Naive Analysis of Variance

W. John Braun
University of Western Ontario

Journal of Statistics Education Volume 20, Number 2 (2012)
http://www.amstat.org/publications/jse/v20n2/braun.pdf

Copyright © 2012 by W. John Braun 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: Data visualization; Graphical ANOVA; Resampling.

Abstract

The Analysis of Variance is often taught in introductory statistics courses, but it is not clear that students really understand the method. This is because the derivation of the test statistic and $p$-value requires a relatively sophisticated mathematical background which may not be well-remembered or understood. Thus, the essential concept behind the Analysis of Variance can be obscured. On the other hand, it is possible to provide students with a graphical technique that makes the essential concept transparent. The technique discussed in this article can be understood by students with little or no background in probability or statistics. In fact, only the ability to add, subtract, compute averages, and interpret histograms is required.

1. Introduction

One of the most important statistical techniques used in the natural and social sciences is the Analysis of Variance (ANOVA). University researchers and students in these areas are often confronted with experimental data coming from a completely randomized design and wish to test for differences among the underlying populations.
Table 1: The Form of the Data

| Treatment 1 | Treatment 2 | ... | Treatment m |
|-------------|-------------|-----|-------------|
| $X_{11}$    | $X_{21}$    | ... | $X_{m1}$    |
| $X_{12}$    | $X_{22}$    | ... | $X_{m2}$    |
| ...         | ...         | ... | ...         |
| $X_{1n}$    | $X_{2n}$    | ... | $X_{mn}$    |

The data consist of $m$ samples, referred to as treatments. In general, the sample sizes could be different, but in this article, we assume each sample has size $n$, as shown in Table 1, where $X_{ij}$ denotes the $j$th measurement in the $i$th treatment.

The research problem is to determine whether there are any systematic differences among the treatments. The usual method to solve this problem is ANOVA, a technique that is often covered in introductory statistics courses. A relatively high level of mathematical sophistication is required of an individual in order to understand ANOVA, or in the words of one textbook author, the “details of ANOVA are a bit daunting” (Moore 2010, p. 641). The following is a list of the required concepts:

- Null hypothesis/alternative hypothesis
- Variance
- Stochastic independence
- Degrees of freedom
- Sums of squares
- $F$-distribution
- $p$-value

Recent statistics education research has shown that most introductory-level students do not fully comprehend these concepts. This was noted by Weinberg et al. (2010) in their review of several studies. The implication is that there are several points at which a breakdown in the understanding of ANOVA can occur. This view is consistent with results reported by Tintle et al. (2011) who spoke of students who have “lost sight of the big picture of statistics (arguably, real data analysis and inference)” by the time they have completed the first two thirds of a traditional introductory statistics course. Such students are not likely to be mentally prepared for the “daunting” experience of classical ANOVA.
David Moore responded to this difficulty in his introductory-level statistics textbook by deferring the ANOVA formulas to an optional section at the end of his ANOVA chapter. His focus is on what he believes to be the principal concept:

The main idea of ANOVA is more accessible and much more important. Here it is: when we ask if a set of sample means gives evidence for differences among the population means, what matters is not how far apart the sample means are but how far apart they are relative to the variability of individual observations (Moore 2010, p. 642).

The goal of this article is to introduce a technique that graphically illustrates the essential idea behind ANOVA. Students gain facility with a technique which will allow them to properly analyze their data while, at the same time, they completely understand the analysis. The technique can be introduced at an early stage in a statistics course, even at the time when descriptive statistics are studied. This approach will help students to see the big picture of statistics. In cases where students will ultimately be introduced to the classical ANOVA, the graphical technique will aid in their understanding of the main concept.

The above goal is in line with Weinberg et al. (2010) who noted that educational researchers have been advocating the development of tools for informal inferential reasoning and who suggested that these tools should facilitate connections between the informal reasoning and the formal concepts. Our objective is similar to that of Lesser and Melgoza (2007) who are also interested in teaching ANOVA to students with a limited mathematical background (i.e. secondary school students). However, their approach is numerical instead of graphical.

Closer in spirit to the present proposal is a graphical ANOVA plot which has been implemented by Barrios (2009) in an R package based on ideas presented by Box et al. (2005). Treatment averages are calculated and subtracted from the observations, yielding residuals which are displayed as a dot plot. A second dot plot displays the treatment averages after centering and scaling so that their average and variance match those of the residuals, under the hypothesis of no treatment effect. That is, the overall average is subtracted from each treatment average, and the results are multiplied by the square root of the treatment sample size (i.e. $\sqrt{n}$) since the standard deviation of each treatment average equals the noise standard deviation divided by $\sqrt{n}$. The spread of the centered-scale treatment averages is then compared with the spread of the residuals; if the treatment averages are more spread out, we have evidence against the hypothesis of no treatment effect. A specific example is pictured in the right panel of Figure 1. This graphical ANOVA plot has much to recommend it: it is very simple to implement and to interpret. However, for a student with a limited understanding of statistical distribution theory, the construction is not completely transparent.
Figure 1: Naive and Graphical ANOVA Plots for the Airplane Data (see Section 2). Left panel: Naive ANOVA plot for airplane experiment data. Right panel: Graphical ANOVA plot (Barrios 2009). Sample sizes are each 4. The $p$-value for the classical ANOVA $F$-test is 0.0282.

We should also note the interesting approach taken by Sturm-Beiss (2005) who used a Java applet to help visualize ANOVA with one and two factors. Danielak et al. (2011) also provided a useful graphical aid to the interpretation of the within and between sums of squares in ANOVA. These approaches, however, ultimately remain concerned with the calculation of test statistics and $p$-values. They are a welcome addition to the data analyst’s toolbox, but they require more sophistication on the part of the student than we are assuming here.

The rest of the article proceeds as follows. In the next section, we present our proposal for a simple ANOVA plot within the context of an Analysis of Variance problem suitable for students and researchers with a limited mathematical/statistical background. In Section 3, simulation results demonstrate that the plot will provide the user with reasonably accurate conclusions. The article continues with additional illustrative examples in Section 4 followed by a short conclusion in Section 5.

2. The Naive ANOVA Plot

We will introduce our graphical technique via the following problem: does the distance traveled by a paper airplane after being thrown depend on the weight of the paper used in its construction? An experiment designed to answer this question can be run in a class or lab setting. Such experiments have long been advocated by statistics educators (e.g. Mackisack
Table 2: Airplane Experiment Data. Flight distances (in meters) for 12 paper airplanes, according to the weight of the paper used in the construction of each airplane.

|       | light | medium | heavy |
|-------|-------|--------|-------|
| 3.1   | 4.0   | 5.1    |
| 3.3   | 3.5   | 3.1    |
| 2.1   | 4.5   | 4.7    |
| 1.9   | 6.1   | 5.3    |

1994). What follows is the description of such an experiment and the subsequent analysis.

Paper airplanes were constructed using a single design, with 20 cm by 27 cm sheets of paper having one of three different weights (light, medium, or heavy). Four sheets of each type of paper were folded into airplanes by a single individual (the author), and each was flown once, launched by the same individual. The 12 flight distances were measured and recorded as in Table 2. Thus, there are $m = 3$ treatments (corresponding to paper weight), each having sample size $n = 4$. In a class discussion, attention should initially be drawn to the fact that not only are the flight distances different for different treatments, but they are also different within treatments. What has caused the variation within each treatment? Factors other than paper weight must be having an effect. Identifying the nature of these other factors is a worthwhile exercise, because it connects the concept of “unexplained variation,” “error,” or “noise” with concrete notions that the students can understand. Some of the possibilities here are: individual airplane construction, initial throwing height, and initial thrust and direction. These unmeasured factors probably vary slightly from throw to throw and thus could account for some or all of the variation observed within each treatment.

The presence of unmeasured factors makes it more difficult to tell if the factor of interest (i.e., paper weight) is responsible for differences in flight distance. For example, if we calculate the average of the distances for each treatment, as in Table 3, we will likely (and, in this case, we do) see differences in those averages, even if the weight of the paper has no real effect. Are the differences we see in the treatment averages due only to the unmeasured factors, or does paper weight really have an effect on mean flight distance?

Table 3: Average Flight Distance by Weight of Paper

|       | light | medium | heavy |
|-------|-------|--------|-------|
| 2.600 | 4.525 | 4.550  |
In order to answer this question, we will create a new, artificial, set of data which is similar to the original dataset in many ways. However, we will create this artificial set of data in a way that ensures that only the unmeasured factors will be responsible for any variation that we see in averages of samples taken from this dataset.

From this artificial dataset, we will repeatedly take samples of size 4, with replacement, and, for each sample, we will compute the average. A histogram of these simulated averages will allow us to see how averages from samples of size 4 should vary if only variation due to unmeasured factors is present. Against this histogram, we can compare the original treatment averages; these should have a similar pattern of variation if paper weight has no effect, and we should see more spread in them if different paper weights lead to different mean flight distances.

The first step in the creation of the artificial dataset is to remove possible effects due to paper weight by subtracting the treatment averages from each flight distance measurement, as shown in Table 4. The effect of these subtractions is to produce a dataset where there is no between-treatment variation at all (all treatment averages are now 0), but where the within-treatment variation remains unchanged (i.e. note that the treatment ranges remain unchanged, for example). Thus, all of the variation observed in this dataset must be due to unmeasured factors. This kind of variation is referred to as “noise” or “residual” variation.

The final step in the construction of our artificial dataset is to add the overall or grand average to each noise measurement. This step ensures that the artificial “flight distances” will be centered at the same location as the original flight distances. This will make it easier when we compare the treatment averages with the simulated averages. Adding the overall average to each of the noise measurements is demonstrated in Table 5. Because the same value has been added to all columns of numbers, we have not altered the pattern of variation in the dataset; all observed variation in this artificial dataset remains due only to unmeasured factors.
Table 5: Creating Artificial Flight Distance Data. Addition of the grand average to the noise measurements in Table 4 to create an artificial “flight distance” dataset where weight of paper does not have an effect.

| light     | medium  | heavy   |
|-----------|---------|---------|
| 0.5 + 3.89 = 4.39 | -0.525 + 3.89 = 3.365 | 0.55 + 3.89 = 4.44 |
| 0.7 + 3.89 = 4.59 | -1.025 + 3.89 = 2.865 | -1.45 + 3.89 = 2.44 |
| -0.5 + 3.89 = 3.39 | -0.025 + 3.89 = 3.865 | 0.15 + 3.89 = 4.04 |
| -0.7 + 3.89 = 3.19 | 1.575 + 3.89 = 5.465 | 0.75 + 3.89 = 4.64 |

From the artificial dataset, we proceed to take a large number of samples of size 4, for example, \{4.04, 2.44, 3.39, 4.64\}, \{2.865, 4.64, 3.865, 4.59\}, .... Taking averages gives 3.63, 3.99, ... A histogram of these simulated averages is plotted in the left panel of Figure 1. The treatment averages are also plotted on the horizontal axis for ease of comparison.

Noteworthy is the appearance of the light paper treatment average as an outlier relative to the histogram. The other two treatment averages are very close together. The interpretation should be clear: the light paper airplanes do not travel as far on average as the other types of paper airplanes. We have strong evidence that the treatment means are not all the same. If they had been the same, we would expect the treatment averages to plot in regions corresponding to higher histogram density.

The right panel of Figure 1 shows the graph resulting from Barrios’ (2009) approach, which was described in the Introduction. The interpretation of this graph is the same as that for the naive ANOVA plot.

We conclude this section with a summary of the naive graphical ANOVA procedure:

1. Compute the treatment averages $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_m$ and the grand average, $\bar{X}$.

2. Compute residuals

   $$e_{ij} = X_{ij} - \bar{X}_i,$$

   for $j = 1, \ldots, n$, and $i = 1, \ldots, m$.

3. Construct a sample of simulated or artificial observations whose variation is not due to the factor of interest by adding the grand average to each residual:

   $$X^*_{ij} = \bar{X} + e_{ij}.$$

4. Take a sample of size $n$ from the complete set of simulated observations and compute the average.
5. Repeat the preceding step \( k \) times, to obtain a collection of \( k \) simulated averages. The value of \( k \) should be taken to be fairly large, say 100 or more.

6. Display the \( m \) original treatment averages and the \( k \) simulated averages on the same graphic, so that the spreads of the two distributions can be compared.

A comparison of the spreads in the distributions of the original treatment averages and the simulated sample averages will now give a concrete impression as to whether the true means must really have been different or not. A number of graphical techniques could be used for this purpose. In this article, we construct a histogram of the simulated averages and plot the treatment averages on the horizontal axis (i.e., a “rug” plot), since students tend to be familiar with histograms so interpretation may be easier. Using dot plots is a reasonably straightforward alternative, as are QQ plots, though these might require more sophistication on the part of the user. With any of these implementations, the strength of evidence against the null hypothesis that the true means are equal will now be visually apparent.

As with any statistical procedure, this one will fail to be valid under certain circumstances. The measurements should be independent and the within-sample variances should be equal. Unlike the Barrios (2009) proposal, normality is not required.

3. Simulation Study

We conducted simulations in order to check the accuracy of the graphical technique and to see if there is a substantial loss to the researcher who would opt for this graphical method over the use of the classical ANOVA.

For the purpose of these simulations, a rule was needed to establish whether the naive ANOVA method was yielding strong evidence against the null hypothesis or not: if the extremes of the treatment averages were in the upper and lower 2.5% tails of the simulated sample average distribution, the null hypothesis of equal means was rejected. This rule is somewhat arbitrary and is not the only possible way to judge whether the treatment average locations are consistent with the null hypothesis; the idea of using reference plots as described at the end of this section is likely to be more useful in practice.

For each of the scenarios we considered, 10,000 “experiments” were simulated. The naive ANOVA “test” was conducted using \( k = 75 \) and \( k = 150 \). For comparison purposes, the \( p \)-value for the classical ANOVA was computed. A result from the graphical ANOVA method of Barrios (2009) is also included; again, this method is not really quantitative, so
an arbitrary rule was set: in this case, rejection of the null hypothesis occurred when the extremes of the (scaled) treatment averages were outside the range of the residuals.

### 3.1 Normally Distributed Scenarios

The first four scenarios we considered were normal-based samples, with \( m = 2, 3, 4 \) and 5, respectively. The standard deviation \( \sigma \) was set to the value 1 each time. For each of these scenarios, we simulated from the null distributions, in order to assess the effective test size. We then simulated from an alternative case where the factor level means were taken to be different; power was estimated at this alternative. The specific scenarios considered were:

- **Scenario 1:** \( m = 2, n = 8 \);
  - parameters at the alternative: \( \mu_1 = 0 \) and \( \mu_2 = 1 \).

- **Scenario 2:** \( m = 3, n = 4 \);
  - parameters at the alternative: \( \mu_1 = 0, \mu_2 = 1 \), and \( \mu_3 = 2 \).

- **Scenario 3:** \( m = 4, n = 10 \);
  - parameters at the alternative: \( \mu_1 = 0, \mu_2 = 0.5, \mu_3 = 1 \), and \( \mu_4 = 1.5 \).

- **Scenario 4:** \( m = 5, n = 6 \);
  - parameters at the alternative: \( \mu_1 = 0, \mu_2 = 0.5, \mu_3 = 1, \mu_4 = 1.5 \), and \( \mu_5 = 2.0 \).

For each scenario, the proportion of Type I errors (\( \alpha \)) and the estimates of power under the given alternative were recorded in Table 6.

### 3.2 Nonnormal Scenarios

We also considered some nonnormal situations, since the technique should work in these cases. Again, test sizes were estimated and power estimates were obtained for a specific alternative. **Table 7** gives results from simulations according to three nonnormal scenarios. For all of these situations we restricted our attention to the cases where \( m = 3 \) and \( n = 4 \). The distributions under the null hypothesis were:

- **Scenario 5:** \( t \) distribution on 3 degrees of freedom
- **Scenario 6:** \( t \) distribution on 2 degrees of freedom
- **Scenario 7:** uniform distribution on the interval \([-3, 3]\)
### Table 6: Simulation Results for Normal Scenarios

| Scenario | Method        | $\alpha$ | power   |
|----------|---------------|----------|---------|
| 1        | $k = 75$      | 0.0198   | 0.3015  |
|          | $k = 150$     | 0.0181   | 0.2875  |
|          | ANOVA         | 0.0461   | 0.4554  |
|          | graphical ANOVA | 0.0166 | 0.2731  |
| 2        | $k = 75$      | 0.0687   | 0.5969  |
|          | $k = 150$     | 0.0603   | 0.5801  |
|          | ANOVA         | 0.0472   | 0.5596  |
|          | graphical ANOVA | 0.0830 | 0.6278  |
| 3        | $k = 75$      | 0.0418   | 0.6855  |
|          | $k = 150$     | 0.0330   | 0.6746  |
|          | ANOVA         | 0.0510   | 0.8138  |
|          | graphical ANOVA | 0.0125 | 0.4903  |
| 4        | $k = 75$      | 0.0760   | 0.7801  |
|          | $k = 150$     | 0.0663   | 0.7784  |
|          | ANOVA         | 0.0481   | 0.8138  |
|          | graphical ANOVA | 0.0371 | 0.6604  |

NOTE: Simulation results for four normal distribution scenarios. The rows labeled “$k = 75$” and “$k = 150$” contain results from the naive ANOVA method, the “ANOVA” rows give the results from classical ANOVA and the “graphical ANOVA” rows give the results from Barrios (2009) graphical ANOVA.

The distributions under the alternative hypothesis were obtained by additively changing the means for the second and third samples. The mean for the second sample was taken to be 1, and the mean for the third sample was taken to be 2.

### 3.3 Discussion

From this simulation study, we see that the ANOVA nominal and actual test sizes match very well in the normal and uniform cases, as expected. For the $t$ cases, the actual test sizes differ from the nominal level, again, as expected. The graphical approaches do not behave quite as well as the classical ANOVA in the normal cases, but their behavior does not seem to deteriorate in the nonnormal cases. Test sizes are not alarmingly different from the nominal (though in some cases they are quite conservative). The risk to the user is primarily through the possible loss of some power, though the design of this simulation study may be exaggerating such effects (see the next subsection for more on this). Overall, the graphical methods both provide a “rough-and-ready” approach to the Analysis of Variance.
Table 7: Simulation Results for Nonnormal Scenarios

| Scenario | Method           | $\alpha$ | power  |
|----------|------------------|----------|--------|
| 5        | $k = 75$         | 0.0586   | 0.3647 |
|          | $k = 150$       | 0.0509   | 0.3475 |
|          | ANOVA           | 0.0359   | 0.3314 |
|          | graphical ANOVA | 0.0611   | 0.3629 |
| 6        | $k = 75$         | 0.0508   | 0.2627 |
|          | $k = 150$       | 0.0446   | 0.2491 |
|          | ANOVA           | 0.0308   | 0.2243 |
|          | graphical ANOVA | 0.0523   | 0.2556 |
| 7        | $k = 75$         | 0.0719   | 0.2350 |
|          | $k = 150$       | 0.0663   | 0.2165 |
|          | ANOVA           | 0.0559   | 0.1968 |
|          | graphical ANOVA | 0.0908   | 0.2867 |

3.4 Use of Reference Plots

The rejection rules defined at the beginning of this section were set in order to simplify the resulting simulation study. These rules are likely causing less than optimal performance of the naive ANOVA plot.

As with QQ-plots, one’s ability to use the naive ANOVA plot effectively will improve with experience. Assessing strength of evidence will depend on the number of treatments under study. Reference plots are a good way to gain some experience before making a judgement based on given data. In this section, four plots are displayed which correspond to cases where the $p$-value from the ANOVA $F$-test is in the vicinity of 0.05. They serve as possible benchmarks for deciding whether a given plot based on actual data provides a strong case that the treatment means differ. These particular plots are based on different sample sizes and different numbers of treatments and are displayed in Figure 2.

4. Additional Illustrative Examples

We now present three more examples of the naive ANOVA plot, further demonstrating that the conclusions that one would draw from it are not all that different from what would be concluded from the ANOVA $p$-value. We use the histogram version of the graphic in our examples.
Figure 2: Naive ANOVA Reference Plots. Top left panel: Naive ANOVA plot for two samples, each of size 8. The $p$-value for the ANOVA $F$-test is 0.0389. Top right panel: Naive ANOVA plot for three samples, each of size 4. The $p$-value for $F$-test is 0.0599. Bottom left panel: Naive ANOVA plot for four samples, each of size 10. The $p$-value for the ANOVA $F$-test is 0.0457. Bottom right panel: Naive ANOVA plot for five samples, each of size 6. The $p$-value for the ANOVA $F$-test is 0.0436.

The top left panel of Figure 3 exhibits the naive ANOVA plot for the noise vibration data which can be found in the Devore5 library (Bates 2004). In this case, five different brands of bearings are compared in terms of the amount of vibration they generate in an electric motor.
Figure 3: Three Examples of Naive ANOVA Plots. Top left panel: Naive ANOVA plot for the motor vibration data. Top right panel: Naive ANOVA plot for the rat arousal data. Bottom Panel: Naive ANOVA plot for the agricultural dataset ex09.66.

The second brand average appears in the extreme right tail of the distribution of simulated averages, suggesting that there is strong evidence of a difference among the different brands. In this case, it seems that the second brand is different from the others, and most particularly, it seems to be different from the fifth brand.

The differences among the means evident in the naive ANOVA plot agree with the $p$-value obtained in the classical ANOVA ($0.0001871$).
The top right panel of Figure 3 exhibits the naive ANOVA plot for the rat arousal data of Danielak et al. (2011), which concerns the effects of four different drug treatments on the behaviour of samples of 10 rats. The Placebo treatment appears to be very different from the combined drug treatment, while the single drug treatments lie in the intermediate region of the distribution. There is clear evidence of a treatment effect here, which is in agreement with the classical ANOVA ($p$-value $= 0.0000417$).

Our final dataset comes from the Devore5 library (Bates 2004). It concerns the effects of a fertilizer on agricultural yield. There are two treatments: a fertilizer treatment and a control treatment. There are eight measurements in each sample. The $p$-value for the classical ANOVA is 0.847, which agrees with the result pictured in the bottom panel of Figure 3, where the treatment averages lie very near the center of the distribution of simulated averages.

5. Concluding Remarks

In this article, a graphical procedure has been proposed as a way of conveying the main point of the Analysis of Variance to students who might otherwise struggle with the classical approach to the Analysis of Variance. The method is based on simple arithmetic and an elementary randomization approach. It can be learned on its own or as a precursor to the classical approach.

The naive ANOVA procedure is based on a nonparametric bootstrap, but the student does not need to make this connection. Ricketts and Berry (1994) made a strong case for the use of resampling in elementary statistical demonstrations. This view was more recently reinforced by Tintle et al. (2011) where the case was made for randomization as a principal means of conveying inference concepts.

The classical approach to ANOVA gives an $F$-ratio and then a $p$-value, inference concepts which require some effort to properly interpret. It should not be surprising when a student fails to see the connection between ANOVA and the original research question.

Making this connection is much easier with the naive approach. The proposed graphic simultaneously conveys the strength of statistical evidence against the null hypothesis while displaying the relative locations of the treatment averages. The student is brought directly back to the original question: a glance at the naive ANOVA plot will reveal if the means look different or not.
Appendix

The following R code (R Core Team 2012) can be used to produce the naive ANOVA plots found in this article:

```r
naiveANOVA <- function(dataset, k=150){
  color <- c("black", "red", "blue", "green4", "purple", "green", "brown")
  dataset <- dataset[order(dataset[,1]),]
  means <- sapply(split(dataset[,2],dataset[,1]),mean)
  n <- sapply(split(dataset[,2], dataset[,1]), length)
  sim.data <- dataset[,2] - rep(means, n) + mean(dataset[,2])
  sim.means <- NULL
  for (i in 1:k) {
    sim.sample <- sample(sim.data, size=n, replace=TRUE)
    sim.means <- c(sim.means, mean(sim.sample))
  }
  hist(sim.means, xlim=range(c(sim.means, means)),
       xlab="simulated treatment averages", main="", cex.lab=1.4)
  points(means, rep(0, length(means)), cex=2.5, col=color[1:length(means)],
         pch=14+1:length(means))
  legend("topleft", legend= c(paste("average", unique(dataset[,1]))),
         pch=14+1:length(means), col=1:length(means))
}
```

Data should be supplied in a two-column data frame, where the first column contains the factor, and the second column contains the response values. As an example, the following data were used to produce the top left panel of Figure 2:

```
x   y
1  0.0
1  0.2
1 -0.6
1 -0.8
1 -1.1
1 -2.0
1  0.5
1 -0.5
2  1.4
2 -0.1
2  1.1
2  1.1
2  0.0
2  0.0
2 -0.5
2 -0.3
```
Acknowledgments

Helpful comments from the editor, associate editor, and two anonymous referees led to a substantial improvement in this article, and are gratefully acknowledged. The author is also grateful to Vincent Marchese, Andrew Noens, Sesen Berhane, Phillip Braun, Andrew Lanz-O’Brien, Lengyi Han, Lisa Hunt, Jennifer Dungavell, Karen Valentine, Duncan Murdoch, David Bellhouse, Matt Davison, Ben Rubin, Bethany White and Doug Woolford for providing feedback on the method proposed in this article. This article is dedicated to Tom Wonnacott.

References

Barrios, E. (2009), “BHH2: Useful Functions for Box, Hunter and Hunter II. R package version 1.0.3, http://CRAN.R-project.org/package=BHH2.

Box G. E. P, Hunter, J. S., and Hunter, W. C. (2005), Statistics for Experimenters II, New York: Wiley.

Bates, D. (2004), Devore5: Data sets from Devore’s “Prob and Stat for Eng (5th ed),” R package version 0.4-5.

Danielak, B. A., Pruzek, R. M., Doane, W. E. J., Helmreich, J. E., and Bryer, J. (2011), granovaGG: Elemental Graphics for Analysis of Variance using ggplot2. R package version 1.0, http://github.com/briandk/granovaGG.

Lesser, L. and Melgoza, L. (2007), “Simple Numbers: ANOVA Example of Facilitating Student Learning in Statistics,” Teaching Statistics, 29, 102–105, doi: 10.1111/j.1467-9639.2007.00272.x

Mackisack, M. (1994). “What is the Use of Experiments Conducted By Statistics Students?” Journal of Statistics Education [Online], 2(1), http://www.amstat.org/publications/jse/v2n1/mackisack.html.

Moore, D. (2010), The Basic Practice of Statistics (5th ed.), New York: W.H. Freeman and Company.

R Development Core Team (2012), R: A Language and Environment for Statistical Com-
Ricketts, C. and Berry, J. (1994), “Teaching Statistics through Resampling,” *Teaching Statistics*, 16, 41–44.

Sturm-Beiss, R. (2005), “A Visualization Tool for One- and Two-Way Analysis of Variance,” *Journal of Statistics Education* [Online], 13(1), http://www.amstat.org/publications/jse/v13n1/sturm-beiss.html

Tintle, N., VanderStoep, J., Holmes, V.-L., Quisenberry, B., and Swanson, T. (2011). “Development and Assessment of a Preliminary Randomization-based Introductory Statistics Curriculum,” *Journal of Statistics Education* [Online], 19(1), http://www.amstat.org/publications/jse/v19n1/tintle.pdf.

Weinberg, A., Wiesner, E., and Pfaff, T.J. (2010), “Using Informal Inferential Reasoning to Develop Formal Concepts: Analyzing an Activity,” *Journal of Statistics Education* [Online], 18(2), http://www.amstat.org/publications/jse/v18n2/weinberg.pdf.

W. John Braun
Dept. of Statistical and Actuarial Sciences
University of Western Ontario
1151 Richmond St.
London, ON N6A 5B7
Canada
braun@stats.uwo.ca