2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Sample size calculations for tasting trials

Promotor :
dr. Francesca SOLMI
Prof. dr. Cristina SOTTO

Tewodros Getinet Yirtaw

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt
KNOWLEDGE IN ACTION
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt | Maastricht University

2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Sample size calculations for tasting trials

Promotor :
dr. Francesca SOLMI
Prof. dr. Cristina SOTTO

Tewodros Getinet Yirtaw

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

universiteit hasselt | Maastricht University

Certification

I declare that this thesis was written by me under the guidance and counsel of my supervisors.

..... Date.....
Tewodros Getinet Yirtaw Student

We certify that this is the true thesis report written by Tewodros Getinet Yirtaw under our supervision and we thus permit its presentation for assessment.

..... Date.....
Dr. Francesca Solmi Internal Supervisor

..... Date.....
Prof. dr. Cristina Sotto Internal Supervisor

Acknowledgements

First of all I would like to thank God with his mother Saint Virgin Marry, next my gratitude goes to my supervisor's Dr. Francesca Solmi and Prof.dr. Cristina Sotto for their guidance and constructive suggestions. My gratitude's will extend to Vlaamse InterUniversitaire Raad (VLIR) for giving me the chance to enhance my skills in the field of statistics. I would like to thank all my master courses professors in Uhasselt, especially Prof.dr. Paul Janssen. And also I would like to thank Mrs. Martine Machiels and Mr. Marc Tholen for their cooperation in every moment without hesitation.

Summary

In most experimental trials, determining the appropriate sample size in advance helps in saving resources. In this report, a simulation study to determine whether it is recommended or not to add extra subjects to future trials are conducted using the information coming from previous trials. The historical data are from tasting trials that are conducted in order to determine the palatability of some products and the preferences of consumers. In these trials, subjects are measured repeatedly over different experimental conditions (each tasting all the products) concerning the collection of a response variable, giving rise to correlated data. For this type of problem, the use of linear models assuming independence among observations taken from the same subject is not appropriate. In this report, random effect model that take the correlation among measurements of the same subject into account is fitted and used as input for the simulation study. Since, there are more than two groups to be compared the type I error rate increases as the number of comparisons increases, and to control this a Tukey type multiple comparison procedure is applied to test for mean differences. The power of the test to detect the mean differences via multiple comparison procedures is defined differently from that of a single comparison or testing procedure. There are a number of ways to define the power of the tests to detect mean differences in multiple comparisons, and in this report, the complete power is used to define the power.

As a conclusion, it is recommended to add extra subjects to future trials in order to detect mean differences of certain values with an acceptable power.

Key word: Linear mixed model, Multiple comparisons, Power, Simulation.

Contents page

List of Figures

List of Tables

Abbrevations

FWER	Family Wise Error Rate
LMM	Linear Mixed Model
g	gram
d	mean differences

1. Introduction

1.1. Background

A tasting trial is a tool used to gather information about the palatability of a product and the preferences of consumers. It is often used as a tool for companies to compare their brand to another brand. Tasting trials are also a tool sometimes used by companies to develop their brand or new product. They may be used by a company to ensure consistency, or differences among products. There are variety of other uses for a tasting trial, which is often carried out on the company level by professional tasters, who have trained to be impartial and valuable tools in the taste profiling process.

When a tasting trial is used to compare or contrast products, it is typically performed blind. In a blind tasting trial, the tasters do not know what they are tasting. They are offered samples of the product in identical presentations and asked to taste and profile the samples. In a double-blind tasting trial, the people offering the samples also do not know what they are. This is designed to ensure impartiality, making the end results potentially more valid.

To run a professional tasting trial, each taster is typically isolated in a compartment. The tasters usually wear no perfumes, and their clothing is laundered in neutral soaps. This is intended to minimize interference with the tasting trial. Usually a palate cleaner is provided as well, so that each taster can start fresh with each taste.

When a company is ready to launch a major product release, tasting trials are very important. A panel of tasters will ultimately determine the formulation of the product, by commenting on flavours they like and do not like. For companies which want to keep their products consistent, a panel of trained tasters familiar with their products is crucial.

In most experimental studies, determining the appropriate sample sizes in advance is of great importance, since it helps in identifying the number of subjects that are needed and sufficient for the study. If the number of subjects in a study is too small, the experiment may lack power to detect important differences (the study will easily result in a false negative conclusion), which would cause a waste of resources. On the other hand, the use of too large number of subjects may also lead to a waste of money, time and effort. For these reasons a

trial should always consider what number of subjects would be appropriate to answer the study question(s) in advance.

The power of a statistical test is the probability that the test will reject the null hypotheses when it is false (i.e. the probability of not making a false negative decision). The power is in general a function of the specific model fitted and test performed, and it depends on the values of the parameters of interest under the alternative hypotheses. The power is used to determine the required sample size that is needed to detect an important difference. When we have more than two groups in a study and we want to make multiple comparisons among those groups, the power is not defined as easily as it is in a single testing situation, since there are multiple parameters and null hypotheses (Westfall, Peter H. et al., 1999).

Calcluting the power of a test beforehand will help to ensure that the sample size is large enough to the purpose of the test. Otherwise, the test may be inconclusive, leading to wasted resources. The power of a test is influenced by sample size, effect size (mean differences), variability in the sample and significance level (alpha) of the test. Increasing all except the variability in the sample will lead to increase in the power. In addition to the aforementioned factors in this report the power is influenced by the number of products to be tasted and the corresponding exposures.

This report focuses on sample size calculations for tasting trials, having different products to be tasted, in order to observe the preferences of consumers. The sample size calculations are done based on the information from historical data (previous trials). The historical data consist of measurements taken from the same subjects repeatedly under different conditions (each subject taking all the products). This leads to a repeated measures design, so the performed analysis must take into account the correlated nature of the data. To account for the correlated nature of the data, a mixed effect model (Laird and Ware, 1982) is applied, which can be considered as an extension of the classical linear model.

Using the results on the historical data as input, a simulation study is performed to determine the appropriate sample sizes for future trials. The powers are computed using the complete power definition for detecting differences among multiple comparisons. In this case multiple comparisons are performed applying Tukey correction.

Section 1.2 introduces the objective(s) of this report followed by the description of the data from the previous trials in section 1.3. The statistical methodology is presented in section 2,

and section 3 covers the results obtained in the simulation study. Finally, conclusions are reported in section 4.

1.2. Objective of the study

Tasting trials are conducted in order to observe the preferences of consumers about the tasted products. These are conducted on 30 subjects each tasting all the products an equal number of times (exposures). The tasting order was randomized over the subjects. The aim of this report is to determine whether it is recommended or not to enrol extra subjects to future trials. In addition to that recommendation on the possible number of exposures per subject is provided, also taking into account the number of products to be tasted in a given trial.

1.3. Data Description

The previous trials inputs for the simulation study are repeated measures studies. In each trial repeated observations are recorded on 30 subjects. In total 17 trials are carried out in order to identify the palatability of the products and the preferences of consumers. In the trials there are m products each to be tasted by the subjects k equal number of times (exposures). Table 1, presents the trials with the corresponding number of products and exposures. For each exposure, the same amount of the products are given to each subject and the amount eaten is recorded. The amount eaten is the response of interest, measured as a continuous outcome (in grams). The products are fed to the subjects using a Latin square design. The preferences are determined by the amount eaten, i.e. the most eaten product is the most preferred.

Table 1. List of trials with the corresponding number of products and exposures

Trial	Products(m)	Exposures(k)
1 to 4	2 (A,B)	2
5 to 6	2 (A,B)	3
7 to 12	4 (A,B,C,D)	2
13 to 14	4 (A,B,C,D)	3
15 to 16	6 (A,B,C,D,E,F)	2
17	6 (A,B,C,D,E,F)	3

2. Methodology

In many experimental trials, subjects are measured repeatedly over time and under different experimental conditions regarding the collection of response variable, which gives rise to correlated data. Here, a statistical model which handles the correlated nature of the data for a continuous response variable is discussed.

2.1. Mixed Models

Mixed effects models are statistical models that incorporate both fixed and random effects. A factor is said to be fixed if its levels in the study represent all levels of interest of the factor, or at least all levels that are important for inference. Therefore, the factors are selected with the purpose of comparing the effects of the levels to each other. On the other hand a factor is said to be random if the levels of the factor in the study can be viewed as random sample from a population of factor levels. Mixed models, like many other statistical models, describe the relationship between a response variable (or dependent variable) and one or more independent variables that are measured or observed alongside the response.

In this study, there are two factors: one fixed (the products to be tasted) and one random (the subject), and a corresponding continuous response variable is the amount eaten. The linear mixed model (LMM) is chosen for this analysis. Let Y_{ijl} be the continuous response variable for the j^{th} subject tasting the i^{th} product at l^{th} occasion (exposures). The corresponding LMM for the trials is given by (Laird and Ware, 1982):

$$Y_{ijl} = \mu + \tau_i + \gamma_j + \varepsilon_{ijl}, \quad i= 1,2,\ldots,m; \quad j= 1,\ldots,n; \quad l= 1,2,\ldots k \qquad (1)$$

Where, μ is the overall mean, τ_i is the effect of the i^{th} level product, γ_j is the random effect for the j^{th} subject ($\gamma_j \sim N(0,\sigma_\gamma^2)$) and ε_{ijl} is the random error term ($\varepsilon_{ijl} \sim N(0, \sigma_\varepsilon^2)$). The γ_j and ε_{ijl} are independent. For further reading on repeated measures with multiple replications per cell, refer to M. Mushfiqur Rashid and Ansuman Bagchi (1997). The indices *i* and *l* in (1) can range from 1 to respectively 2,4, or 6 and 2 or 3, according to the number of products and of exposures.

2.2. Multiple Comparisons

When there are more than two groups to be compared, next to the classical F test that detects whether there is a significant difference among the factor level means, it is often of interest to perform tests to detect which groups are different. Multiple comparison procedures are methods that allow us to compare the different groups and to detect which differences exist. The global type I error (the probability of rejecting the global null hypothesis when it is true) increases when performing multiple comparisons or tests simultaneously, at a given significance level. In particular, the actual global significance level is not controlled at the nominal value, and it increases as the number of comparisons increases. The global type I error for multiple comparisons is known as family wise error rate (FWER). There are a lot of multiple comparisons techniques which are designed to control the family wise error rate at a certain nominal level. Among them the Tukey multiple comparison procedure is a standard choice when interest is on all pair wise comparisons. Moreover, when the sample sizes per group are equal, the method controls the FWER on an exact way.

In this study the sample sizes per group are equal and the interest is on all pairwise comparisons. Therefore, the Tukey multiple comparison procedure, that uses the studentized range distribution to control the family wise error rate, is applied here. For more information on multiple comparison procedures, refer to Hochberg and Tamhane (1987).

2.3. Power

In single comparison or testing situations the power is defined as the probability of rejecting the null hypotheses when it is actually false. When we perform multiple comparisons, we are faced with a number of null hypotheses. In this case the definition of power for a single test becomes too restrictive, and we need an extension to deal with the multiplicity of the problem. The definition of power will now depend on the number of false null hypotheses that we need to reject. The definitions include (Westfall, Peter H. et al., 1999):

Complete Power is the probability that the test will reject all false null hypotheses (to reject the null hypotheses all tests related to a comparison under the alternative hypotheses should be significant).

Minimal Power is the probability that the test will reject at least one false null hypotheses.

Individual Power is the probability to reject a particular (single) null hypotheses that is false.

Proportional Power is the average proportion of false nulls that are rejected.

As mentioned earlier the interest is on all pairwise comparisons of the means. These means the comparisons consist of all tests of the form:

$$H_{0il}: \mu_i - \mu_{i'} = 0 \text{ versus } H_{ail}: \mu_i - \mu_{i'} \neq 0 \qquad (i = 1,2,\ldots,m;\ l = 1,2,\ldots k) \qquad (2)$$

The global null and alternative hypotheses for the complete power can be formulated from (2) as follows:

$$H_{0l}: \bigcup_{i=1}^{m} H_{0il} \text{ versus } H_{al}: \bigcap_{i=1}^{m} H_{ail} \qquad (3)$$

The indices i and l in (2) and (3) can range from 1 to respectively 2,4, or 6 and 2 or 3, according to the number of products and of exposures. The hypotheses involves testing the union of the the global null hypotheses (H_{0l}) against the intersection of the alternative hypotheses (H_{al}).

In this report the Tukey multiple comparison procedure is used to perform all pairwise comparisons and a family of pairwise comparisons is considered as significant, if every pairwise comparison in the family is significant. In other words the global null hypotheses (H_{0l}) is rejected if all H_{ail} are significant. Therefore, in this report the power is defined as a complete power.

2.4. Simulation Study

In order to run the simulation, for determining appropriate sample sizes for future trials, data generation is a crucial step. The data generation is performed based on the formulation of (1). To obtain the inputs for the data generation, (1) is fitted using the lmer function in the R software on all 17 trials. For each trial, the means for the different products and the variances (error and random subject) are obtained from the model estimates. The mean differences that needs to be recognized in future trials with an acceptable power (above 80%) are given by the field experts, and range between 2.5 to 15 g.

The previous trials are conducted on 30 subjects, and as mentioned earlier, the aim of this study is to determine whether it is recommendable to add extra subjects to future trials. The company is willing to add up to 25 extra subjects hence, sample sizes considered in the simulation range from 30 to 55 for each trial.

In order to generate the data using the formulation in (1), overall mean, product effects (which represent the mean differences), random subjects effects (the same for all observation coming from the same subject) and random error term (unique for each observation) are

required. Before discussing how to obtain these required values, one thing to notice is that the data generation is performed for each of the six products by exposure combination (Table 1).

For the simulation, the overall means for each combination are obtained as an average of the estimated means for all the products, but the last levels, B, D and F, respectively, in the case of 2,4 and 6 tasted products. Then the means of the last products (B, D and F) excluded from the computation of the overall means for each combination are built as the overall mean plus the differences (2.5 to 15 g), while the other products means are taken equal to the overall mean. Hence, the mean differences are between the last products (B, D and F) and the previous ones, and not among the previous products themselves. The means of the random error and random subject variances for each combination from the corresponding trials are used to generate the random error terms and random subjects effect from the normal distribution with mean zero and corresponding variances respectively.

Table 2, presents values for the overall means and variances used to generate the random error terms and random subjects effect that are used as an input for the simulation study in their corresponding product-by-exposure combination. Values of the mean differences range from 2.5 g to 15 g by 2.5 g (i.e. 2.5, 5, 7.5, 10, 12.5 and 15 g) and the sample sizes range from 30 to 55 by 5 (i.e. 30, 35, 40, 45, 50 and 55 subjects) for each product-by-exposure combination. Finally, the data are generated, for each product-by-exposure combination, by changing the mean differences for every single sample size. In addition to this in order to see the effects of the variances, the data generation is repeated by reducing the estimated variances to half and double respectively.

Table 2. Inputs for data generation with the corresponding combinations

Products(m)	**Exposures(k)**	**Means**	σ_γ^2	σ_ε^2
2	2	87	20	68
2	3	85	24	68
4	2	90	40	102
4	3	84	34	97
6	2	89	91	177
6	3	90	87	160

After the data are generated a LMM is fitted, as formulated in (1) using the lmer function, and the Tukey multiple comparison procedure is performed using the glht function in the R software. Finally, the powers are estimated, using the complete power definition to determine appropriate sample sizes for future trials. For every single combination 1000 data sets are simulated.

3. Results

3.1. Results of the Simulation Study

In this section the results of the simulation study for each product-by-exposure combination are presented and described. In this paper powers obtained to detect the significant mean differences are classified as either insufficient (below 80%) or acceptable (above 80%). Insufficient powers are not used to determine sample sizes for future trials. Since, would cause a waste of resources. The sample sizes are determined based on acceptable powers (80 to 90%) since, higher powers could also cause waste of resources.

Figure 1, displays the power estimates as the sample size increases, for the different variances (halved, estimated and doubled) respectively, for the combination Product=2 and Exposure=2 and the plots reveal that, when half of the estimated variances are used, an acceptable power of about 80% to detect a mean difference of d=2.5 g is reached with the sample size of about 45 subjects. For almost all the remaining higher mean differences, the complete powers are estimated as 100% for every sample size. This indicates that in all the simulation runs the tests detect the mean differences. Instead, when the estimated variances are used, with an acceptable power of about 90%, a mean difference of d=5 g is detected for a sample size of about 30. The powers obtained are insufficient, to detect a mean difference of d=2.5 g. When the estimated variances are doubled, the powers are insufficient to detect a mean difference of d=2.5 g, but for a mean difference of d=5 g, an acceptable power of about 80% is reached for a sample size of about 45 subjects.

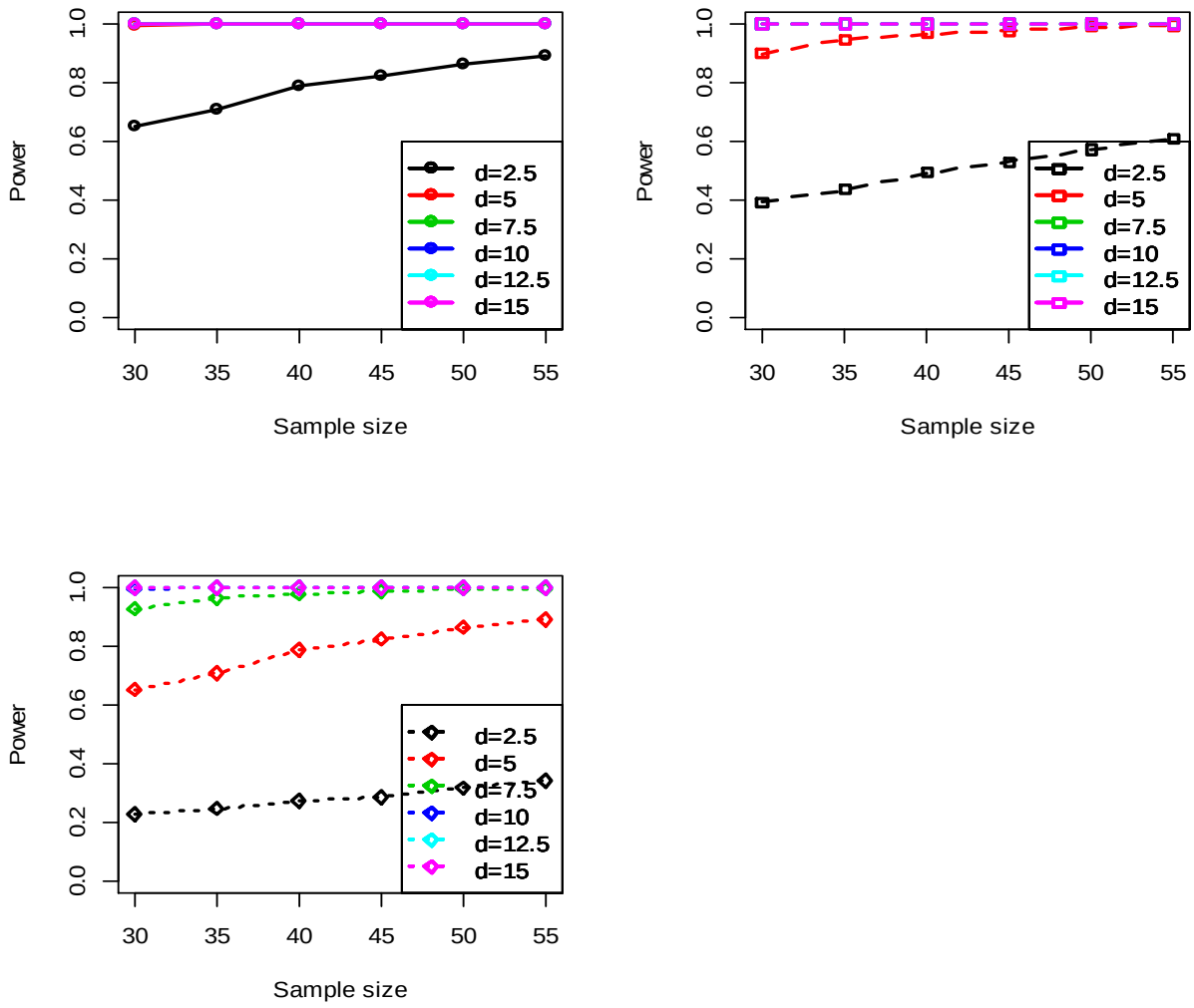

Figure 1: Power by sample size for Product= 2 and Exposure= 2 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

Results for the combination Product=2 and Exposure=3 are displayed in Figure 2. They reveal that, when the estimated variances and half of them are used, with an acceptable power of about 80% a mean difference of d=2.5 g is detected for sample sizes of about 55 and 30 subjects respectively. Instead, when the estimated variances are doubled an acceptable power of about 80% to detect a mean difference of d=5 g is reached for a sample size of about 30, while to detect a mean difference of d=2.5 g, the powers obtained are insufficient.

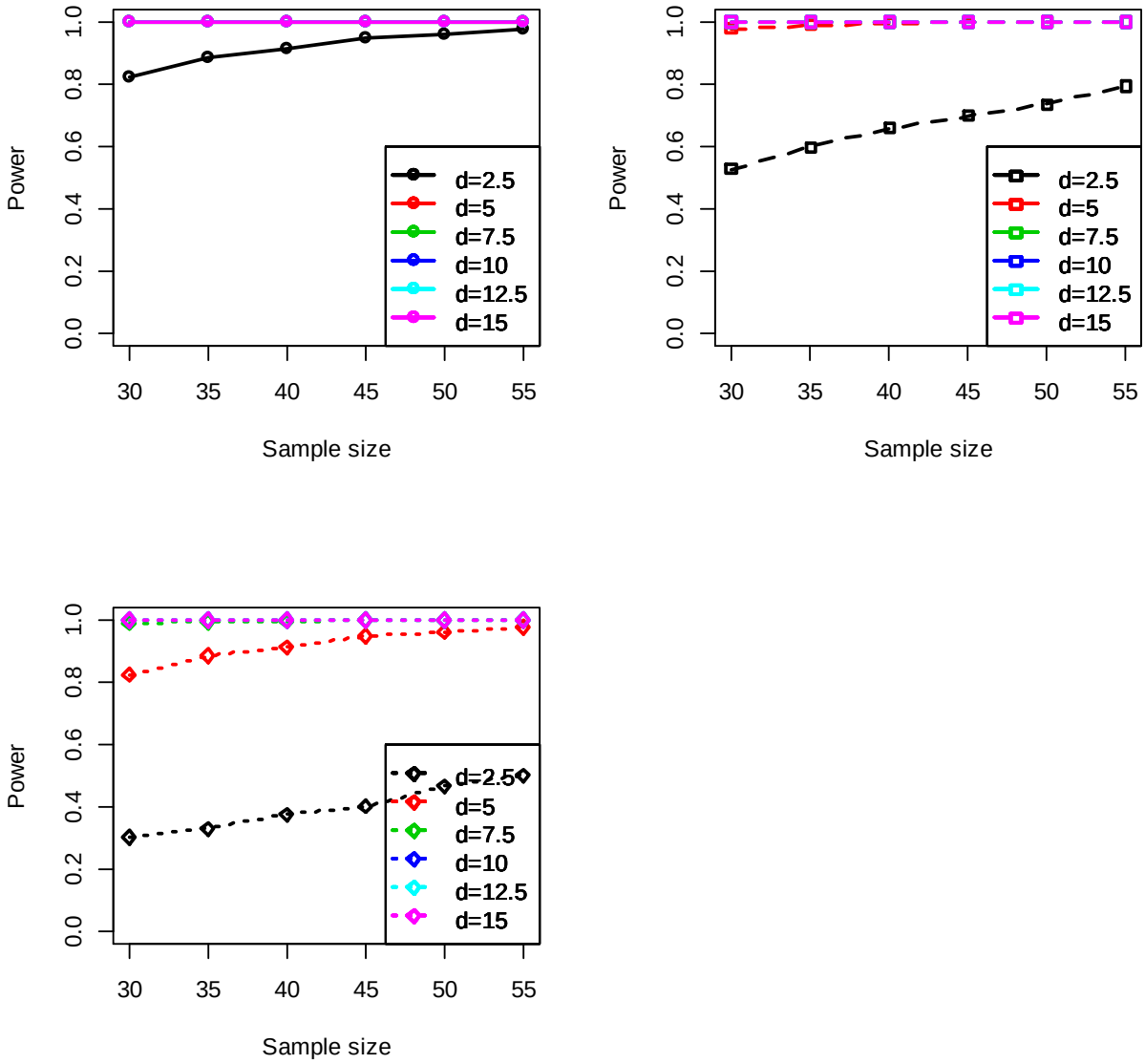

Figure 2: Power by sample size for Product= 2 and Exposure= 3 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

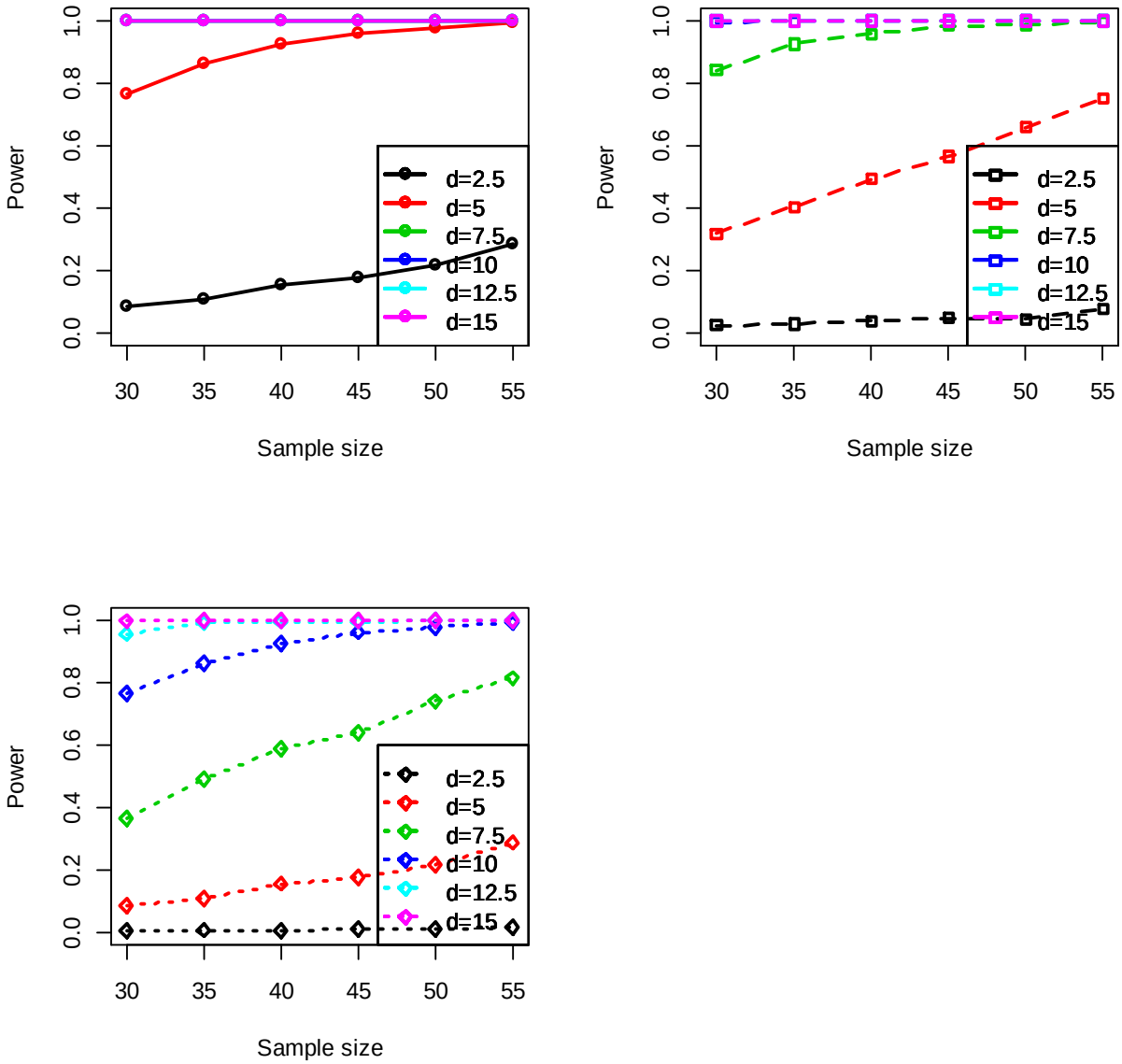

Figure 3: Power by sample size for Product= 4 and Exposure= 2 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

Figure 3, displays the results for the combination of Product=4 and Exposure=2. They reveal that, when half of the estimated variances are used, to detect a mean difference of d=5 g, an acceptable power of about 90% is reached for a sample size of about 35 subjects. The obtained powers are insufficient to detect a mean difference of d=2.5 g. Instead, when the estimated variances are used, an acceptable power of about 80% to detect a mean difference of d=7.5 g is obtained for a sample size of about 30 subjects. Powers obtained are insufficient to detect the remaining lower mean differences. Doubling the estimated variances results in

reaching an acceptable power of about 80% to detect a mean difference of d=10 g for a sample size of about 35 subjects. Obtained powers are insufficient to detect the remaining lower mean differences.

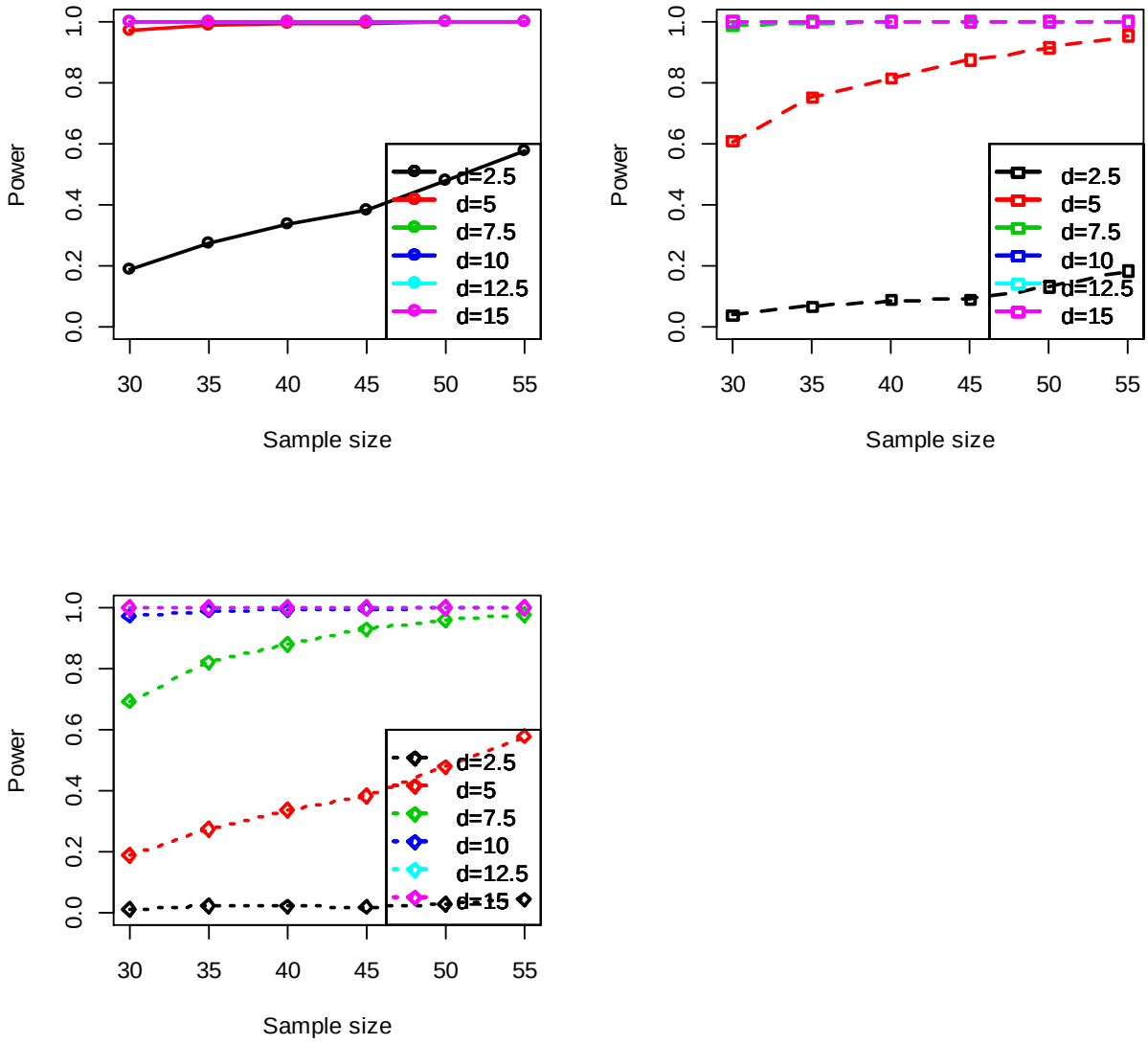

Figure 4: Power by sample size for Product= 4 and Exposure= 3 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

Let us look at results of the combination Product=4 and Exposure=3 displayed in Figure 4. When the estimated variances reduce to half, the obtained powers are insufficient to detect a mean difference of d=2.5 g. Results for the estimated variances indicate that a mean difference of d=5 g can be detected with an acceptable power of about 80% for a sample size

of about 40 subjects, while to detect a mean difference of d=2.5 g, powers obtained are insufficient for any of the sample sizes. Instead, when the estimated variances are doubled, only a mean difference of d=7.5 g is detected with an acceptable power of about 80% for a sample size of about 35 subjects, and all the remaining lower mean differences are detected with insufficient powers.

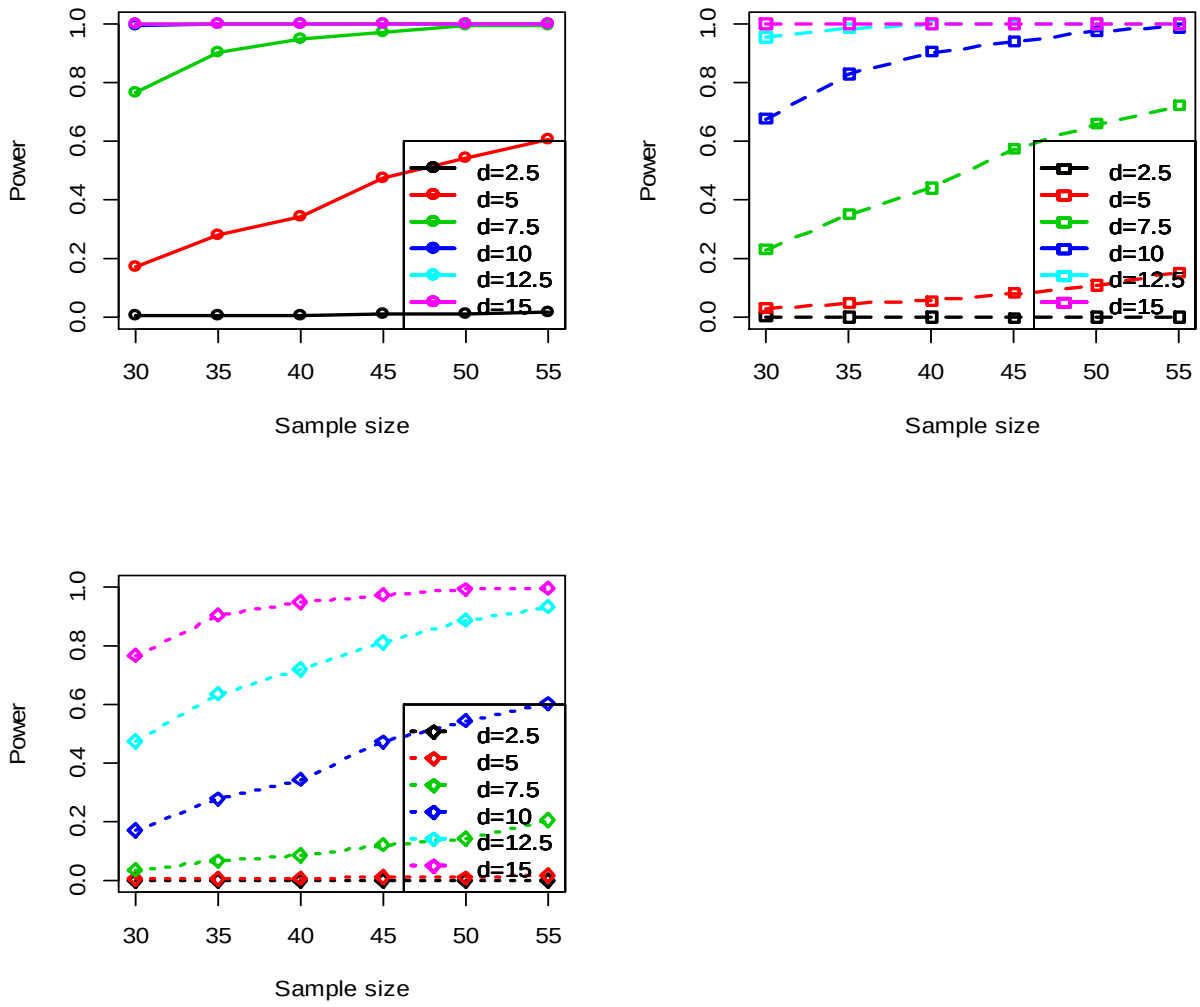

Figure 5: Power by sample size for Product= 6 and Exposure= 2 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

Results of the combination Product=6 and Exposure=2 are displayed in Figure 5. They reveal that, reducing the estimated variances to half, leads to the detection of a mean difference of d=7.5 g, with an acceptable power of about 90% for a sample size of about 35 subjects, while all the remaining lower mean differences are detected with insufficient powers. The use of estimated variances leads to the detection of a mean difference of d=10 g with an acceptable

power of about 80% for a sample size of about 35 subjects, while all the remaining lower mean differences are detected with insufficient powers. Instead, when the estimated variances are doubled, an acceptable powers of about 80% and 90% are reached to detect the mean differences of d=12.5 g and d=15 g respectively with sample sizes of about 45 and 35 subjects. Insufficient powers are reached to detect all the remaining lower mean differences.

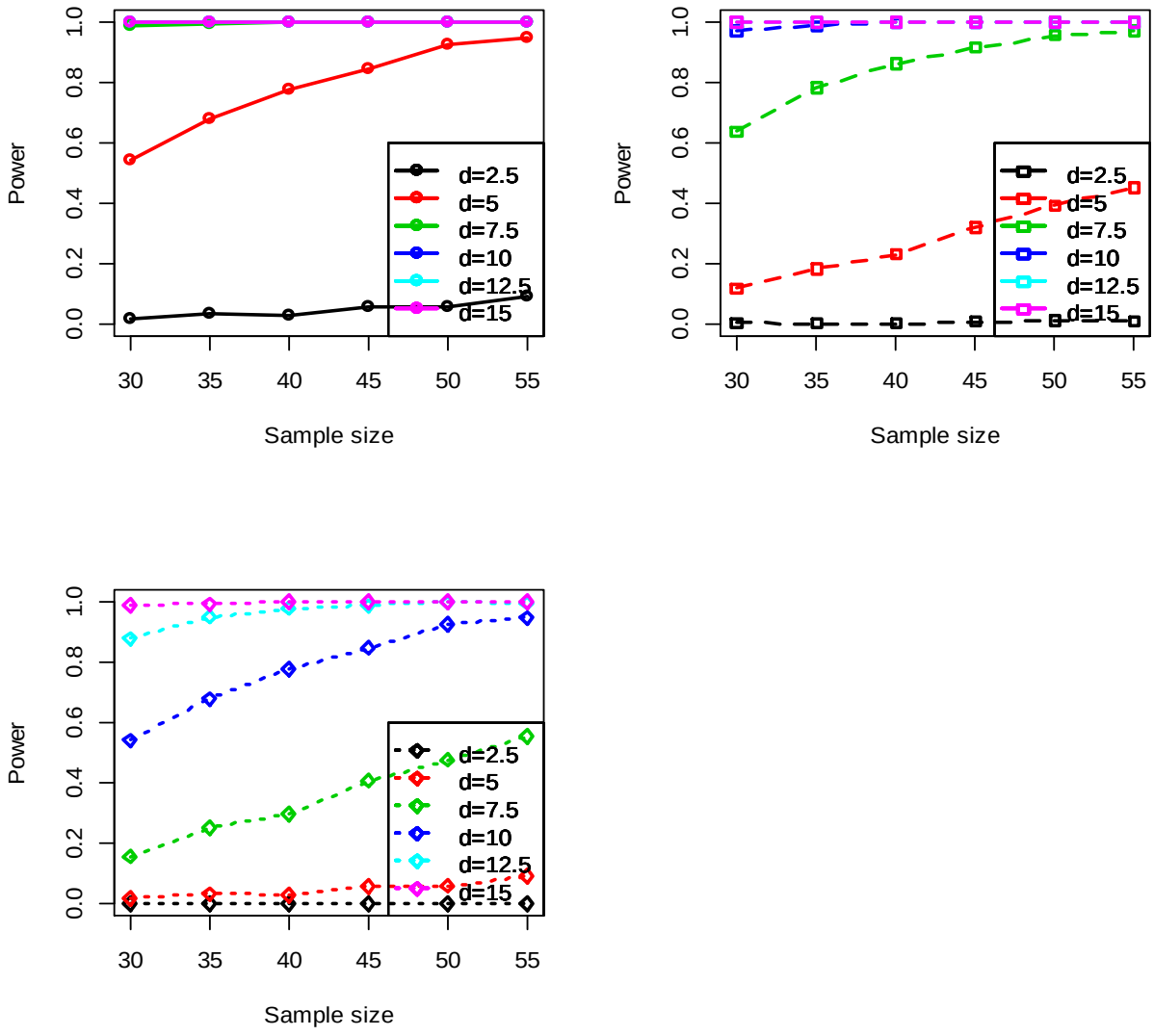

Figure 6: Power by sample size for Product= 6 and Exposure= 3 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

Results regarding the combination of Product=6 and Exposure=3 displayed in Figure 6. They reveal that, when half of the estimated variances are used, to detect a mean difference of d=5 g an acceptable power of about 80% is reached for a sample size of about 45 subjects, while

obtained powers are insufficient to detect a mean difference of d=2.5 g. When estimated variances are used, an acceptable power of about 90% to detect a mean difference of d=7.5 g is obtained for a sample size of about 40 subjects, while powers obtained are insufficient to detect the mean differences of d=2.5 g and d=5 g. Finally, when the estimated variances are doubled a mean difference of d=10 g is detected with an acceptable power of about 80% for a sample size of about 45 subjects, and also an acceptable power of about 90% is obtained for a sample size of about 30 subjects to detect a mean difference of d=12.5 g. The powers obtained are insufficient to detect all the remaining lower mean differences.

4. Discussion and Conclusions

In this section the results of the simulation study, that are presented in section 3, are used as a basis to give concluding remarks for future trials. In order to see the effect of variances on the sample size determination, as mentioned earlier, estimated variances are reduced to half and doubled. Results showed that, despite, the similar powers obtained in each combination, in general, as the variances increase the appropriate sample sizes required to detect the mean differences also increase, which means that, more subjects are needed as the variances increase.

The recommended sample sizes based on the simulation study are summarized using tables for each product-by-exposure combination. Table 3, Table 4 and Table 5, displays the required sample sizes for future trials, to detect the corresponding mean differences in each product-by-exposure combination for the different variances.

Table 3. Recommended sample sizes with the corresponding mean differences for half of estimated variances

Products(m)	Exposures(k)	differences(d)	Sample sizes
2	2	2.5	45
2	3	2.5	30
4	2	7.5	35
4	3	-	-
6	2	10	35
6	3	5	45

Table 4. Recommended sample sizes with the corresponding mean differences for estimated variances

Products(m)	Exposures(k)	differences(d)	Sample sizes
2	2	5	30
2	3	2.5	55
4	2	7.5	30
4	3	10	40
6	2	10	35
6	3	7.5	40

Table 5. Recommended sample sizes with the corresponding mean differences for doubled estimated variances

Products(m)	Exposures(k)	differences(d)	Sample sizes
2	2	5	45
2	3	5	30
4	2	10	35
4	3	7.5	35
6	2	12.5 & 15	45 & 35
6	3	10 & 12.5	45 &3 0

From the results of the simulation study, regardless of, similar powers obtained to detect some of the mean differences in each combination. In general, powers decrease as the number of tasted products increase, which means that, as the number of tasted products increase more subjects are needed to detect the mean differences. Unlike the number of products tasted, powers increase as the exposure increases, that is, as the exposure increase fewer subjects are needed.

The conclusions given on the number of tasted products and exposures might be influenced by the change in values of the variances in each combination. That is, in each combination the variances are different and increase with the number of tasted products and exposures (Table 1). In order to clear this suspicion, the simulation study were repeated, using equal values for the overall mean and variances in each combination as an input. The overall mean and variances are obtained by computing the respective averages in Table 1. Results are displayed in appendix; shows that, powers decrease as the number of products to be tasted increase, and powers increase as the number of exposures increase. That is, more subjects are needed, when the number of products to be tasted increase and less subjects are required, when the exposures increase.

As recommendation, if the interest is in detecting all or some of the mean differences that are not detected with an acceptable power with a range of sample sizes considered in this study, then the possibility of adding more subjects (than 55) should be looked at.

References

Douglas M. Bates. 2010. Lme4 : Mixed-effects modelling with R. Springer.

Duchateau L., Janssen P. and Rowlands G.J. 1998. *Linear mixed models. An introduction with applications in veterinary research.* ILRI (International Livestock Research Institute), Nairobi, Kenya. 159 pp.

Frank Bretz, Torsten Hothorn, and Peter Westfall. 2011. Multiple Comparisons Using R. Chapman & Hall.

George Zhengzhi Xia. 2009. Multiple Comparisons for Mixed Models. VDM.

Hochberg, Y., and A.C. Tamhane. 1987. Multiple Comparison Procedures. New York: Wiley.

Jerry L. Hintze. 2011. PASS User's Guide III: Power Analysis and Sample Size System. NCSS.

Littell, Ramon C., George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, and Oliver Schabenberger. 2006. SAS for Mixed Models, Second Edition. Cary, NC: SAS Institute Inc.

Michael H. Kutner, Christopher J. Nachtsheim, Jhon Neter, and William Li. 2005. Applied Linear Statistical Models, Fifth Edition. Mc Graw-Hill Irwin.

M. Mushfiqur Rashid and Ansuman Bagchi. 1997. Robust Analysis of One-Way Repeated Measures Designs With Multiple Replications Per Cell. Statistica Sinica 7, 647-667.

Nan M. Laird, and James H. Ware. 1982. Random-Effects Models for Longitudinal Data. Biometrics, Vol. 38, No.4.

Verbeke, G. and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. New York, Springer.

Wendy Bergerud and Vera Sit. (2001). Power Analysis Wrkshop. Biometrics Information.

Westfall, Peter H., Randall D. Tobias, Dror Rom, Russell D. Wolfinger, and Yosef Hochberg. 1999. Multiple Comparisons and Multiple Tests Using SAS. Cary, NC: SAS Institute Inc.

Appendix

Appendix A: Figures (Based on equall means and variances for each combination)

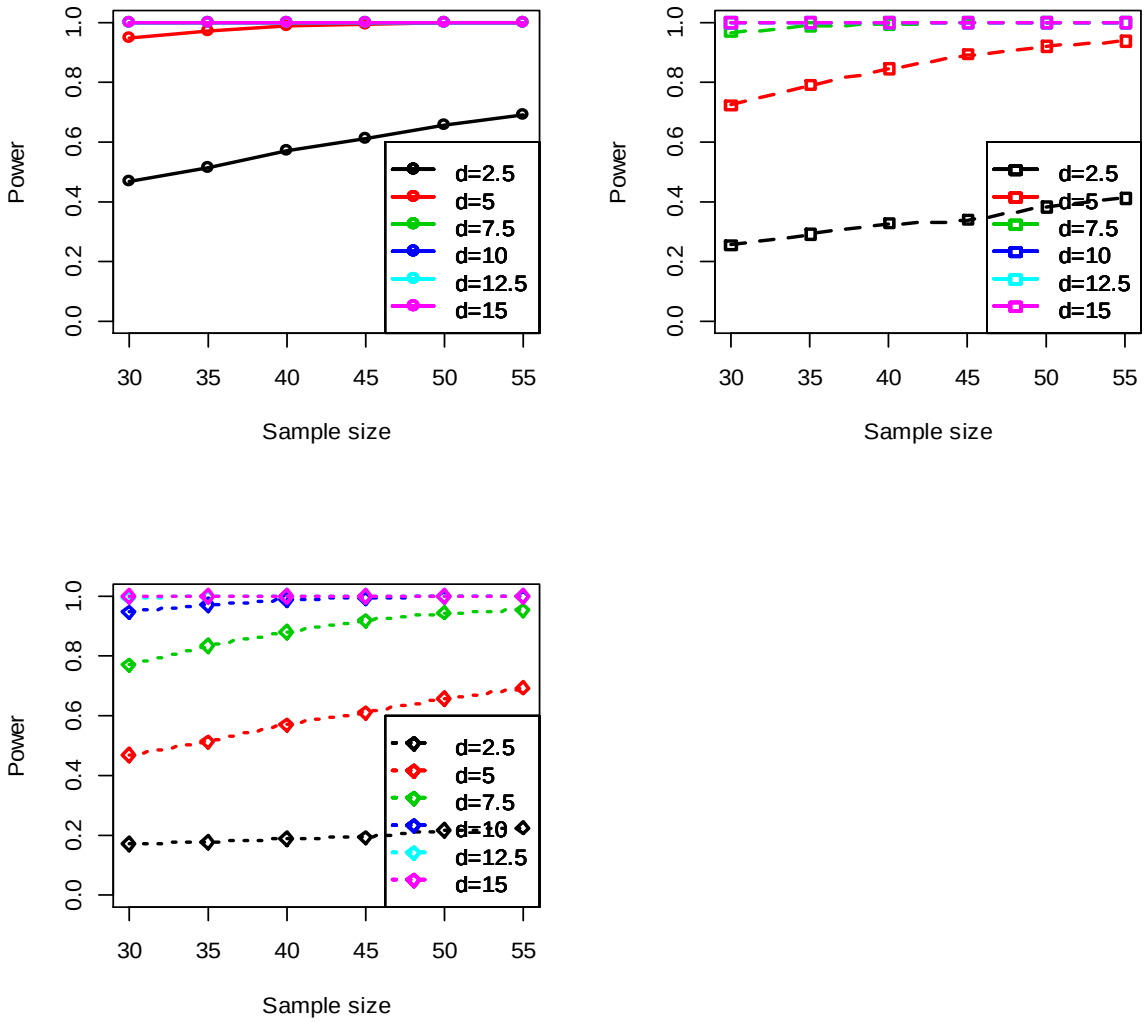

Figure 7: Power by sample size for Product= 2 and Exposure= 2 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

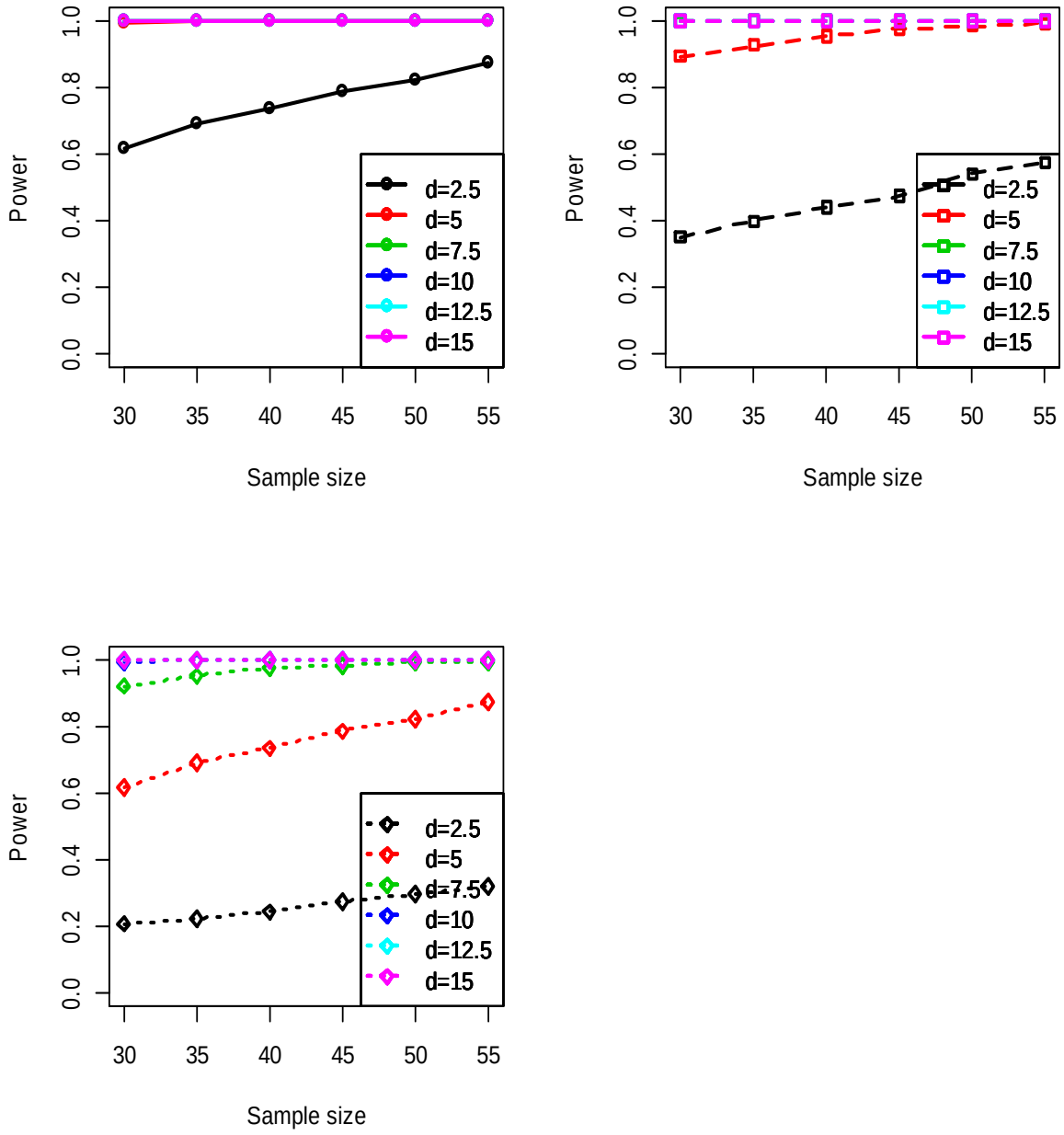

Figure 8: Power by sample size for Product= 2 and Exposure= 3 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

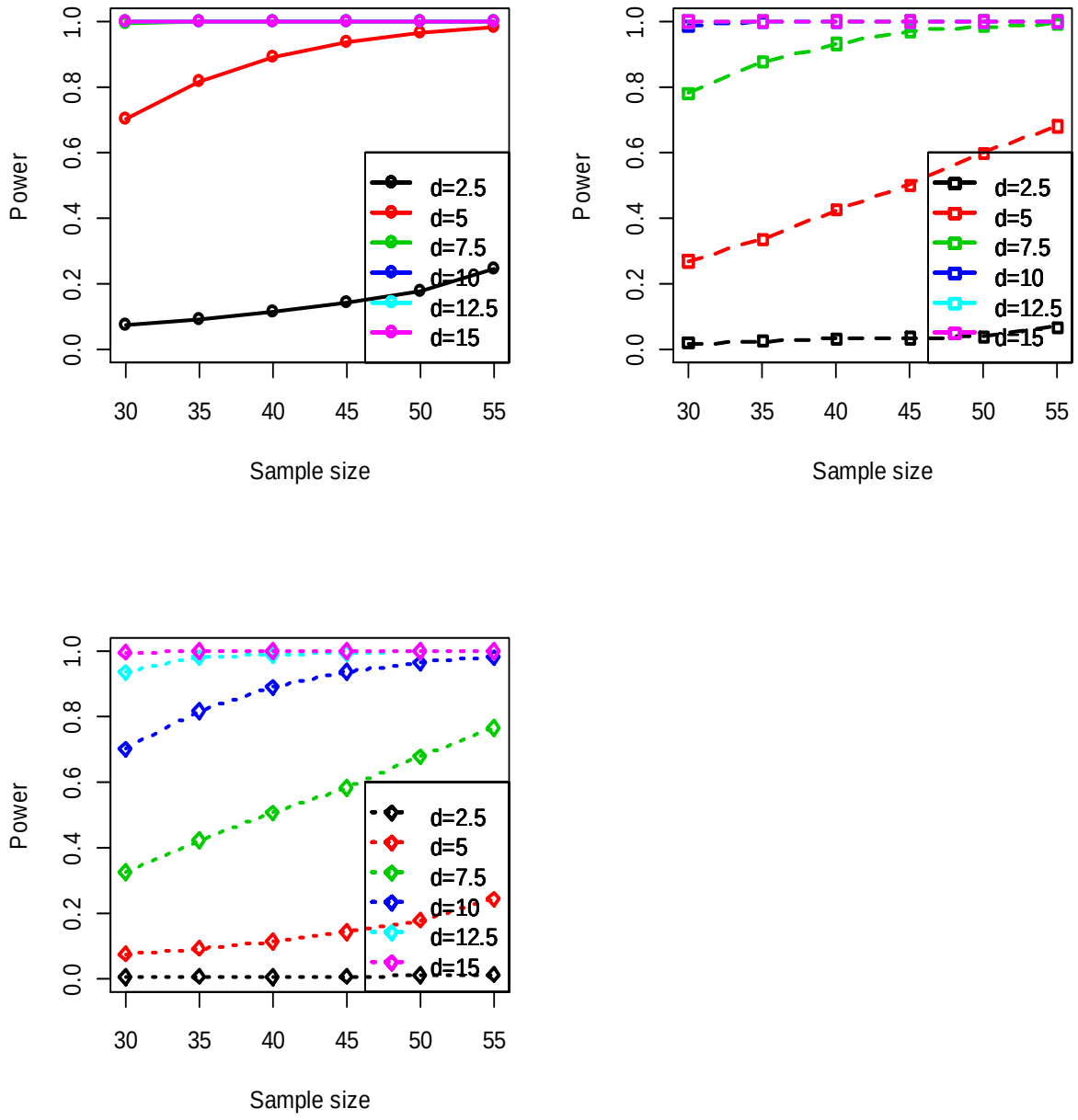

Figure 9: Power by sample size for Product= 4 and Exposure= 2 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

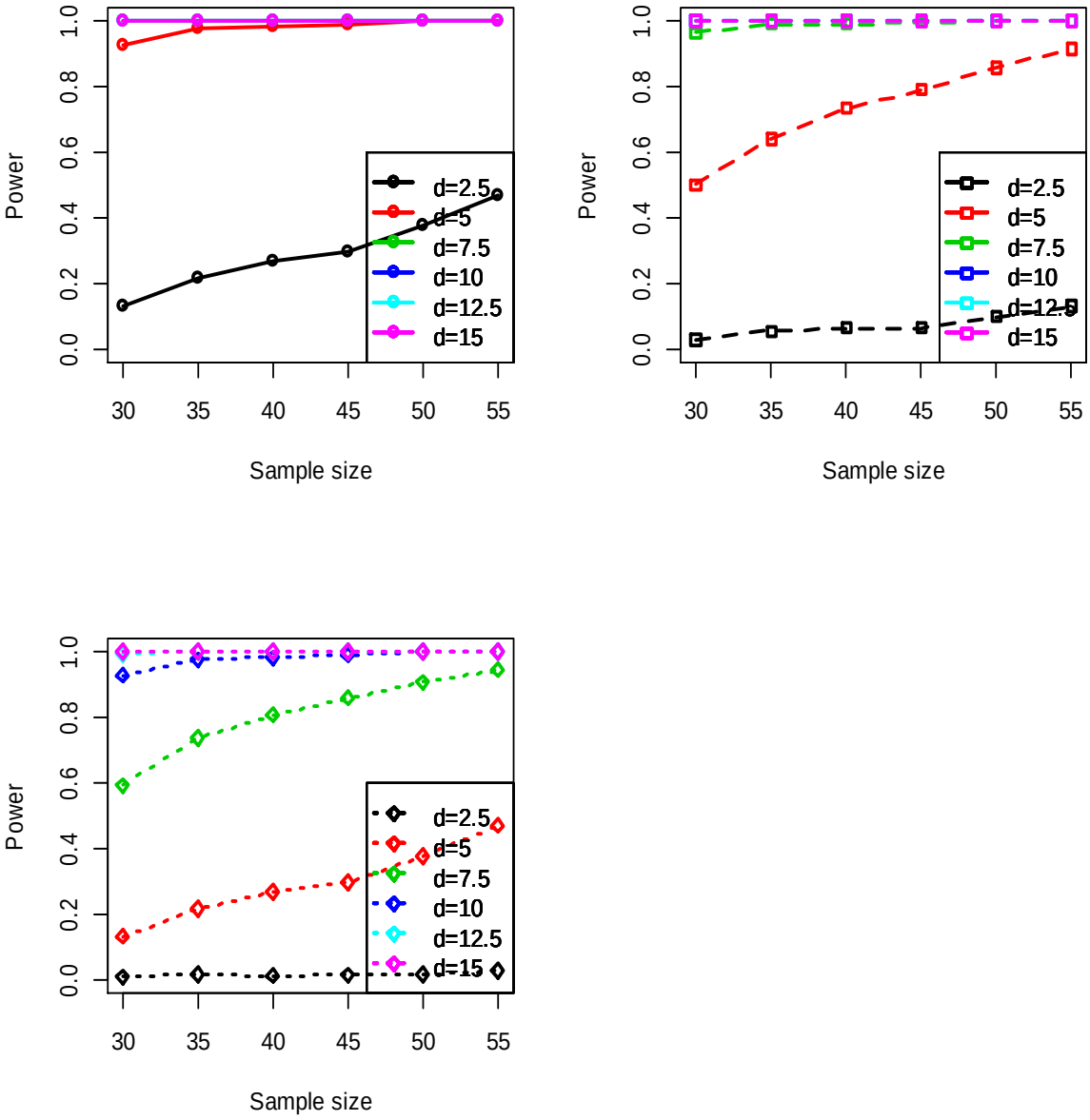

Figure 10: Power by sample size for Product= 4 and Exposure= 3 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

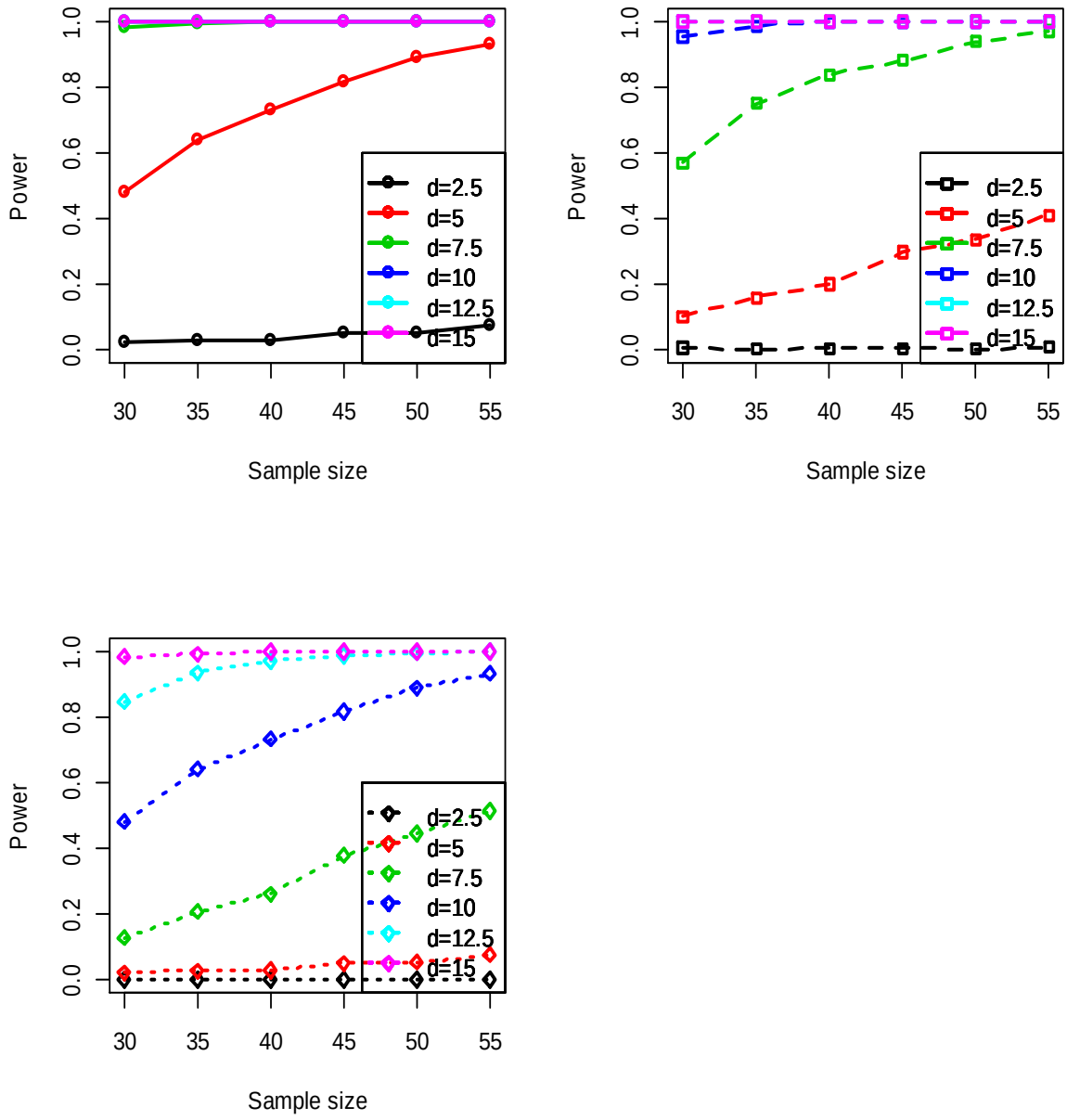

Figure 11: Power by sample size for Product= 6 and Exposure= 2 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

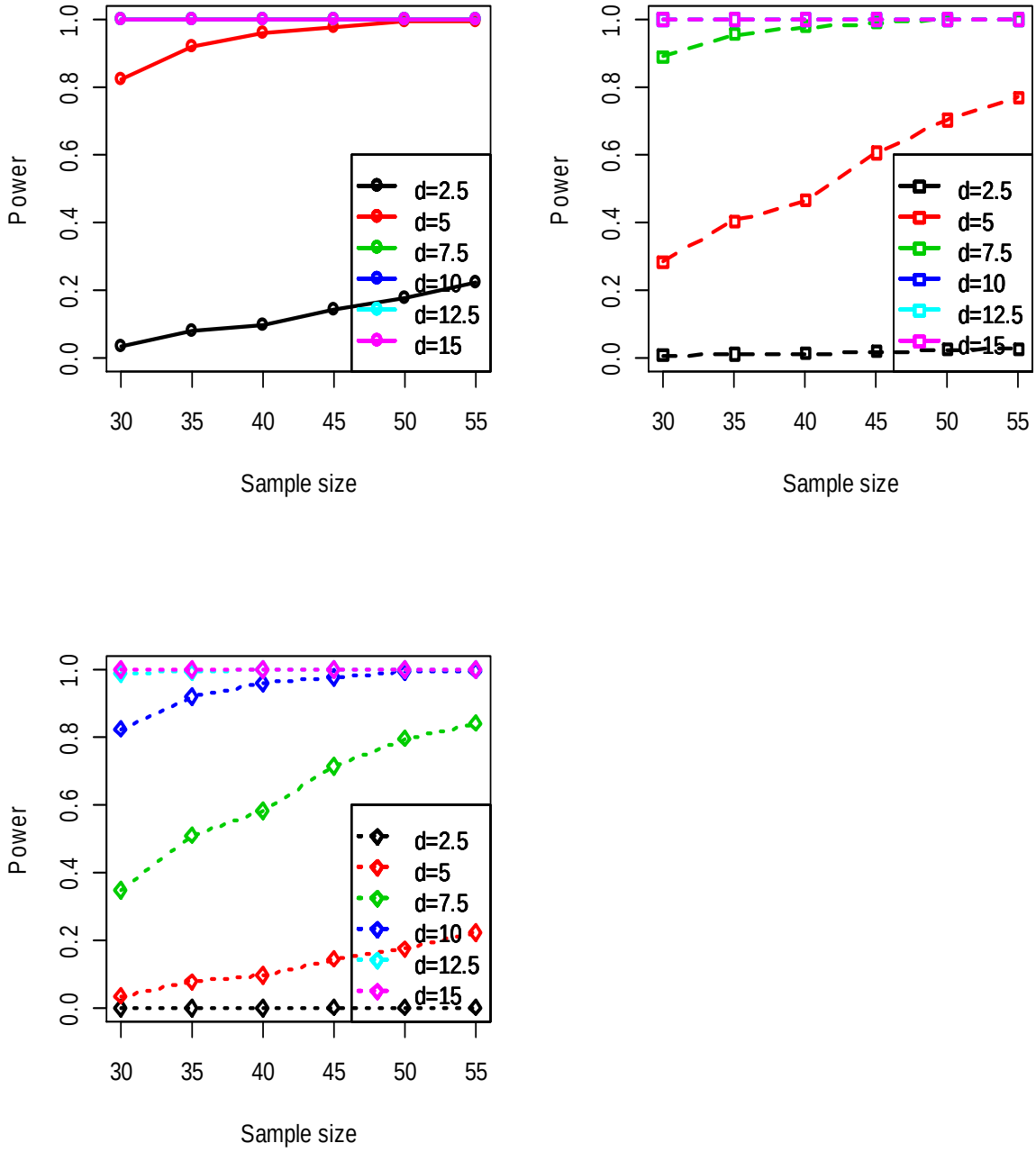

Figure 12: Power by sample size for Product= 6 and Exposure= 3 with d= mean differences (Top left= halved, Top right= estimated variances and Bottom = doubled)

Appendix B: R Simulation code

```
####partial code for the simulation study####

require(multcomp)
library(lme4)
generate_data = function(
  n # number of units
  , asim # number of simulation
  , t # number of exposure
  , d # mean difference
  , mu # population intercept
  , Sigma_s # sd of subject effect
  , Sigma_e # sd of residual
  , r  # number of product
){
  ### creating the effect size for two products#####
  p = rep(c("A","B"),t*n)
  f=c()
  for (i in 1:length( p)){

    if(p[i] =="B") {
      f[i]<-  d
    }else {
      f[i]<-  0
    }
  }
### creating the effect size for four products#####
p =rep(c("A","B","C","D"),t*n)
  f=c()
  for (i in 1:length(p)){
    if(p[i] =="D") {
      f[i]<-  d
    }else {
      f[i]<-  0
    }
  }

### creating the effect size for six products#####
p =rep(c("A","B","c","D","E","F"),t*n)
  f=c()
  for (i in 1:length(p)){
    if(p[i] =="F") {
      f[i]<- d
    }else {
      f[i]<-  0
    }
  }
  ### creating store that save p values for two products##
 Pv=rep(NA,asim)
```

```
### creating store that save p values for more than two products##
pv=rep(c(NA),asim)
  #for loop for creating data sets for all products####
  for(i in 1:asim)
  {
    set.seed(i)
    rs <- rnorm(n, 0, Sigma_s)
    rr <- rnorm(r*t*n, 0, Sigma_e)
# data generation format for two products#
    dta <- within(expand.grid(Product=c('A','B'), Subject= 1:n , Exposure = 1:t), Eat <- mu + f
+ rs[Subject] + rr)
# data generation format for four products#
dta <- within(expand.grid(Product=c('A','B','C','D'), Subject= 1:n , Exposure = 1:t), Eat <- mu
+ f  + rs[Subject] + rr)
# data generation format for six products#
dta <- within(expand.grid(Product=c('A','B','C','D','E','F'), Subject= 1:n , Exposure = 1:t), Eat
<- mu + f  + rs[Subject] + rr)

# mixed model#
    newdata.model= lmer(Eat ~ Product + (1|Subject), data=newdata)
# multiple comparisons#
    pva <-summary(glht(newdata.model,linfct=mcp(Product="Tukey")))
# extracting p values of the tests for two products##
    Pv[i] <- pva$test$pvalues[1]
#  extracting p values of the tests for four products##
pv[i]= list(c(pva$test$pvalues[3] ,  pva$test$pvalues[5] , pva$test$pvalues[6]))
# extracting p values of the tests for six products##
pv[i]= list(c(pva$test$pvalues[5] ,  pva$test$pvalues[9] , pva$test$pvalues[12],
pva$test$pvalues[14] , pva$test$pvalues[15]))
  }

  return( Pv)
}
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Sample size calculations for tasting trials

Richting: **Master of Statistics-Biostatistics**
Jaar: **2013**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Getinet Yirtaw, Tewodros

Datum: **12/09/2013**