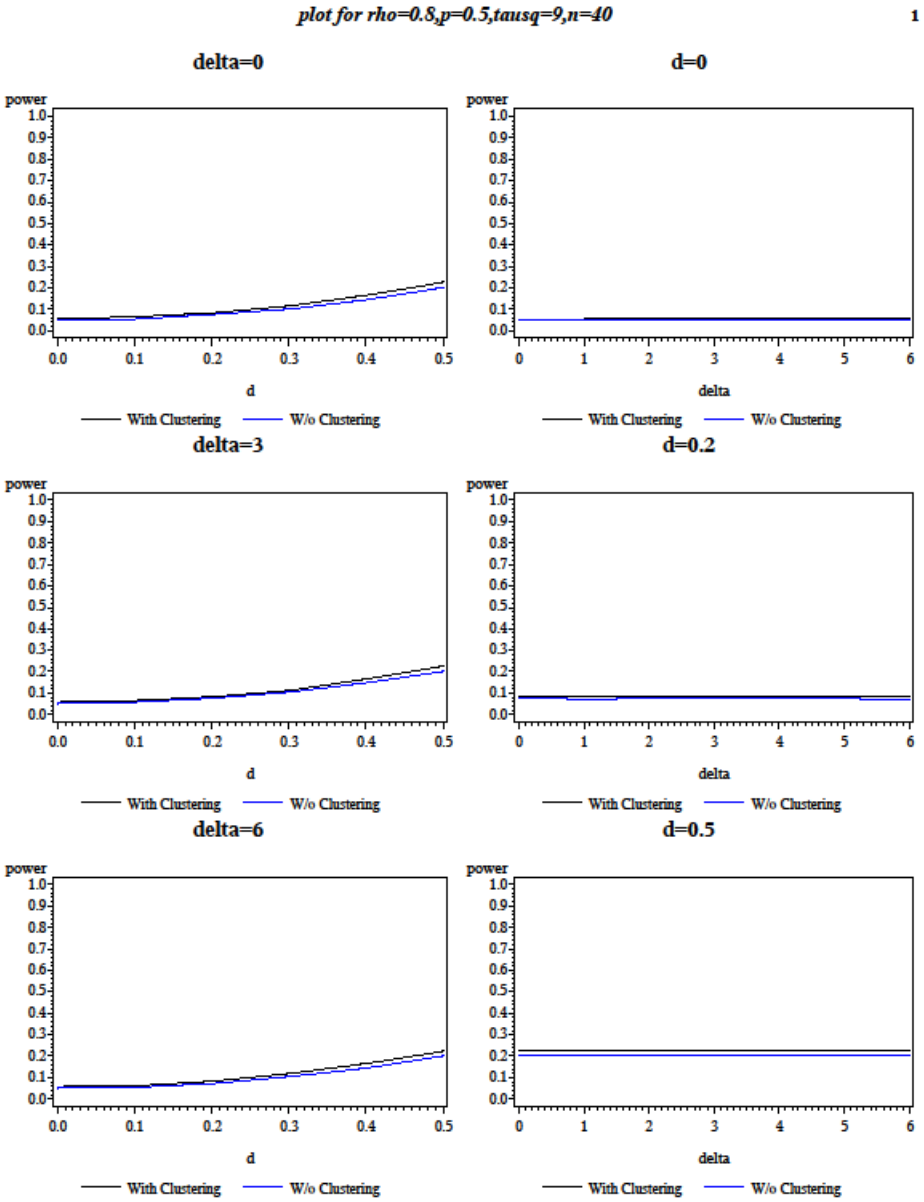


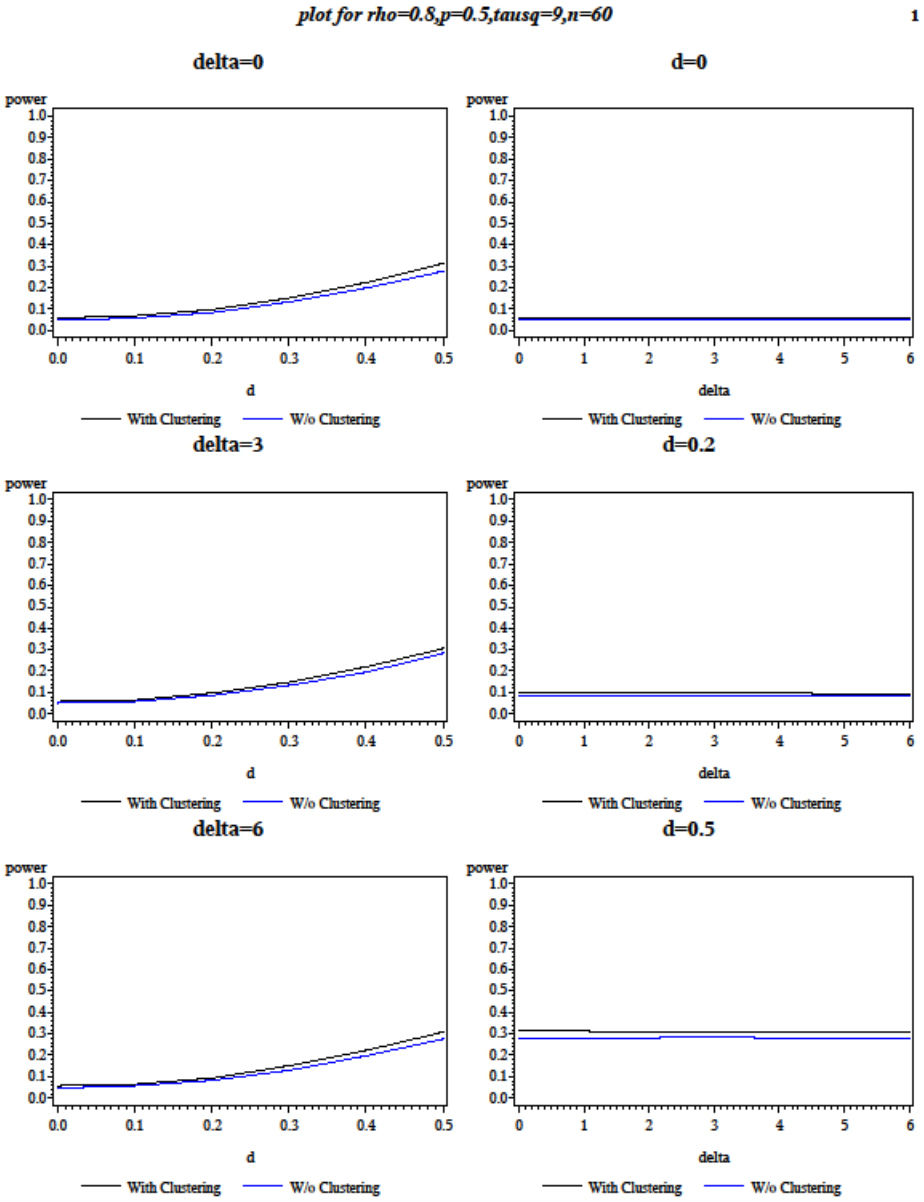












VITA

Fengjiao Hu was born January, 17, 1986 in Wuhan, China. Before entering school at Virginia Commonwealth University, she earned a master degree in Mathematics from Georgia Southern University in 2010 and a Bachelor degree in Mathematics from Huazhong Normal University in Wuhan, China in 2008.