An Experiential Approach to Integrating ANOVA Concepts

LeRoy A. Franklin
Rose-Hulman Institute of Technology

Belva J. Cooley
University of Montana

Journal of Statistics Education Volume 10, Number 1 (2002)

Copyright © 2002 by LeRoy A. Franklin and Belva J. Cooley, all rights reserved.
This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Cell means; Interaction; Parameter estimation; Power; Replication; Sample size.

Abstract

This paper presents a data set based on an industrial case study using design of experiments. The data set is pedagogically rich because it has a rather large total sample size from an industrial setting that naturally yields a large third order interaction term. The experiment is a $2^3$ design and is initially presented with no replications. The sample size of the data is then doubled and the analysis repeated, comparing these results with previous results. The process is repeated until eight replications are available for each combination of factors and all parameters are estimated. With eight replications, the analysis shows all main effects and all interactions are statistically significant at the $\alpha = 0.05$ level. With smaller sample sizes, various main effects and interactions are not found to be statistically significant. Through this presentation the instructor can lead students in discussions about the effect of increased sample sizes, power, statistical significance (or insignificance), interaction terms, Type I and Type II errors as well as the importance and the role of the error term. In addition, students can manipulate the data set in a computer laboratory setting to illustrate many of the concepts inherent in the design of experiments and analysis of variance.

1. Introduction

Concepts such as power, Type I and Type II errors, model assumptions, statistical significance, statistical non-significance, estimates of the model variance, and the impact these concepts have on interpretation of the data are crucial features for analysis of variance and design of experiments. All too frequently, these concepts are presented in a disjointed fashion without showing their interrelationships.

The authors have found that an "in class discussion" of a data set (based on an industrial case study)
which is repetitively presented with progressively larger samples has been a very helpful tool to integrate and synthesize many of these concepts. This discussion occurs after the fundamental concepts, calculations and a few simple examples have been presented to the class. The case is used in an intermediate course in applied statistics with an audience of senior business majors, MBA students and various graduate-level science majors. When it has been used, the reception of the class has been enthusiastic and, although student feedback has been informal, several students have encouraged the authors to continue using such an exercise.

2. The Industrial Case Setting

A company engaged in the manufacture of large industrial fans recently acquired a smaller competitor that was also engaged in manufacturing fans. However the two companies had very different manufacturing methods and each company championed its method as superior to the other. The merging of the two companies presented an important opportunity to choose "best practices," standardize operations, and reduce inventories. Since many factors were potentially involved and the decision to standardize on a particular process had substantial long run manufacturing and financial ramifications, the authors were contacted to assist in the design and analysis of the experiments.

After much discussion, the company engineers agreed to look at three important factors in the manufacture of the fans: the type of hole in the fan "spyder" (that is, blades), the type of barrel onto which the spyder was placed, and the type of assembly method for these two components. Each of these factors was qualitative in nature and at two levels. Thus the experiments became a straightforward $2^3$ design that is presented in most design of experiments texts (Cochran and Cox 1957; Box, Hunter, and Hunter 1978; Hicks 1993). Because of industrial and competitive considerations, company representatives have requested that the precise nature of these factor levels not be presented. The two levels of each factor are simply coded -1 and 1. However, Figure 1 "suggests" the type and nature of the factors.

The single most important metric in measuring the quality of a manufactured fan was the torque (in foot-pounds) at which the spyder could be broken from the barrel. When the fan blade broke free of the hub, the shaft of the turning motor would rotate but the blades would not. Such a failure had serious ramifications for the large industrial units being manufactured.
Figure 1. Factors in the manufacture of the fans.
3. The Discussion

The key technique used in this discussion is the use of one data set that is revisited a total of four times (about all that a single class period will allow). Each time the experiment is revisited the sample size is doubled so that the same experiment is done with 8, then 16, then 32, and finally 64 data points. A powerful approach to teaching this material is to allow students to work through the analysis in a computer laboratory, using prepared diskettes containing the datasets. Students can also get the data from the Internet. If the hands-on approach is not practical, the authors recommend that tables containing the data and the analysis be used as a class handout to focus each student’s attention on interpretation and synthesis of concepts and to de-emphasize the hand calculations necessary to achieve those results.

The first experiment is comprised of one observation for each combination of hole type, assembly method, and barrel design. Table 1 shows the MINITAB® results for ANOVA and the parameter estimates for a sample size of eight. Using the printed cell sample means and the standard set of equations, the class can compute the estimates of each of the model terms that are shown in the table. The three-factor interaction term is used to provide an estimate of the error term with one degree of freedom, a course of action sometimes recommended by statisticians (Montgomery 2001). Using the calculated values for F or the p-values, students will find that only the surface of the barrel is statistically significant at the 0.05 significance level. The instructor might note that when MINITAB® reports p-values of 0.000 they are not truly "zero" p-values but rather statements that the actual p-values are less than 0.0005.

The instructor can lead discussion concerning the use of the three-way interaction term as a surrogate for the error term. The analyst is, in fact, assuming that the three-way interaction term is not significant. If the assumption is incorrect and the interaction term is large, main effects of substantial size may be "hidden." Since the fans were destroyed in the testing process, the engineers may have been tempted to test a small number of fans. Had they stopped at eight and used the analysis in Table 1, they may have concluded that only the shape of the barrel had a significant effect on the strength of the fan. They could have concluded that neither the shape of the hole in the spyder nor the assembly method was important. Therefore, inventory and equipment using both types of hole and both types of assembly method could have been routinely used to produce fans. The analysis shows that the average breaking torque of the fans using the best barrel shape was 176.25 foot-pounds.

Table 1. Experiment 1 (N = 8)

| Source            | DF | SS      | MS   | F     | p    | Power |
|-------------------|----|---------|------|-------|------|-------|
| Hole (A)          | 1  | 1035.1  | 1035.1| 18.78 | 0.144| 0.266 |
| Assembly (B)      | 1  | 1485.1  | 1485.1| 26.94 | 0.121| 0.316 |
| Barrel (C)        | 1  | 26565.1 | 26565.1| 481.91| 0.029| 0.940 |
| Hole*Assembly (A*B)| 1  | 325.1   | 325.1| 5.90  | 0.249| 0.151 |
| Hole*Barrel (A*C) | 1  | 55.1    | 55.1 | 1.00  | 0.500| 0.073 |
| Assembly*Barrel (B*C)| 1  | 45.1    | 45.1 | 0.82  | 0.532| 0.069 |
| Error             | 1  | 55.1    | 55.1 |       |      |       |
This is a good time for the instructor to review Type I and Type II errors. The hypothesis being evaluated is that the treatment effects are equal to zero. If a conclusion is made that an effect is significant when it is not, the analyst has committed a Type I error. The probability of a Type I error, $D$, can be controlled by setting a value of $D = 0.05$, for instance, in conducting the hypothesis test. The p-values, in the table above, give the probability of obtaining more extreme test results by chance if the null hypothesis was true. A Type II error is committed when the analyst concludes that the treatment effects are zero when an effect actually does exist. Students should understand that the probability of committing a Type II error is $E$, while the probability of identifying a significant effect when there actually is one is $1 - E$ (the power of the test). For a given level of $\alpha$, the power to detect real differences depends on the size of the sample taken and the magnitude of the difference that must be detected. At this point the instructor should emphasize that factors and terms displaying p-values larger than 0.05 and...
judged not statistically significant could actually have a significant effect in producing stronger fans. The sample size may simply not be large enough to detect their actual (non-zero) effect.

The last column in the ANOVA portion of Table 1 gives an approximation of the observed power for the hypothesis test, assuming \( \alpha = 0.05 \). (SPSS® provides these estimates as optional output.) With the exception of the variable for barrel shape, all power estimates are quite low. For instance, if the observed effect size for the first variable, hole, is a true difference, the null hypothesis would be rejected only 26.6% of the time in repeated experiments of this design. If the power is small, the analyst cannot be confident in a decision not to reject the null hypothesis.

The instructor should next challenge the students to predict what might happen if the sample size is doubled to 16 observations, two at each of the factor combinations. Table 2 gives the results of the MINITAB® analysis for a sample size of 16. Students should note that replicating the experiment allows an estimate of an error term with eight degrees of freedom. The mean squared error is now smaller than in the previous experiment. Doubling the sample size results in an ANOVA showing all main effects to be significant. In addition, the Hole*Assembly interaction is significant at the 0.05 level and so is the three-factor interaction.

### Table 2. Experiment 2 (N = 16)

| Analysis of Variance | Observed Source                  | DF | SS   | MS   | F     | p     | Power |
|----------------------|----------------------------------|----|------|------|-------|-------|-------|
|                      |                                  |    |      |      |       |       |       |
| Hole (A)             |                                  | 1  | 2025.0 | 2025.0 | 58.91 | 0.000 | 1.000 |
| Assembly (B)         |                                  | 1  | 3080.3 | 3080.3 | 89.61 | 0.000 | 1.000 |
| Barrel (C)           |                                  | 1  | 47089.0 | 47089.0 |1369.86 | 0.000 | 1.000 |
| Hole*Assembly (A*B)  |                                  | 1  | 870.2  | 870.2  | 25.32 | 0.001 | 0.992 |
| Hole*Barrel (A*C)    |                                  | 1  | 144.0  | 144.0  | 4.19  | 0.075 | 0.437 |
| Assembly*Barrel (B*C)|                                  | 1  | 56.2   | 56.2   | 1.64  | 0.237 | 0.204 |
| Hole*Assembly*Barrel (A*B*C) |                 | 8  | 275.0  | 34.4   |       |       |       |
| Total                |                                  | 15 | 53846.0 |      |       |       |       |

| Hole N               |         | Means   |
|----------------------|---------|---------|
| -1                   | 8       | 106.25  |
| 1                    | 8       | 128.75  |

| Assembly N           |         | Means   |
|----------------------|---------|---------|
| -1                   | 8       | 131.38  |
| 1                    | 8       | 103.62  |

| Barrel N             |         | Means   |
|----------------------|---------|---------|
| -1                   | 8       | 63.25   |
| 1                    | 8       | 171.75  |

| Hole Assembly N      |         | Means   |
|----------------------|---------|---------|
| -1 -1                | 4       | 112.75  |
| -1 1                 | 4       | 99.75   |
| 1 -1                 | 4       | 150.00  |
| 1 1                   | 4       | 107.50  |
It is helpful to recalculate the estimates of the model parameters from the larger sample and note that these values are close to but not identical to those calculated from the original sample of eight. Prior to experimentation, the company engineers debated how much of a difference in treatment effects would be of material or practical difference. For instance, a person’s fingers can easily exert +/- 1.5 foot-pounds of torque in placing a nut on a bolt. The engineers finally decided that they wanted to detect differences of 10 foot-pounds or more. Each of the main effects in this experiment is estimated at greater than 10 foot-pounds.

The instructor should point out the new estimates for observed power in Table 2 and how they differ from those in Table 1. The Hole*Barrel and Assembly*Barrel interactions are not significant at the \( \alpha = 0.05 \) level. However, the observed estimates of power for the tests are 0.437 and 0.204, respectively. Since the power is low, the analyst should be wary of a decision not to reject the hypothesis. The instructor should point out, however, that the estimates of effects for all the interaction terms are less than the 10 foot-pounds of force that the engineers wanted to detect. Even if true differences of the magnitudes observed in the interaction effects are detected, they may not be large enough to be considered material differences in the context of the problem. Depending upon the level of class being taught, the instructor might want to point out the availability of tables to determine the appropriate
sample size for analysis of variance given specified levels of $\alpha$, $\beta$, and the minimum acceptable size of the treatment effect (Nelson 1985). Formulas for calculating power are available for those instructors wishing to include a more rigorous treatment of power (Desu and Raghavarao 1990).

Again, the instructor should ask the students to predict what will happen if the sample size is doubled to 32 (four observations at each treatment combination). Table 3 presents the results of the third experiment. The instructor should point out that the larger sample size results in more degrees of freedom for the error sum of squares, resulting in an estimate of mean square error of 30 instead of 34.4. The Hole*Barrel interaction is now significant at $\alpha = 0.05$, but the Assembly*Barrel interaction term is not. The instructor might point out that the Hole*Barrel interaction had a p-value of 0.075 in the previous experiment with $n = 16$, but is now estimated to be 0.02. The instructor should ask the students to explain what has happened.

**Table 3. Experiment 3 (N = 32)**

| Source                     | DF | SS     | MS     | F       | p     | Power |
|----------------------------|----|--------|--------|---------|-------|-------|
| Hole(A)                    | 1  | 4028   | 4028   | 134.48  | 0.000 | 1.000 |
| Assembly(B)                | 1  | 6699   | 6699   | 223.69  | 0.000 | 1.000 |
| Barrel(C)                  | 1  | 93853  | 93853  | 3133.87 | 0.000 | 1.000 |
| Hole*Assembly(A*B)         | 1  | 1526   | 1526   | 50.96   | 0.000 | 1.000 |
| Hole*Barrel(A*C)           | 1  | 185    | 185    | 6.19    | 0.020 | 0.665 |
| Assembly*Barrel(B*C)       | 1  | 116    | 116    | 3.88    | 0.060 | 0.473 |
| Hole*Assembly*Barrel(A*B*C)| 24 | 719    | 30     |         |       |       |
| Total                      | 31 | 107765 |        |         |       |       |

---

| Source     | N | Means  |
|------------|---|--------|
| Hole       |   |        |
| -1         | 16| 105.81 |
| 1          | 16| 128.25 |
| Assembly   |   |        |
| -1         | 16| 131.50 |
| 1          | 16| 102.56 |
| Barrel     |   |        |
| -1         | 16| 62.88  |
| 1          | 16| 171.19 |

---

| Source       | N | Means  |
|--------------|---|--------|
| Hole Assembly|   |        |
| -1 -1        | 8 | 113.37 |
| -1 1         | 8 | 98.25  |
| 1 -1         | 8 | 149.62 |
| 1 1          | 8 | 106.87 |
| Hole Barrel  |   |        |
| -1 -1        | 8 | 49.25  |
| -1 1         | 8 | 162.37 |
| 1 -1         | 8 | 76.50  |
| 1 1          | 8 | 180.00 |
Table 4 gives the ANOVA results and parameter estimates for a sample size of 64 (8 observations for each treatment combination). With a sample size of 64, all main effects and all treatment interactions are significant at $\alpha = 0.05$. This is a good time to introduce the concept of inflating the overall probability of Type I error by making several hypothesis tests. The instructor might suggest using a more conservative value for the significance level, such as 0.01, a Bonferroni-type correction. With seven tests (three main effects and four interactions), the probability of making at least one Type I error would be $1 - (0.99)^7$ or approximately 0.07. The instructor should ask the students to review the results of the four experiments to determine what their conclusions would have been using $\alpha = 0.01$.

Table 4. Experiment 4 (N = 64)

| Source           | DF | SS   | MS   | F    | p    | Power |
|------------------|----|------|------|------|------|-------|
| Hole (A)         | 1  | 8258 | 8258 | 266.68 | 0.000 | 1.000 |
| Assembly (B)     | 1  | 13369| 13369| 431.73 | 0.000 | 1.000 |
| Barrel (C)       | 1  | 193050| 193050| 6234.17 | 0.000 | 1.000 |
| Hole*Assembly (A*B) | 1  | 2849 | 2849 | 92.00 | 0.000 | 1.000 |
| Hole*Barrel (A*C) | 1  | 594 | 594 | 19.19 | 0.000 | 0.990 |
| Assembly*Barrel (B*C) | 1  | 135 | 135 | 4.36 | 0.041 | 0.537 |
Hole*Assembly*Barrel           1397    1397    45.11  0.000   1.000
(A*B*C)

Error                    56    1734      31
Total                    63  221387

| Hole | N  | Means |
|------|----|-------|
| -1   | 32 | 106.66|
| 1    | 32 | 129.38|

| Assembly | N  | Means |
|----------|----|-------|
| -1       | 32 | 132.47|
| 1        | 32 | 103.56|

| Barrel | N  | Means |
|--------|----|-------|
| -1     | 32 | 63.09 |
| 1      | 32 | 172.94|

| Hole | Assembly | N  | Means |
|------|----------|----|-------|
| -1   | -1       | 16 | 114.44|
| -1   | 1        | 16 | 98.87 |
| 1    | -1       | 16 | 150.50|
| 1    | 1        | 16 | 108.25|

| Hole | Barrel | N  | Means |
|------|--------|----|-------|
| -1   | -1     | 16 | 48.69 |
| -1   | 1      | 16 | 164.63|
| 1    | -1     | 16 | 77.50 |
| 1    | 1      | 16 | 181.25|

| Assembly | Barrel | N  | Means |
|----------|--------|----|-------|
| -1       | -1     | 16 | 79.00 |
| -1       | 1      | 16 | 185.94|
| 1        | -1     | 16 | 47.19 |
| 1        | 1      | 16 | 159.94|

| Hole | Assembly | Barrel | N  | Means |
|------|----------|--------|----|-------|
| -1   | -1       | -1    | 8 | 53.25 |
| -1   | -1       | 1     | 8 | 175.62|
| -1   | 1        | -1    | 8 | 44.13 |
| -1   | 1        | 1     | 8 | 153.62|
| 1    | -1       | -1    | 8 | 104.75|
| 1    | -1       | 1     | 8 | 196.25|
| 1    | 1        | -1    | 8 | 50.25 |
| 1    | 1        | 1     | 8 | 166.25|

Estimates of the Model Parameters

\[ \mu = 118.0 \]

\[ A_{-1} = -11.4 \quad A_1 = 11.4 \]

\[ B_{-1} = 14.5 \quad B_1 = -14.5 \]

\[ C_{-1} = -54.9 \quad C_1 = 54.9 \]

\[ AB_{-1-1} = -6.7 \quad AB_{-11} = 6.7 \quad AB_{1-1} = 6.7 \quad AB_{11} = -6.7 \]

\[ AC_{-1-1} = -3.0 \quad AC_{-11} = 3.0 \quad AC_{1-1} = 3.0 \quad AC_{11} = -3.0 \]

\[ BC_{-1-1} = 1.5 \quad BC_{-11} = -1.5 \quad BC_{1-1} = -1.5 \quad BC_{11} = 1.5 \]

\[ ABC_{-1-1-1} = -4.7 \quad ABC_{-11-1} = 4.7 \quad ABC_{1-1-1} = 4.7 \quad ABC_{11-1} = -4.7 \]

\[ ABC_{-111} = 4.7 \quad ABC_{111} = 4.7 \quad ABC_{111} = 4.7 \]
The instructor should again call the students’ attention to the estimated means for the various treatment combinations. With the original sample size of eight, only the barrel shape was significant. If the other factors are ignored and the engineers concentrate only on the best setting for barrel shape, the average breaking torque is estimated to be 176.25. The larger sample size of 64 helps identify significant differences for all factors. If the engineers choose the best settings for all factors, the average breaking torque is estimated at 196.25. Although destructive testing was necessary, the larger sample size helped identify a combination of factor settings that substantially improved the strength of the industrial fans. Figures 2 and 3 show plots of the sample means for the eight different factor combinations.

Figure 2. Breaking Torque (for Hole = -1).
Looking back to the beginning of the exercise when only eight fans were tested, the third-order interaction was used as a surrogate for error in the analysis since the experiment was unreplicated. Because the sum of squares for that interaction was large and there was only one degree of freedom, only one main effect (barrel shape) was significant. In fact, using the large interaction term as error hid the impact that the hole shape and assembly method had in producing a stronger fan. A better practice may have been to combine the interactions terms to serve as error and only screen for main effects in the model as seen in Table 5. All main effects are significant at the $\alpha = 0.05$ level, which is consistent with the findings of the $n = 64$ experiment.

Table 5. Experiment 1 Revised (N = 8)

| Source       | DF | SS   | MS   | F       | p       | Power |
|--------------|----|------|------|---------|---------|-------|
| Hole (A)     | 1  | 1035 | 1035 | 8.617   | 0.043   | 0.602 |
| Assembly (B) | 1  | 1485 | 1485 | 12.363  | 0.025   | 0.749 |
| Barrel (C)   | 1  | 26565| 26565| 221.15  | 0.000   | 1.000 |
| Error        | 4  | 481  | 31   |         |         |       |
| Total        | 7  | 29566|      |         |         |       |

As a possible homework assignment, the instructor may want each of the students to randomly choose observations from the full dataset to create new datasets of 8, 16, and 32. Requiring this will result in each student’s work being slightly different. However the authors have run several different subsetting...
patterns. Although the associated p-values do change moderately, it is rare to have a dramatic difference, and the essential results remain the same. However, the instructor may wish the students to explain the cause of any differences that do occur in their analysis as part of the homework assignment.

4. Some Other Issues for Discussion

The instructor may also choose to remind the students that assumptions should be checked when using ANOVA techniques. In this case, a Kolmogorov-Smirnov test of normality on the error terms failed to find a significant non-normality. Tests for homogeneity of variances were mixed. Bartlett's Test resulted in a p-value of 0.116, indicating no significant heterogeneity of variances. However, Levene's Test resulted in a p-value of 0.000, indicating that the variances are not the same at all factor levels. Discussions of these two tests can be found in Montgomery (2001), Box (1954) discusses the robustness of ANOVA to departures from the homogeneity of variances assumption if the sample sizes are equal at all factor levels. In this case, the smallest standard deviation of breaking torques for a factor level is 3.01, while the largest is 8.78. The standard deviations can differ by as much as a factor of three without invalidating the results of the ANOVA analysis.

The instructor could also lead a discussion about other issues, for which information is not available, that may impact the decision. Students might suggest that costs of the various methods be collected and compared. They might suggest that an engineer review the recommended design. Students might ask if breaking torque is the appropriate measure to use and, if so, what value of breaking torque do the customers require. They might explore the existence of other measures of quality that might be appropriate. They should also look at the standard deviation of each of the combinations to determine if the combination of factors that gave the highest breaking strength also resulted in a relatively small standard deviation. A complete discussion of each of these issues will lead to discussion of many more topics in the quality literature and substantially increase the amount of time necessary to complete the exercise. The instructor may wish to assign these topics to small groups of students for further discussion or for short papers.

5. Summary and Conclusions

The repeated analysis of data based on an industrial case study by successively increasing the total sample size has been found to be a useful tool by the authors in summarizing and unifying many analysis of variance concepts. Concepts such as power, statistical significance, statistical non-significance, material differences, Type I and Type II errors, and the effects of sample size are clarified and illustrated progressively.

6. Getting The Data

The file fadata.dat.txt contains the raw data. The file fandata.txt is a documentation file containing a brief description of the dataset.

---

Appendix - Key to Variables in fandata.dat.txt

Data columns 1, 5, 9, and 13 are values for the two types of hole tested (coded -1 and 1)
Data columns 2, 6, 10, and 14 are values for the two types of assembly method used (coded -1 and 1)
Data columns 3, 7, 11, and 15 are values for the two barrel shapes used (coded -1 and 1)
Data columns 4, 8, 12, and 16 are the breaking torques in foot-pounds for the various factor combinations

Data columns 1 - 4 give the results of the experiment with one observation for each factor combination
Data columns 5 - 8 give the results of the experiment with two replications for each factor combination
Data columns 9 - 12 give the results for the experiment with four replications for each factor combination
Data columns 13 - 16 give the results for the experiment with eight replications for each factor combination

Acknowledgements

The authors wish to thank the referees, the associate editor, and the editor for their constructive comments leading to improvements in this paper.

References

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978), *Statistics for Experimenters*, New York: John Wiley and Sons.

Box, G. E. P. "Effects of Inequality of Variance and Correlation Between Errors in the Two Way Classification," *Annals of Mathematical Statistics*, 25, 484-498.

Cochran, W. G., and Cox, G. M. (1957), *Experimental Designs*, New York: John Wiley and Sons.

Desu, M. M., and Raghavarao, D. (1990), *Sample Size Methodology*, Boston: Academic Press.

Hicks, C. R. (1993), *Fundamental Concepts in the Design of Experiments* (4th ed.) New York: Saunders College Publishing.

Montgomery, D. C. (2001), *Design and Analysis of Experiments* (5th ed.) New York: John Wiley and Sons.

Nelson, L. S. (1985), "Sample Size Tables for Analysis of Variance," *Journal of Quality Technology*, 17 (3), 167-169.

LeRoy A. Franklin
Rose-Hulman Institute of Technology
5500 Wabash Avenue
Terre Haute, IN 47803
USA
LeRoy.A.Franklin@Rose-Hulman.edu
