The Analysis of Experimental Data

Why some studies cannot be replicated

Donald J. Wheeler

Last month we looked at the analysis of observational data. Here we will consider experimental data and discover a weakness in the way they are obtained that can contribute to the problem of non-reproducible results.

BACKGROUND

The discipline of statistics grew up in agricultural and biomedical research. There a major problem for researchers is the fact that their basic experimental units are fields, livestock, and people. Since these units all differ, any researcher has to find a way to keep these differences from masking the effects of any treatments being studied. And the classic solution for this problem is some form of randomization.

In a completely randomized design we use as many experimental units as possible and randomly assign these experimental units to the various treatments being studied. This randomization seeks to average out, within each treatment, the differences between the experimental units. When this happens the averages for the different treatments can be used to compare the treatment effects. Thus, to isolate and minimize the effects of extraneous variables, randomization and blocking have been the mainstay of experimental designs for the past 100 years. The success of this approach has made the analysis of experimental data one of the major topics in statistical education.

Figure 1 illustrates a completely randomized design. Three treatments are to be compared using 204 experimental units. These units are randomly assigned to the treatments, resulting in 68 units per treatment. The differences between the experimental units creates the spread of responses seen in each histogram. By randomizing the units used with each treatment it is hoped that the differences between the units will average out within each treatment so that the three treatment averages will show the true effects of the treatments.

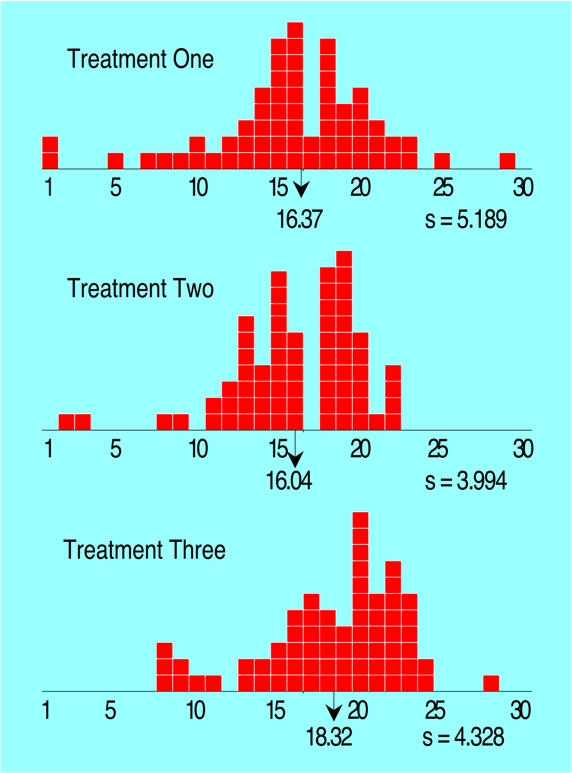

Figure 1: An Experiment Comparing Three Treatments

To compare the treatment averages modern statistical techniques use the variation present within each treatment to filter out the noise. This within treatment variation is intended to incorporate the effects of all of the extraneous variables not considered in the study. And the question is whether the differences between the treatment averages are large enough to be detectable above this background noise within the treatments.

The traditional analysis here would be a one-way analysis of variance (ANOVA). For those who are interested, the appendix summarizes the computations involved. Here we find a p-value of 0.007, which is small enough to satisfy everyone's criterion for what constitutes a detectable difference between the treatments.

A GRAPH IS ESSENTIAL

ANOVA does not provide a picture of the results. However, this can be remedied by using the analysis of means (ANOM). The grand average of all 204 data in Figure 1 is 16.91. The average of the three standard deviation statistics is 4.504. ANOM limits with a 1% probability of exceedance are:

$$\text{Grand Average} \pm 0.291 * \text{Avg. Std. Dev.}$$
$$= 16.91 \pm 1.31 = 15.60 \text{ to } 18.22$$

(For details on this computation see the appendix.) Figure 2 shows this 1% ANOM comparing these three treatment averages.

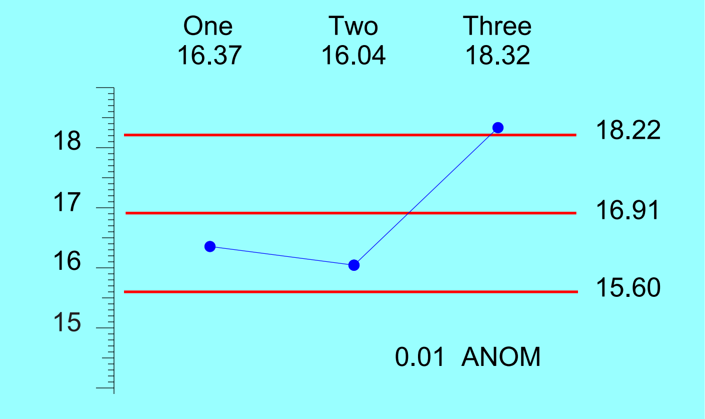

Figure 2: 1% ANOM for Figure 1

ANOM compares each average with the grand average, and treatment three is found to be detectably larger than the grand average. This graph shows the potential signals (the averages) versus the probable noise (the limits) so that we can evaluate just what our p-value of 0.007 is actually telling us. Here the "signal" of Treatment Three is seen to be just slightly outside the limits. This means that, at the 1% level, this signal is just barely detectable above the background noise.

REPLICATION

Here both ANOVA and ANOM show "statistically significant results" at the 1% level. So, can someone else replicate this result?

Replication might be hard in this case since the three treatments here were actually one and the same treatment. All 204 data were collected under the same conditions. The signals found here by both ANOVA and ANOM are entirely due to differences in the three sets of experimental units. When we place the 204 values on an *XmR* chart we get Figure 3.

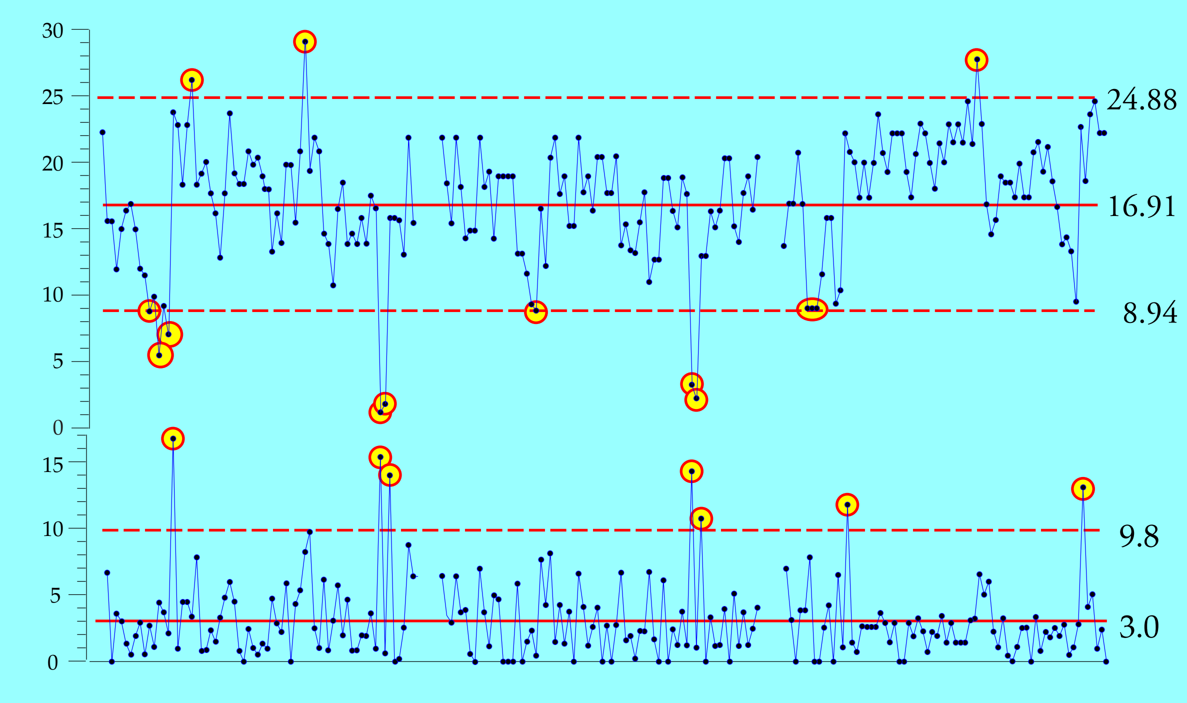

Figure 3: *XmR* Chart for Data of Figure 1

The *XmR* chart characterizes these data as being non-homogeneous. As expected, the experimental units are different, but there is also a difference between the three groups of experimental units. And it is this difference between the groups that was detected by ANOVA and ANOM.

While randomization may allow us to conduct experiments using experimental units that are different, and while it will often work as intended, there is no guarantee that randomization will completely average out the effects of the extraneous variables. As a result, a completely randomized design based on over 200 experimental units can give a "statistically significant result at the 0.01 level" when there is no real difference between the three treatments in the study.

This is why the non-trivial replication of results is so important. A "statistically significant result" is merely an indicator of a *possible* difference. As seen here, the panacea of randomization may not work within each treatment to effectively average out the differences in the experimental units. Randomization may be better than doing nothing, but it does not always work as intended.

Other experimental designs using blocking and even more complex restrictions on randomization have been widely used, but the essence of all randomization is the quest to average out the extraneous factors.

ANTI-PERSPIRANT

Thirty-six subjects were used to study two anti-perspirant compounds. The two compounds were used with each subject (a randomized complete block design). Using a random assignment, the two compounds were applied to each subject's forearms and after a specified period of time the galvanic skin response for each forearm was recorded. The researchers threw the data into the computer, performed a one-way ANOVA, got a small p-value, and claimed that they had found a detectable difference between the two compounds. (By failing to take the blocks into account the researchers had obtained a flawed analysis.)

Their boss did not understand their analysis or their explanation of the results, so she plotted the data, by subject, in a simple running record. Her plot revealed that, for every subject, right arms perspired more than left arms. Upon inquiry it turned out that every subject in the study was right-handed. (Dominant arms tend to have more muscles and therefore perspire more.)

Next the boss took the data and her analysis to two different statisticians, both of which backed her up. The only signal in the data was the difference between left and right arms. The "statistically significant result" and small p-value found originally had nothing to do with the two compounds. The randomization and blocking had failed to remove the effects of an extraneous variable, and this problem was only discovered by using a simple running record of the original data.

EIGHTEEN MONTHS WASTED

A study funded by the U.S. Department of Energy and the Tennessee Valley Authority was carried out in the Tennessee Valley between Asheville N.C. and Paducah, Ky. This study spent 18 months collecting data, but when the analysis was done no one could make sense of the results. In an effort to salvage this study I was brought in. I worked with an economist as we sought to find explanations that fit the data. Nothing worked. Finally, in desperation, I started looking through the stack of original data sheets. When I found a change in handwriting, I finally asked the right question. The only signal in the data was the fact that the data had been collected by two different teams. The difference between the two teams perfectly aligned with the graphs of the data. The two teams had understood their jobs differently, and consequently had obtained their data differently.

The next day in an auditorium full of clients, many of whom had flown in from Washington, the principal investigator had to get up and explain that the whole study was a case of garbage-in, garbage-out. There was no useful information about the factors studied over the past 18 months. The difference between the two teams completely obscured everything else in the study.

Most statisticians can tell stories like these of how some extraneous variable managed to sabotage the results of an experimental study. We are even taught to look for these things. Yet many more experiments are performed without the aid of professional statisticians than are done with such help. As long as the p-value is small enough, researchers rush to publish. As a result the journals are full of unreproducible results.

SUMMARY

When we carry out any experiment there are only three things we can do with factors that may influence the response variable: We may study them in the experiment; we may hold them constant during the experiment; or we may ignore them. Blocking is the way we hold some factors constant. Randomization is the way we cross our fingers and hope the factors we have ignored will not mess up our results.

Randomization allows experiments to be conducted in situations where the experimental units are all different. By including many different experimental units for each treatment, randomization seeks to build the non-trivial replication into a single iteration of the experiment. When a single iteration takes months to complete, this built-in replication offers an additional advantage that has made the agricultural/biomedical experimental design model extremely popular. However, there are times when a different model is useful.

In industrial experiments and in clinical studies we are interested in what happens with a single experimental unit. "How can we improve the operation of our machine?" or "Is a given treatment helping a particular patient?" Here randomization has no role to play, and we need to use either replication over time, or a time-series based analysis.

Returning to the traditional experimental designs, there is no guarantee that randomization will work every time. Good experimenters are aware of this fact, and will be on the look-out for things that might have gone wrong. But many researchers remain unaware of this problem. I have had client after client whose only question was "What is the p-value?" As long as it is small, they think they have proved their point Nevertheless, the non-trivial replication of experimental results is the criterion for what is true, correct, and useful. We cannot achieve this by simply using a smaller p-value. Statistical analysis can indicate *potential* signals, but potential is not the same as *proven*.

APPENDIX ANOVA and ANOM FOR FIGURE 1

In ANOVA we characterize the variation within each treatment by computing a variance statistic for each treatment and then averaging them together. The variance statistics for the three treatments are 26.923, 15.953, and 18.730. The average is 20.535. This average within-subgroup variance is known as the mean square within (MSW). Since each treatment variance has 67 degrees of freedom, the MSW is said to have a total of 201 degrees of freedom. This MSW defines the noise-component for our signal-to-noise ratio.

To compute a signal component we use the treatment averages to compute a variance statistic. The three averages are 16.37, 16.04, and 18.32, and the variance statistic is 1.5210. To compensate for the fact that this statistic was computed using the averages, we multiply it by the number of data summarized by each average (n = 68). The result is known as the mean square between (MSB). Here the MSB is 68 x 1.521 = 103.43. This quantity has two degrees of freedom since it was based on three averages.

Our signal-to-noise ratio is known as an F-ratio in honor of Sir Ronald Fisher. It is found by dividing the MSB by the MSW. In this case we get an F-ratio of 5.04. Since this value exceeds the 99th percentile of an F-distribution with 2 and 201 d.f. we have a detectable difference between the treatment averages. (The observed F-ratio of 5.04 falls at the 99.27th percentile, so that it is said to have a p-value of 0.0073.)

In ANOM we plot the treatment averages against detection limits that filter out a specified amount of the background noise. The formula for these limits is:

$$\text{Grand Average} \pm \text{Scaling Factor} * \text{Est. SD(X)}$$

Here our estimate for SD(X) will be the average of the three standard deviation statistics shown in Figure 1. This value is 4.504 and is said to have 199.6 degrees of freedom. The scaling factor for the ANOM shown in Figure 2 is found according to the formula:

$$\text{Scaling Factor} \quad = \quad \sqrt{\frac{k-1}{k\,n}} \quad H$$

Where H is a critical value which depends upon the alpha level (0.01), the number of treatment averages being compared (k = 3) and the degrees of freedom for the estimated standard deviation of the subgroup averages (d.f. = 200). Here, using published tables we find that 2.914 < H < 2.939. Using the larger value, with n = 68, we get the scaling factor of 0.291 to use with the average standard deviation statistic.