Journal of Data Science **7**(2009), 129-138

A Replicated Experiment Used in Manufacturing

Roger L. Goodwin
US Government Printing Office

Abstract: Controlled experiments give researchers a statistical tool for determining the yield from subjecting an experimental unit to various treatments. We will discuss a replicated, block design applied to the experimental unit yeast. We subjected the yeast to six treatments. The purpose of the experiment is to extract a compound to be used in the manufacturing industry. We considered an ANOVA and a MANOVA model to analyze the data. The rationale for selecting one model over the other will be discussed. Results and recommendations of which treatments to use when processing the yeast will be presented, also.

Key words: ANOVA, controlled experiment, correlation analysis, MANOVA, manufacturing, residual analysis, sphericity test.

1. Introduction

Beta Glucan has been researched since the 1960's. It is extracted from baker's yeast cell walls. Historically 1,3-beta-glucan has been used to treat cancer by activating white blood cells. By activating the white blood cells, it enhances the body's immune system. Additionally, 1,3-beta-glucan (found in oat brand) has been linked to lowering cholesterol.[1] [2]

In this experiment, a known weight of powered yeast is treated with two factors: a disruption method(Factor A) and a digestion method(Factor B). The disruption method has three levels: A1) liquid nitrogen, A2) ground mechanically, and A3) un-ground (control). The digestion method has two levels: B1) water(H_2O), and B2) sodium hydroxide (NaOH). The purpose of treating the yeast is to break-up the cell walls of the yeast and to extract a compound called 1,3-beta-glucan to be used in manufacturing makeup. Given this, the ideal method or combination of methods should produce the highest yield of 1,3-beta-glucan. For brevity, selected analyzes will be presented through-out the paper.

[1] Healthnotes Inc. (2004), http:// www.vitacost .com/science /hn/Supp/ Beta_Glucan. htm

[2] The Cancer Cure Foundation (1976), http:// www.cancure. org/ beta_glucan.htm.

Table 1: One treatment block

	Replicates		
	1	2	3
5 Minutes	xxx	xxx	xxx
10 Minutes	xxx	xxx	xxx
...	...	...	...
90 Minutes	xxx	xxx	xxx

Yeast was exposed to each level of the two factors in the following manner: A1B1, A1B2, A2B1, A2B2, A3B1, A3B2. Thus, there are six treatments in this study, the yeast being the experimental unit. Yeast exposure to the treatments was replicated 3 times each. Measurements of the solution(in terms of area under the absorbance curve cm^2) occured at 5 minute intervals starting with 5 minutes and ending at 90 minutes. Thus, there are 18 measurements taken of each treatment per replication. This gives a total of $18 \times 3 = 54$ measurements for each treatment and a total of $54 \times 6 = 324$ measurements in the experiment. The questions to be answered are as follow:

1. Which disruption method(Factor A) is the most effective?
2. Which digestion method(Factor B) is the most effective?
3. Which of the factors or combination of factors produced the fastest rate of extraction?
4. What is the optimal extraction time?
5. Are either of the two extraction methods more effective than the control?

Questions 1, 2, and 5 can definitely be answered with a statistical model. Consider time as another factor in the experiment, and Question 4 can be answered. However, Question 3 suggests finding the reaction rate of each treatment which would fall in the realm of Chemistry not Statistics. Moreover, the treatment with the fastest reaction rate does not necessarily imply the highest yield will be extracted. Question 3 will be omitted from the analysis.

2. Model Selection

Measurements were consistently taken from the same solution in time increments. This may lead to a repeated measures model depending on the correlation coefficient. Modeling the 18 time increments as another factor(Factor C) gives the following choice of models with possible interaction:

1. In the case that the correlation coefficient is *insignificant,* the data can be represented in a 3-way crossed design(ANOVA):

$$\left.\begin{aligned} y_{ijkm} = \ & \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + \\ & (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkm}, \\ & i = 1,2,3; j = 1,2; k = 1,2,...,18; m = 1,2,3. \end{aligned}\right\} \quad (2.1)$$

   μ is the overall mean. α_i is the effect of the i-th level of Factor A. β_j is the effect of the j-th level of Factor B. γ_k is the effect of the k-th level of Factor C. $(\alpha\beta)_{ij}$ is the interaction of the i-th level of Factor A with the j-th level of Factor B. $(\alpha\gamma)_{ik}$ is the interaction of the i-th level of Factor A with the k-th level of Factor C. $(\beta\gamma)_{jk}$ is the interaction of the j-th level of Factor B with the k-th level of Factor C. $(\alpha\beta\gamma)_{ijk}$ is the interaction of the i-th level of Factor A, the j-th level of Factor B and the k-th level of Factor C. ϵ_{ijkm} is random error and $\epsilon \sim N(0, \sigma^2)$.

2. In the case that the correlation coefficient is *significant,* the data can be represented in a multivariate design(MANOVA) as:

$$\left.\begin{aligned} y_{ijk} = \ & \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \\ & i = 1,2,3; j = 1,2; k = 1,...,18; \end{aligned}\right\} \quad (2.2)$$

   μ is the overall mean. α_i is mean effect of the i-th level of Factor A. β_j is the mean effect of the j-th level of Factor B. $(\alpha\beta)_{ij}$ is the mean effect of interaction between the i-th level of Factor A and the j-th level of Factor B. ϵ_{ijk} is random error and $\epsilon \sim N(0, \sigma^2 I)$.

The term *correlation coefficient* used in the context above is used to specify the correlation between two random variables. The random variables are the 18 measurements of 1,3 beta glucan for a single replication. Thus we will measure the correlation of the first measurement with the second measurement; the first measurement with the third measurement; upto the first measurement with the eighteenth. Then, repeat for the second measurement with the third measurement; and so on. This is done for each replicate. Obviously, the correlation of a measurement to itself is always equal to one. This is why the correlation matrix has one's along the diagonal.

Since the random error terms in either model are $N(0, \sigma^2)$ in the ANOVA model and $N(0, \sigma^2 I)$ in the MANOVA model, this also implies that the error terms (also called *residuals*) are uncorrelated. Most authors do not explicitly state this assumption. It is an important assumption to verify in this experiment given that measurements were taken from the same chemical solution in time increments. Software packages ensure that the residuals ϵ_{ijkm} sum to zero. So, both of the expected values (means) of the residuals, $E(\epsilon_{ijkm}) = 0$ and $E(\epsilon_{ijk} I) =$

0 are equal to zero in the normality assumption. This is not a problem. If the variance of each residual $Var(\epsilon_{ijkm})$ and $Var(\epsilon_{ijk}I)$ is not approximately equal to σ^2, then we have several venues to investigate. Is it because of abnormalities in the data or the data collection procedures? Is it because the correlation has some structure?

Once the correlation coefficient has been quantified, the proper model can be selected and an analysis can be performed on the data. Note that if any of the interaction terms are found to be statistically significant, in this case $(\alpha\beta)_{ij}$ or $(\alpha\gamma)_{ik}$, in either of the models, then we can not discuss the factors independently. This is because the response of one factor affects the mean response of the other factor. Indeed, in this paper we will see interaction as a significant variable after the model has been selected.

2.1 Correlation analysis

Table 2: This table shows the correlation analysis. The variables $Y1$ through $Y6$ are the first six dependent random variables measuring 1,3 beta glucan taken in 15 minute time intervals. The analysis shows that these variables are highly correlated. $H_0 : \sigma^2 I$ versus $H_1 : \Lambda \Rightarrow$ reject H_0.

DF = 12	Y1	Y2	Y3	Y4	Y5	Y6
Y1	1.000000 0.0001	0.805055 0.0009	0.684215 0.0099	0.797613 0.0011	0.739900 0.0038	0.779293 0.0017
Y2	0.805055 0.0009	1.000000 0.0001	0.910334 0.0001	0.959974 0.0001	0.967580 0.0001	0.951462 0.0001
Y3	0.684215 0.0099	0.910334 0.0001	1.000000 0.0001	0.924955 0.0001	0.945084 0.0001	0.932897 0.0001
Y4	0.797613 0.0011	0.959974 0.0001	0.924955 0.0001	1.000000 0.0001	0.974034 0.0001	0.931556 0.0001
Y5	0.739900 0.0038	0.967580 0.0001	0.945084 0.0001	0.974034 0.0001	1.000000 0.0001	0.952571 0.0001
Y6	0.779293 0.0017	0.951462 0.0001	0.932897 0.0001	0.931556 0.0001	0.952571 0.0001	1.000000 0.0001

Test for Sphericity: Mauchly's Criterion = 0.0066855
Chisquare Approximation = 50.57899 with 14 df Prob > Chisquare = 0.0000

Each treatment is an 18×3 matrix (time has been included). There are two factors A and B. But, these factors have different levels: Factor A : $A1$, $A2$, and $A3$. Factor B : $B1$, and $B2$. Cross the two factors to get a grand total of $18 \times 3 \times 2 \times 3 = 324$ observations. For now, we are dealing with matrices and vectors. The dependent variable y_1 is an 18×1 vector. The dependent variable y_2 is an 18×1 vector and so on upto y_{18} is an 18×1 vector. It just so happens that 18^2 equals to 324 also. So, this 18×18 matrix is on the left hand side of the model statement. This is how it is implemented in SAS. The variance among this matrix is what we are trying to explain statistically. It's not until a univariate analysis is performed that these vectors need to be copied to a 324×1 vector, in which case, a variable for TIME needs to be created.

Eighteen dependent SAS variables, **Y1, Y2, ...,Y18,** were created to represent each measurement at time $i, i = 1, 2, ..., 18$. Two independent SAS variables, **DISRUPT** and **DIGEST,** were created to represent Factor A and Factor B. A repeated measures analysis was performed to determine if correlation existed among the 18 dependent variables and to determine if the correlation was constant. Looking at a subset of the partial correlation matrix in Table 2, it is obvious that the first 6 random variables are correlated (partial correlation was calculated for all 18 variables and all the partial correlations were high). See Table 2. Why only 6 variables appear on the printout will become apparent in the next paragraph. For now, it can be concluded that a model similar to equation (2.2) should be used.

The next problem is to determine if the correlation structure is constant. The sphericity test can be used to test for equal correlation among the 18 random variables. In a multivariate model, the hypotheses are $H_0 : \sigma^2 I$, versus H_1 : correlation is not equal(Λ.) Specifically with the given data, there were not enough degrees of freedom to run the sphericity test. However, if H_0 is rejected on some subset of the 18 random variables, then it can be concluded that H_1 is true. This approach gives enough degrees of freedom for the sphericity test. Since the p-value of the sphericity test in Table 2 is 0.0000, reject H_0 using the subset **Y1, Y2, Y3, Y4, Y5, Y6.** From this, it can be concluded that a different approach must be taken to analyze the data. PROC MIXED in SAS will be used.

3. Data Analysis

PROC MIXED gives many ways of choosing the structure for ϵ when $\epsilon \sim N(0, \Lambda)$. Among those structures are the first-order autoregressive AR(1), the ARMA(1,1), compound symmetry, factor analytic, banded, etc. In all, there are twenty three named covariance structures in SAS. Given all these choices, we tried the first-order autoregressive AR(1) covariance structure. Autoregressive models are used often in repeated measures experiments. Since we did continually draw measurements from the same chemical solution in time increments, this experiment lends itself to the AR(1) model. To analyze the data using PROC MIXED, a new data set was created. The dependent variable **Y** was created and the dependent variables **Y1–Y18** were dropped from the data set. **Y** is a 324×1 column vector containing all the information in **Y1–Y18.** AR(1) was used to model the correlation structure. The parameter estimate of the correlation is $\widehat{\rho} = 0.8922$ and the parameter estimate of σ^2 is $\widehat{\sigma}^2 = 1362.33$. We will need both of these estimates in Section 3.1 to uncorrelate the residuals. PROC MIXED did give an estimate of ρ which PROC GLM did not do. The hypotheses tests using PROC GLM match those in Table 3 using PROC MIXED: the digestion methods and time levels are both significant, and the disruption methods are insignificant. The next problem is to determine which digestion level and which time level are most significant using PROC MIXED.

Table 3: Estimates for the correlation and sigma squared from PROC MIXED

Covariance Parameter Estimates (REML)					
Cov Parm	Ratio	Estimate	Std Error	Z	Pr > $\|Z\|$
DIAG AR(1)	0.000655	0.89221	0.0257	34.66	0.0001
Residual	1.000000	1362.3261	318.0393	4.28	0.0001

The disruption methods (Factor A) had an overall p-value of 0.0677. See Table 4. None of the different levels of disruption are statistically significant at the 95% confidence level. So, the recommended disruption method is unground yeast.

The digestion methods (Factor B) had an overall p-value of 0.0001. See Table 4. Digestion is a significant factor in the experiment. Using the Tukey pairwise comparison test, there is a statistically significant difference between H_2O and NaOH. Since NaOH has the higher mean of 255.305 compared to H_2O of 118.765, use NaOH in the manufacturing process. The Tukey comparison tests are quite lengthy, and have been omitted for brevity.

The time levels (Factor C) had an overall p-value of 0.0002. See Table 4. Time is a significant factor in the experiment. Time level 16 resulted in the highest yield. Using the Tukey pairwise comparison test, time level 16 is statistically different from levels 1 thru 7. At time level 8, the yields become insignificant when compared with level 16. Thus, use time level 8 (40 minutes) in the manufacturing process.

3.1 Residual analysis

The purpose of performing a residual analysis is the verify the normality assumption of the model. Since the residuals were correlated, they had to be transformed to make them uncorrelated. The transformation involved creating the 18×18 matrix $\widehat{V}$ such that $\widehat{V} = \widehat{\sigma}^2 \widehat{\rho}^{|i-j|}$, where $i = 1, 2, ..., 18$; $j = 1, 2, ..., 18$. Then, the following transformation matrix is derived:

$$\widehat{\Lambda}^{-1/2} = \left(I \otimes \widehat{V} \right)^{-1/2}.$$

$\otimes$ is the direct product of the 18×18 identity matrix I and $\widehat{V}$ and produces a 324×324 matrix. The transformed residuals are obtained by multiplying $\widehat{\Lambda}^{-1/2}$ by the residuals from PROC MIXED in Section 3. The transformation matrix was created in PROC IML.

The hypotheses tests are: H_0 : uncorrelated residuals are normally distributed, vs H_1 : uncorrelated residuals are not normally distributed. When examining the PROC UNIVARIATE output of the residuals, the p-value of the Wilkins test for normality is 0.0001. Thus, we reject H_0. The residuals do not come from a normal population. Many of the extreme residuals came from the A2B2 treatment(ground yeast and NaOH).

Table 4: Tests of hypotheses of the factors in the experiment using PROC MIXED

		Tests of Fixed Effects				
Source	Num DF	Den DF	Chi Sq	F	Pr > Chi Sq	Pr > F
DIGEST	1	12	109.56	109.56	0.0001	0.0001
DISRUPT	2	12	6.80	3.40	0.0334	0.0677
DIGEST*DISRUPT	2	12	5.35	2.67	0.0690	0.1095
TIME	17	204	48.67	2.86	0.0001	0.0002
DIGEST*TIME	17	204	26.86	1.58	0.0601	0.0718
DISRUPT*TIME	34	204	33.03	0.97	0.5148	0.5187
DIGEST*DISRUPT*TIME	34	204	34.86	1.03	0.4270	0.4374

Table 5: Tests of hypotheses of the factors in the experiment after removing the A2B2 block. The significance levels of the interaction terms have changed from the previous hypotheses testing.

		Tests of Fixed Effects				
Source	Num DF	Den DF	Chi Sq	F	Pr > Chi Sq	Pr > F
DIGEST	1	11	487.63	487.63	0.0001	0.0001
DISRUPT	2	11	54.25	27.13	0.0001	0.0001
DIGEST*DISRUPT	2	11	49.15	24.57	0.0001	0.0001
TIME	17	187	92.36	5.43	0.0001	0.0001
DIGEST*TIME	17	187	42.96	2.53	0.0005	0.0012
DISRUPT*TIME	34	187	38.88	1.14	0.2593	0.2824
DIGEST*DISRUPT*TIME	34	187	39.52	1.16	0.2368	0.2611

4. Normality Remedy

Since many of the extreme residuals came from the A2B2 block, one or more of the replications from that block should removed. By trial and error, it was decided to remove the third replication (later, the biologist admitted that she had problems with the equipment). This still leaves two replications to estimate the mean response of the A2B2 treatment. Upon removing the third replication, the correlated residuals became normally distributed. However, the interaction[3] between Factor A(disruption) and Factor B(digestion) is now significant. The interaction between Factor B(digestion) and Factor C(time) is now significant. See Table 5.

Granted that the entire A2B2 block was removed. This can be justified due to equipment failure. If one or more of the observations in that block are unreliable, then most likely, the validity of the other observations in that replicate are questionable also. So, instead of picking through each individual observation in a replicate to obtain the normality assumption, it was decided to remove the entire replicate. This still left two replicates for estimation purposes.

5. Re-Analysis of the Data

Currently, the model equation is similar to that of equation 2. The mean interaction responses are being estimated by :

$$\widehat{(\alpha\beta)}_{ij} = \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}.$$

The mean interaction responses should be estimated by:

$$\bar{y}_{ij\cdot} = \frac{1}{r}\sum_{k=1}^{r} y_{ijk}.$$

which corresponds to the following model $y_{ijk} = \mu_{ij} + \epsilon_{ijk}$. PROC MIXED was run again with just the interaction terms to obtain the estimates.

A Tukey pairwise comparison test was run on the AB interaction. The highest mean occurred at the ground yeast and NaOH combination. The mean is 332.56, and the mean is statistically different from the other various combinations of each level of Factor A and Factor B. Thus, ground yeast in sodium hydroxide should be used in the manufacturing process.

A Tukey pairwise comparison test was run on the BC interaction. The highest mean being 306.28 occurred at the NaOH level of Factor B and level 13 of Factor C. Holding NaOH constant, level 13 of Factor C is not statistically different from

[3] As in the statistical sense: the level of one factor affects the mean response of another factor.

levels 9 thru 12 but is statistically different from levels 1 thru 8. The combination NaOH and level 9 of Factor C is statistically different from the H_2O level of Factor B and level 17 of Factor C. The yeast should be exposed to NaOH for 45 minutes.

6. Manufacturing Recommendations

1. Which disruption method(Factor A) is the most effective? Holding the digestion levels constant, there is a statistical significance and difference among the levels of the disruption methods at the 95% confidence level. Recommendation: ground yeast.

2. Which digestion method(Factor B) is the most effective? Holding the disruption levels constant, there is a statistical significance and difference between H_2O and NaOH at the 95% confidence level. Recommendation: NaOH(sodium hydroxide).

3. Which of the factors or combination of factors produced the fastest rate of extraction? Omitted.

4. What is the optimal extraction time? Holding the digestion levels constant, there is a statistical significance and difference between the various levels of time. Recommendation: 45 minutes.

5. Are either of the two extraction methods more effective than the control? Holding NaOH constant, ground yeast results in a statistically significant higher mean response than either un-ground yeast or yeast treated with liquid nitrogen. Holding H_2O constant, there is no statistical difference in the various levels of the extraction methods.

7. Summary

We chose the MANOVA model over the ANOVA because the correlation coefficient was statistically significant. We modeled the data, but rejected the assumption that the residuals (uncorrelated) have a normal distribution. After studying the extreme residuals (outliers), we decided to remove one of the replications in the experiment. This still left two replications for hypothesis testing, and estimation purposes in that block. We re-analyzed the data, and recommended how to manufacture the yeast for extracting the compound 1,3-beta-glucan.

Acknowledgments

This project originated at the Old Dominion University Center for Biotechnology. The author thanks Dr. Naik (Old Dominion University), and Ann Wilson (Virginia Tissue Bank) for their help.

References

Khattree, Ravindra, and Naik, Dayanand N. (1995), *Applied Multivariate Statistics with SAS Software.* SAS Institute Inc., Cary, NC.

Petersen, Roger G. (1985), *Design and Analysis of Experiments.* Marcel Dekker.

Received February 23, 2007; accepted August 16, 2007.

Roger L. Goodwin
US Government Printing Office
732 North Capitol Street NW
Washington DC 20401
roger_goodwin@dc-sug.org or rgoodwin@gpo.gov