Journal of Statistics Education

# Enhancing Dependent Sample Analyses with Graphics

Robert M. Pruzek
State University of New York at New Albany

James E. Helmreich
Marist College

*Journal of Statistics Education* Volume 17, Number 1 (2009), www.amstat.org/v17n1/helmreich.html



---

**Key Words:** Dependent samples; Graphical analyses; Matching; Blocking; Efficient designs; Repeated measures; R software; `granova`

## Abstract

A standard topic in many Introductory Statistics courses is the analysis of dependent samples. A simple graphical approach that is particularly relevant to dependent sample comparisons is presented, illustrated and discussed in the context of analyzing five real data sets. Each data set to be presented has been published in a textbook, usually introductory. Illustrations show that comprehensive graphical analyses often yield more nuanced, and sometimes quite different interpretations of data than are derived from standard numerical summaries. Indeed, several of our findings would not readily have been revealed without the aid of graphic or visual assessment. Several arguments made by John Tukey about data analysis are seen to have special force and relevance.

## 1. Introduction

Nearly all modern introductory statistics texts cover the analysis of dependent sample (or paired) data sets. Generally, some discussion of independence and dependence issues between samples is given and illustrated with an example. The differences between scores in the paired samples are found and examined using single-sample methods that have been discussed earlier in the text. It is rare for the paired data to be plotted or graphed, even when the authors have advocated the practice of looking at data graphically in the context of analysis. It is also rare to see any detailed discussion of the circumstances in which such data sets arise, or what the questions might have been that led to the data collection, questions that could facilitate discussions of what the reader should take from the analysis.

We provide a brief overview of different situations where dependent sample data arise and use real data examples to indicate how or where graphical displays can yield insights in such analyses. We use a function from our R package, `granova`, to depict the data graphically, software that is freely available to anyone with access to the internet. None

of the concepts on which we focus are above the level of discussion and analysis that is possible in a first semester introductory course in statistics for undergraduates; moreover, the discussion aims to illustrate how and why graphical displays deserve to be incorporated in classroom discussions of such data.

In our discussion, we focus on how the dependent paradigm can to advantage be extended beyond two groups. We also highlight special virtues of the dependent sample paradigm, depending on design features associated with data collection. Further, we comment on issues that arise in teaching of dependent sample analysis with graphics.

Four main types of paired dependent samples data should be distinguished:

**1a** Comparisons of two measurement instruments or scales for the same individuals or entities (time of measurement not seen as relevant);

**1b** Examination of trends or effects for repeated measures data (often with treatment intervention between measurements);

**2a** Comparisons of two experimental treatments, or one treatment and a control, for blocks (perhaps pairs) that were initially matched on the basis of prior information, then (perhaps randomly) assigned to treatment;

**2b** Comparisons of matched individuals, perhaps for two observed treatments, where any of various methods were used to form the matched pairs.

In the first two cases, 1a and 1b, the same individuals are measured twice, creating a natural dependency between the groups. In our experience these are the most likely situations to be used as examples of dependent samples in an introductory text. For the next pair of situations, 2a and 2b, associations are created between pairs or blocks of individuals using information available about the experimental cases. Category 2a is clearly most important in terms of facilitating efficient, constructive and informative *causal analyses,* notably when randomization has been used for assignment of units to treatment groups within pre-defined homogeneous blocks. Randomization within blocks underpins experimental comparisons and can minimize doubts about selection bias. For observational data, in contrast to experimental data, it is chiefly selection bias that tends to confound interpretations of between-group comparisons. Category 2b is important vis-à-vis analysis of observational data, particularly in applications of propensity score methods where matching can be used on variables that may have been identified at the end of the treatment comparisons to mitigate selection bias (as example 5 in section  below illustrates). Note that extension beyond pairs to triplets, quads, etc., is generally straightforward, although this idea is not routinely taught.

Executed effectively, it would not be unreasonable to describe dependent sample comparisons (cf., 2a above) as gold standards for experiments. When blocking has been based on prior information that is strongly related to the ultimate response variable, i.e., when blocks are notably homogeneous, then dependent sample designs and their corresponding analyses can provide highly effective paradigms for comparing experimentally defined groups. In other words, dependent sample designs can lead to especially efficient and scientifically informative studies with near-optimal statistical properties since they can often use the prior information in particularly effective ways. Although these (2a) designs are quite versatile, and can account for many real-world complexities (and lead to interaction discovery), they are not widely taught, nor apparently broadly understood. More will be said about the possibilities in the discussion section later.

When blocks of size two are homogeneous, then as will be illustrated in our first two examples, experimental comparisons of treatments within blocks can lead to clear and effective answers about treatment effects even when samples are relatively small. In cases where experimental effects have been found to be homogeneous across blocks,

the mean difference score is most easily interpreted. If treatment assignments have been made at random for each block, then each post-treatment, within-block response difference may be taken as independent evidence of experimental effects. When difference scores vary little across blocks, as in our second example below, strong and potentially generalizable statements about experimental effects may be warranted even when sample sizes are quite small. Alternatively, when effect estimates vary substantially across blocks this may be evidence of block-treatment interactions, which can also be informative. Heterogeneous treatment effects will tend to mitigate against overly-simple interpretations of the mean difference score. Graphical displays are generally helpful to discern the difference between results where interpretations may be simple, or straightforward, and when interactions seem apparent. Additionally, these displays can help investigators to discern whether non-linear transformations of response variables may be appropriate (as in the tobacco example, cf. Appendix A, section 6). The interpretation of particular interactions can also be illuminated by careful study of graphical results; some of our illustrations below will show how detailed studies might proceed.

In the following section we present five examples that correspond to categories 2a, 2a, 1b, 1b and 2b, respectively. These will be discussed in turn, with special emphasis on the value of graphical presentations.

## 2. Examples of Dependent Sample Data

We present figures that illustrate use of a particular graphical method, here called a Dependent Sample Difference Score Assessment Plot, to study dependent sample data (for a similar graphical method, see Rosenbaum). Detailed descriptions of the features of this plot may be found below in Appendix B, section 6. The first two examples use real data from true experiments; these are the kinds of studies that can yield relatively strong scientific inferences, especially when certain easily checked conditions are met. The next two examples compare pre- and post-measurements of weights for girls who were involved in a therapy program to treat anorexia, followed by a study of stress related to the prospect of surgery. The final example exhibits use of matching to aid the analysis of observational data.

It has been surprising to learn that in many cases like these, there appear to be interesting patterns, trends, irregularities, etc., that call into question the appropriateness of more standard approaches to analysis, where the mean difference is either tested for significance, or a confidence interval is generated. The main ideas in what follows will be to attempt to reveal as much as possible about what each dependent sample data set may have to say, and to help ensure that interpretation(s) are both as clear and comprehensive as possible.

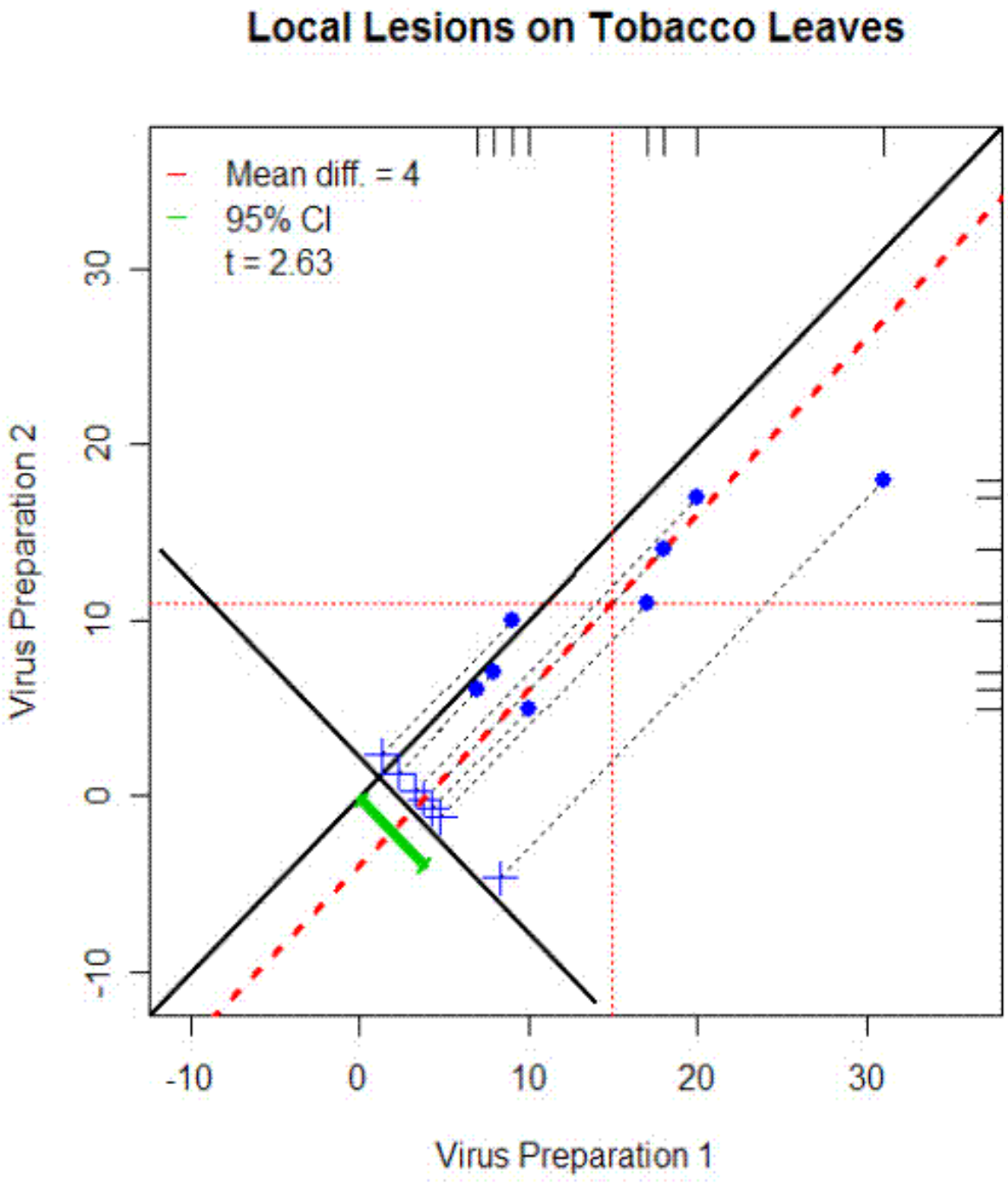


**Figure 1**: Two virus preparations were rubbed on different sides of tobacco leaves. The number of lesions per side for eight leaves are plotted. Youden and Beale.

## 2.1 Example 1: Virus Preparation on Tobacco Leaves

The data shown in Figure 1 were taken from Snedecor and Cochran (1980) and correspond to a true matched pairs experiment. The data originally came from Youden and Beale in 1934 who "wished to find out if two preparations of a virus would produce different effects on tobacco plants. Half a leaf of a tobacco plant was rubbed with cheesecloth soaked in one preparation of the virus extract, and the second half was rubbed similarly with the second extract." (Page 86, Snedecor and Cochran, 1980) Each of the 8 points in the figure corresponds to the numbers of lesions on the two halves of one leaf with sides that had been treated differently. The standard dependent sample *t*-test yields a significant result ($p < .05$), despite the small sample size. (See Appendix A to see the relevance of score transformations in this

context.) The standard analysis focuses only on numerical results and summaries, and a simple test of significance ignores the trend that can be discerned in Figure 1. In particular, the plot shows a tendency for the departures of points from the heavy diagonal line to become larger as the mean ($X$,$Y$) values become larger. That is, the more lesions that are manifest on any one tobacco leaf, the more the $X$ viral extract counts are likely to exceed those for $Y$. The correlation between $X$ and $Y$ is .90. Although this is a small data set, standard analysis shows a statistically significant difference between the two viral extracts, but the $X$ extract shows a tendency to yield more lesions than the $Y$ extract for leaves with more lesions. Before concluding this is the best analysis that one could do, however, see Appendix A.

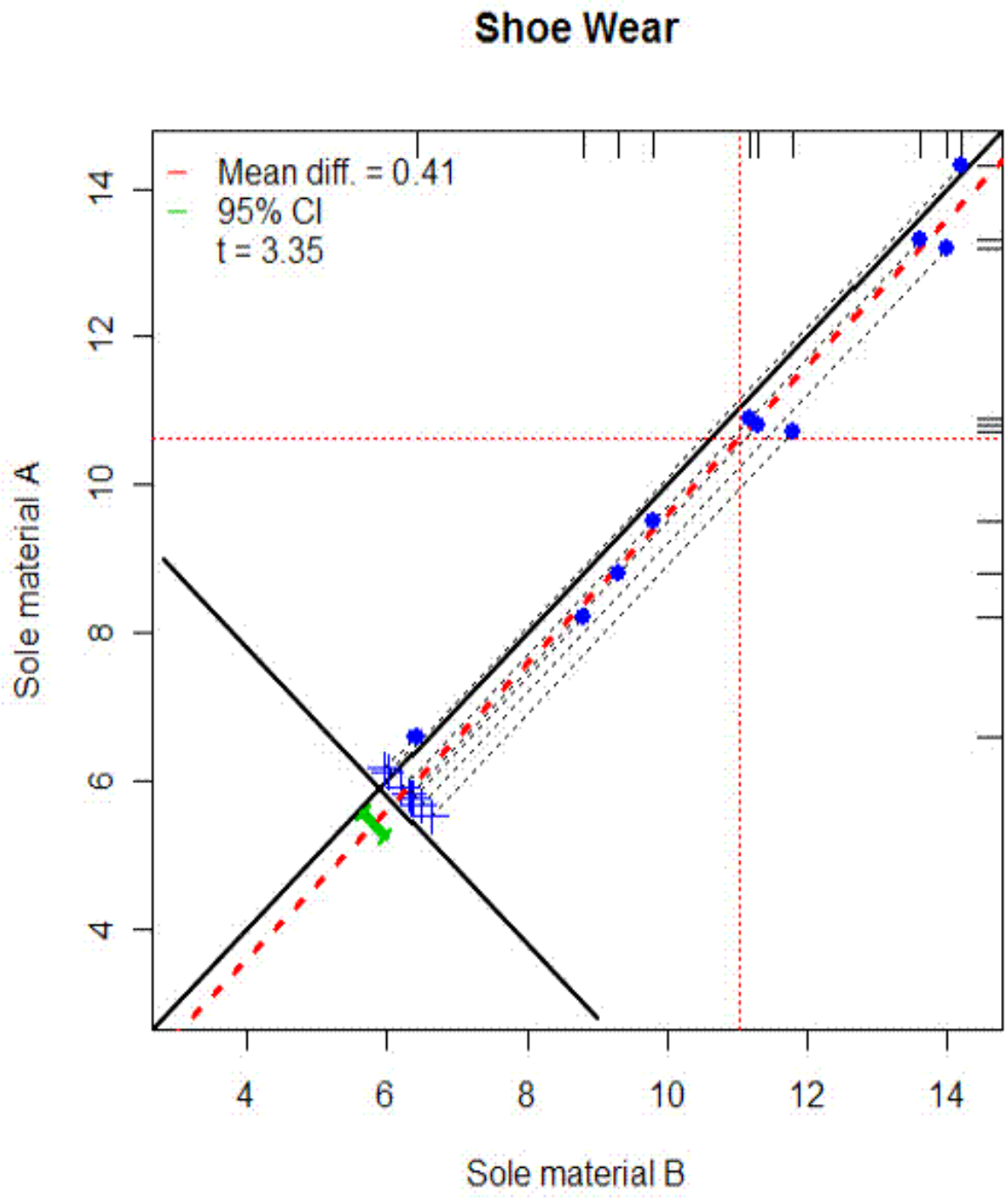


**Figure 2**: Different shoe sole materials for each of eight pairs of shoes test relative wear rates. Box, Hunter and Hunter.

## 2.2 Example 2: Shoe wear

Figure 2 displays the well-known data set on shoe wear initially given by Box, Hunter and Hunter. One sole made of material *A*, and another made of material *B*, were randomly assigned to the two shoes of 10 boys. The *A*, *B* scores are measures of wear for the two materials. This is a 'true experiment' (type 2a) based on matched pairs where a relatively strong cause/effect conclusion is justified, using observations of wear taken some time after the soles had been attached. The numerical summary and the plot show a large standardized effect size, where the *B*-material wore longer than the *A*-material. Indeed, statistical significance is noted (see the *t*-statistic in the legend) despite the small sample size, this being a consequence of the small variation in the differences. Although the high correlation ($r$ = .99) between *A* and *B* scores is related to the relatively small variance of the differences, note that a high correlation alone is not sufficient, since as the preceding example showed, the major axis of the ellipse associated with the *A*, *B* point swarm may not be parallel to the identity line in the plot. Because random assignments had been used with matched pairs, and the shoe data results are so clear, this kind of design can be seen as an exemplar of a true experiment. Blocking was highly effective in reducing variation of the differences, and the response variable metric was also well chosen.

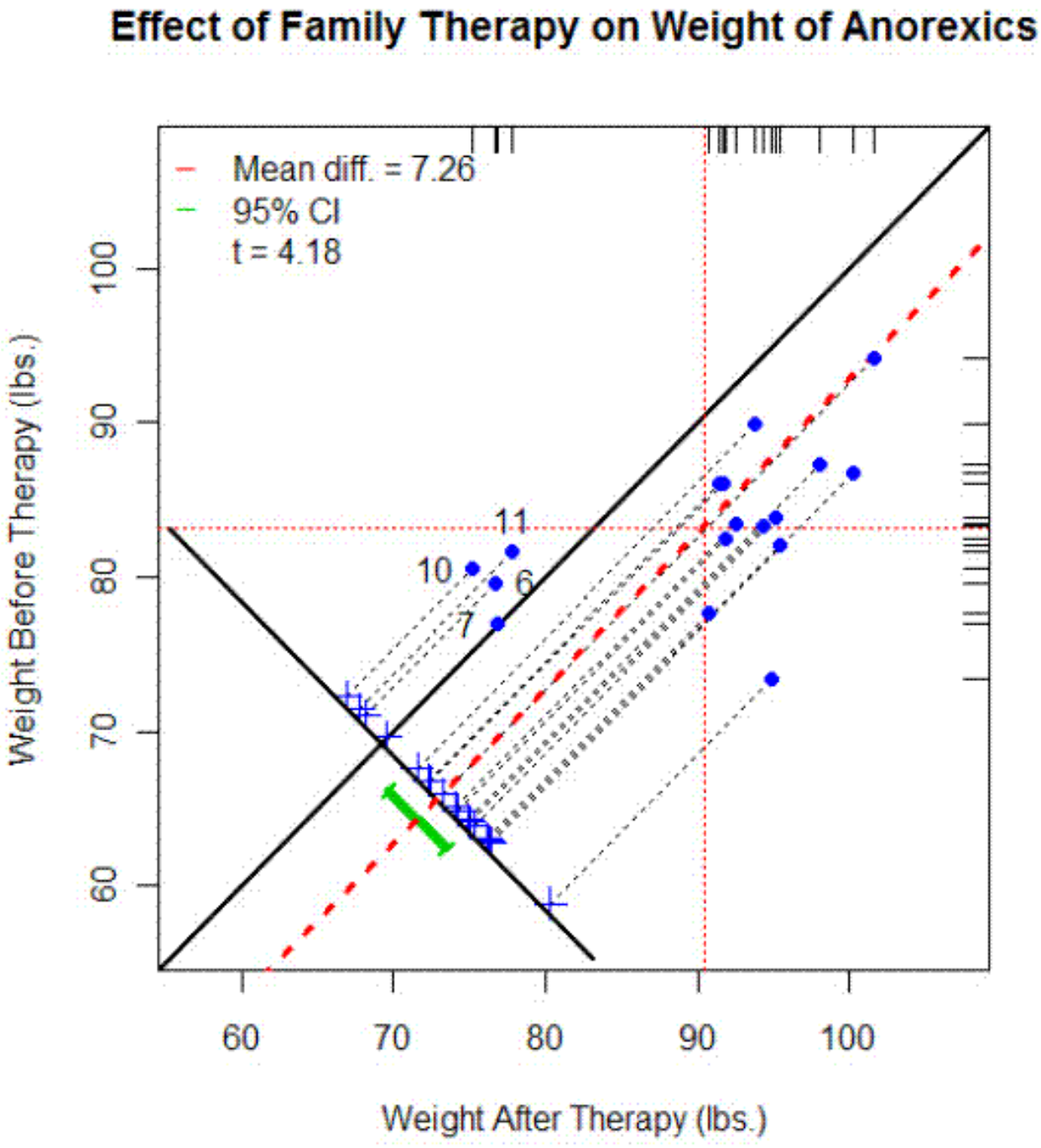


**Figure 3**: Effect of Family Therapy on weights of anorexic girls. Hand et al.

## 2.3 Example 3: Family Therapy as Treatment for Anorexia

The data set used to produce Figure 3 consists of weights in pounds for 17 girls who were weighed before and after treatment for anorexia. These data were originally published by Hand, et al, and were reprinted in Howell. *X* scores are weights (in pounds) after family therapy; *Y* scores are corresponding weights before therapy. A difference score is positive (and below the main diagonal line) for a girl who gains weight, negative if she lost weight. Summary statistics are given in the legend for the usual omnibus question: Is there evidence that girls gained weight following therapy, and if so, is the effect statistically significant? The broad answer is in the affirmative since the average weight gain was 7.26 pounds and the corresponding *t*-statistic is 4.18. Even the standardized effect size, 1.01, is notable.

The plot in Figure 3, however, tells a more nuanced story. The cluster of points at the extreme left (labeled 6, 7, 10, 11) show that these four girls actually lost weight. Indeed, the remaining 13 girls had an average weight gain of 10.4 pounds, and for them the standardized effect size was an impressive 2.26. (For this subgroup, the *t*-statistic moves up to 8.22, although post-hoc data selection undermines probabilistic interpretation of this statistic.) Note also that the rug plot for the post-test scores at the top of the figure shows two distinctive subgroups of scores/weights, while no clusters can be seen on the right-side rug plot that reflects pre-experimental weights. One has to wonder what was different for the four girls who did not profit from their family therapy, and indeed lost weight over its course.

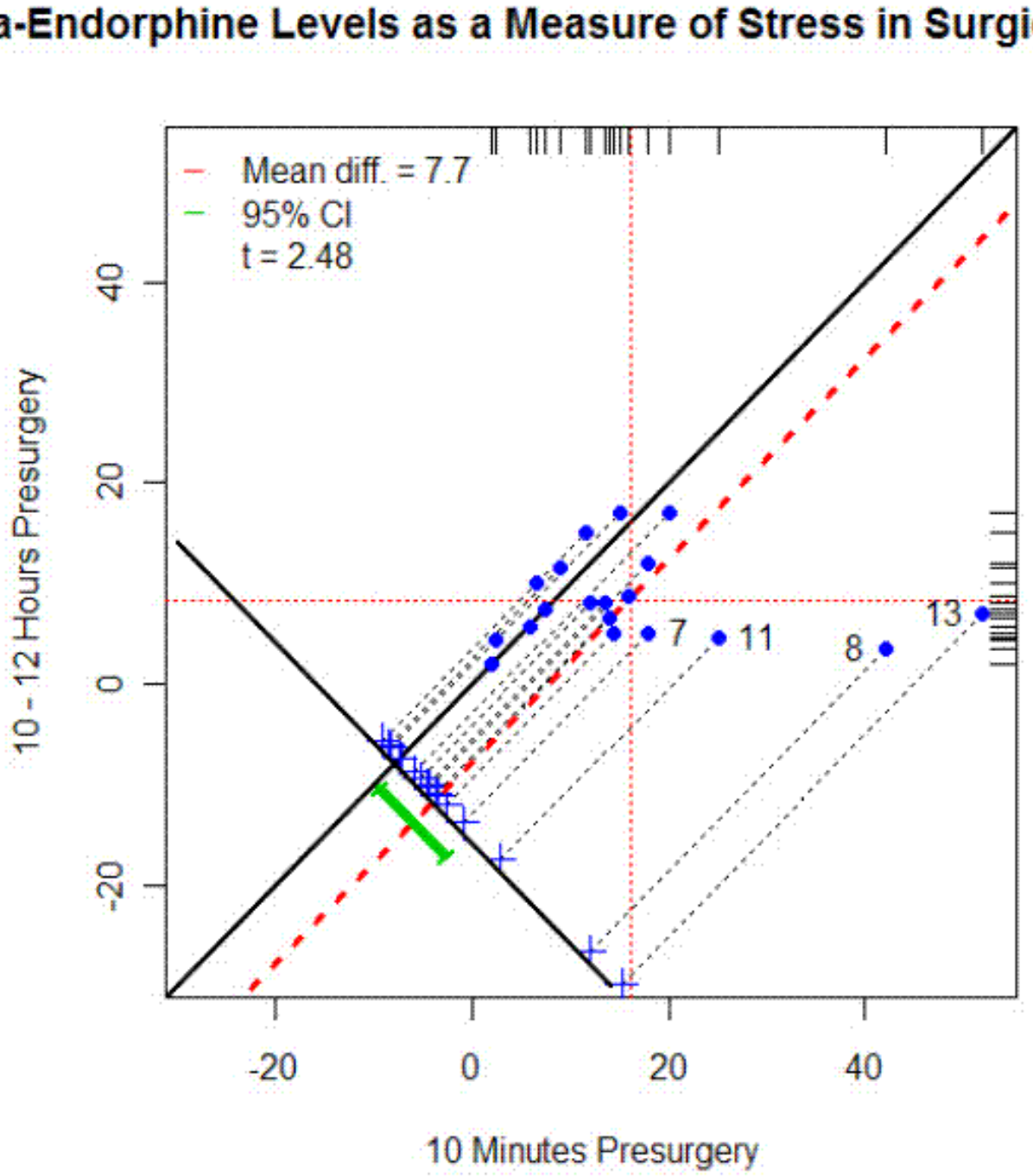


**Figure 4**: Beta-endorphine levels in surgical patients 10 to 12 hours before surgery versus ten minutes before. Hoaglin, Mosteller, Tukey as found in Hand et al.

## 2.4 Example 4: Beta-Endorphine Levels in Surgical Patients

Figure 4 appears to yield new insights for analysis of time-related data, taken from Hand et al, though originally published in Hoaglin, Mosteller, and Tukey. These data concern blood levels of beta-endorphin for 19 patients prior to surgery; beta-endorphin scores are intended to measure stress. The *Y* scores show beta-endorphin levels 12 hours before surgery, *X* scores show levels 10 minutes before surgery. Conventional analysis yields a *t*-statistic equal to 2.48, suggesting that stress levels rise significantly just prior to surgery. Indeed, the standardized effect size of .57 is moderate-to-large by conventional standards. The plot in Figure 4 tells a different story: three or four patients (numbered in the figure 13, 8, 11 and perhaps 7) had beta-endorphin levels much higher just prior to surgery, compared

with earlier scores. If data for these four highest differences are removed, the remaining data depict a much lower stress effect, non-significant ($p > .05$), with a lower standardized effect size of .45. In fact, five of these fifteen people had lower beta-endorphin levels just prior to surgery than 12 hours earlier. Note also that there is little correspondence between these two measures of stress, actually a negative correlation ($r = -.06$), so that use of repeated measures did not have the usual effect of reducing the variance of differences in this situation.

An interaction seems evident; in particular, one would like to know what distinguishes persons who manifest notably higher levels of stress just prior to surgery from those whose beta-endorphin levels were similar on the two occasions. At the least, it seems inappropriate to leave the analysis with nothing more than the conclusion: "surgery significantly increases stress (as evidenced by elevated beta-endorphin levels)" as is stated by Howell.

## 2.5 Example 5: Lead Levels in Children's Blood

Data shown in the Figure 5 are based on an observational study by Morton, et al. Children of parents who had worked in a factory where lead was used in making batteries were matched by age, exposure to traffic, and neighborhood with children whose parents did not work in lead-related industries. Whole blood was assessed for lead content yielding measurements in mg/dl; results shown compare the exposed with control children. Conventional dependent sample analysis shows that the effect size was about 1 standard deviation unit, the 95% confidence interval was far from zero, and the *t*-statistic for the mean of the difference scores is 5.78, so the results support the interpretation that parents' lead-related occupations tend generally to influence how much lead is found in their children's blood.

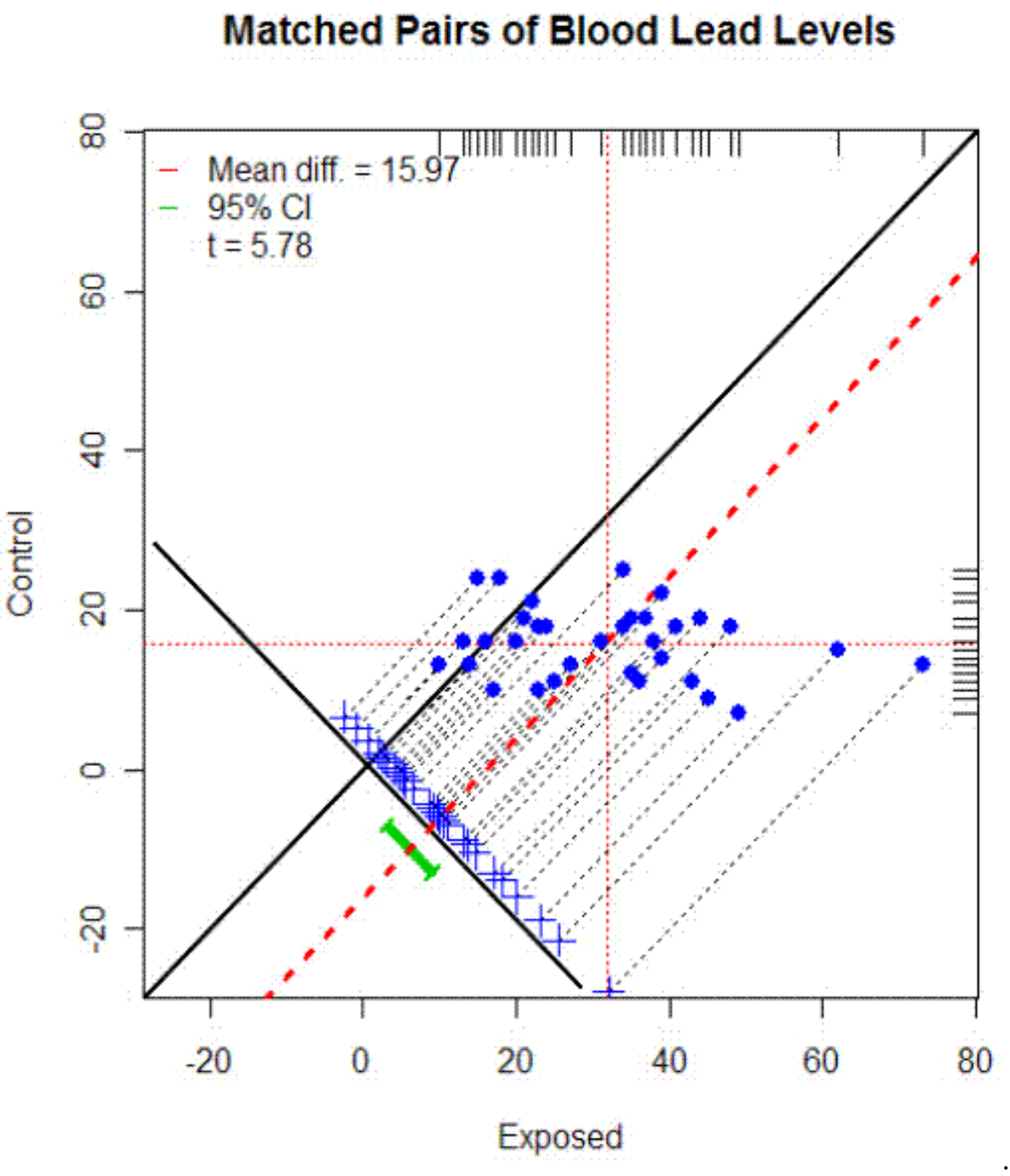


**Figure 5**: Blood lead levels of lead workers' children matched with similar control children. Morton et al [7]

Examination of the graphic adds information. Note the wide dispersion of lead measurements for exposed children in comparison with their control counterparts. One interpretation of this result is that particular features of parent experiences at work or home with their children need to be taken into account to make the comparison most informative. That is, given the wide variation in blood levels across the exposed group, where those with the lowest levels are quite comparable with their control counterparts, but not those corresponding to points on the far right side, it seems reasonable to say that the general hypothesis should be reformulated in terms of specifics. Indeed, the original authors of this study, Morten et al, were conscientious in recording relevant details, so that this further feature of the analysis could be examined with respect to the graphic results shown in Figure 5. Children whose parents worked at the factory in jobs with lower levels of exposure to lead, and parents who practiced good hygiene despite having high exposure to lead via their jobs were among those who had children with lower blood lead levels. Although it is not

certain that control and exposed children did not differ in other ways, Rosenbaum uses a sensitivity analysis to show that hidden bias would have to be quite extreme to explain away differences this large. This example can be seen as a kind of propensity score analysis.

## 3. Some Further Points

Matching is virtually never extended beyond pairs in textbook treatments. Indeed, an informal survey found but one experimental design text that noted the possibility (Maxwell and Delaney, 1990). Yet the two dependent sample paradigm is easily extended to more complex situations. For example, suppose a matched sample procedure were used, but instead of forming pairs, triples were constructed in the context of comparing three treatments. Planned comparison contrasts may be used to generate two or more pairs of difference scores for the system of contrasts. One might use coefficients, say $c_1$ = [1,–1,0] to generate a difference score for the first in relation to the second treatment; then a second contrast, $c_2$ = [1/2, 1/2, –1] yields difference scores based on comparison of the average of the first two treatments, and the third. For each such contrast, an assessment plot has potential to provide visual evidence of treatment effects, going beyond standard summaries for planned comparison contrasts. While effect size computation, formal hypothesis tests, and confidence intervals are easily generated using one or more pooled variance estimates, graphical analysis in a search for patterns, trends and anomalies can also add value to any such comparison.

Extension to other more complex designs is also straightforward. For example, for matched quads, assignments to four treatments arranged in a 2×2 factorial may be used to generate data organized in four columns to correspond to four 'cells' in a factorial arrangement. Two main effect contrasts and one for interaction are generally easy to construct and analyze in such a case (an interesting point for those who teach that contrasts are for 'one-way' analysis of variance). Again, however, the potential of plots to show details of real data may be such that an emphasis on hypothesis tests, confidence intervals or even effect sizes will be seen as inappropriate or notably incomplete. If there are clusters, patterns or outliers (most easily exposed using graphs) then interpretation of conventional summary statistics will be incomplete, and possibly relatively unimportant in comparison to revised questions that follow from this analysis.

Reflecting on the shoe dataset in Figure 2 above, one is struck by the effectiveness of blocking in situations where blocks are especially homogeneous. Compared to the (naive) design that would assign sole material randomly to two groups of boys, the dependent sample paradigm yields enormous dividends, here seen as notable treatment differences within boys, while making differences between boys apparent too. This leads to a particular suggestion about how one may often be able to create blocks that are homogenous (with respect to the ultimate response measure). The idea is to measure and then *rank* units or entities at the outset of an experiment on some variable that, absent treatment effects, is likely to correlate highly with the ultimate response measure. Then, say for blocks of size two, randomly assign units to the two treatments within adjacent pairs after ranking. As indicated in Figure 2, and to a lesser extent in Figure 1, within-block differences may be made quite small using such a strategy. Extending this idea, if the analyst seeks to learn about treatment effects for separate subgroups of units, defined by other covariates (age, sex, demographics, prior experience, etc.), the ranking method could be executed within distinctive subgroups, taking advantage of the homogeneity feature for each one. Various details about treatment effects, and possible interactions defined either within or across subgroups, seem likely to be brought into sharp relief by an approach to design and analysis that capitalizes effectively on prior information. It seems curious that such strategies seem rarely to be tried in applied experimental research.

# 4. Discussion

Experience suggests that even statisticians are often not overly familiar with dependent sample analyses, so it follows that experience with corresponding graphic displays is likely to be even less common. The advent of modern software that facilitates more comprehensive analyses, and especially better graphical displays, offers possibilities that simply did not exist until recently. Given the dependent sample graphical method on which we have focused, and a few examples such as have been examined above, it seems that both statisticians and their students may benefit from use of comprehensive numerical and graphical analysis methods. Examples such as those seen above are aimed partly at improved understanding of methods, but even more importantly at effective applied research where the dependent sample paradigm may offer previously unforeseen benefits.

Given a display of dependent sample data, perhaps especially data from a true experiment where effective blocking has been used, the first thing to recognize is that individual block effects correspond to points that fall away from the identity diagonal. The strongest general evidence of treatment differences would, as in the case of Figure 2, correspond to seeing only minor variation among effect sizes across pairs or blocks. Such a finding leads to a relatively small denominator for the test statistic, and to a narrower confidence interval. When there are no outliers or irregularities among data points, the strongest possible generalizations about treatment effects are supported. But experiments often yield evidence of discernable, and perhaps distinctive, differences in experimental effects across blocks (pairs). When there are patterns, trends, clusters or outliers (most easily seen in plots) this can be viewed as evidence of interactions between treatments and the variables used to form blocks. Indeed, there is reason to believe that even most well-prepared and knowledgeable investigators who design highly efficient experiments may not be able to make accurate predictions about particular interactions.

The basic idea that has been emphasized is that of focusing on the details of what data have to say. That is, refraining from summarizing quickly and focusing on formal inference at the expense of ignoring particulars of data. The recently departed John W. Tukey spent much of his professional life discussing and demonstrating the value of graphs, plots, and visualization in data analysis, trying to ensure that numerical summaries, inferential statistics, and graphics would be used in service of understanding data rather than becoming ends in themselves. Numerous articles, chapters, and talks by John Tukey provide elaboration of this central point; see the several collected works of Tukey (e. g. Jones, ed. 1996). Yet much of the strongest evidence to be found that will reinforce this central message of Tukey can come from passing real data sets through software designed to expose details of dependent sample experimental data. It is unfortunately fairly rare, even for the most well-designed experimental studies, to see data so clean that standard summary statistics are wholly adequate for describing what real data have to say.

Furthermore, at least from the perspective of studying the dependent sample paradigm, students rarely see examples of sound and comprehensive real data analyses in their primary textbooks. Even authors of books on (experimental) design seem to concentrate on data-driven questions infrequently, nor is the use of data to refine or modify initial hypotheses or research questions comprehensively demonstrated. Authors almost inevitably focus on methods, qua methods, even when it would be a worthwhile 'digression' to focus on data.

Pedagogically, the dependent sample plot is a natural specialization of a scatterplot. This means the plot can be introduced quickly, with the proviso that the $X$ and $Y$ variables are now seen as commensurable; difference scores make sense only if measurements are on the same scale. The plot can be meaningfully discussed as soon as the students have been exposed to the idea of dependent samples. The plot naturally reinforces the admonition that investigators need to look at their data; and, particularly when used to examine data of the kind seen in Figure 1 and Figure 2, it can help the

instructor to make the case for blocking. Also, as indicated above, it can lead to useful discussions of data and the questions that motivated data collection.

In our courses, the analysis of dependent sample data and simple blocking is generally the last topic taught at the end of the first semester of introductory statistics. The methods part of the chapter can be reduced to one-sample techniques covered earlier in the semester, while the key concept of blocking extends ideas introduced in a previous discussion of experimental design. The dependent sample plot also tends to reinforce lessons taught during the earlier discussion of residual plots, viz., that patterns and trends can provide useful information. By plotting matched values as points, and difference scores as derived from these points, all information is visually related in the same display. In turn, deficiencies of conventional dependent sample analysis may also be exposed. A plot like this helps to bring the course full circle; it provides an excellent tool to stress interrelationships among descriptive and inferential aspects of statistics, as well as something of the deeper function of exploration. The role of design can also be highlighted. For more advanced students, the dependent sample plot can provide an introduction to two-way ANOVA, propensity score analysis, and even longitudinal data analysis. When connected to the ideas noted in the preceding section such plots can also facilitate an introduction to hierarchical methods and analyses.

---

## Appendix A: Another View of the Tobacco Leaf Data

Consider what happens if the question about experimental effects were modified in the case of the tobacco leaf data seen in Figure 1. In particular, suppose that rather than computing the difference between X and Y scores for each pair, that the ratio of $X$ to $Y$ scores were found. There is nothing magical about using algebraic differences to assess

experimental effects. For example, one might just as well ask whether ratios of the form $\frac{X}{Y}$ tend systematically to

differ from unity as evidence that one treatment is different than the other. A natural variant of this idea is to use logarithms of ratios, whence it is recalled that $\log(X/Y) = \log(X) - \log(Y)$. It follows that logarithms, and differences $D_{\log}$ between logarithms may be used as a basis for analysis.

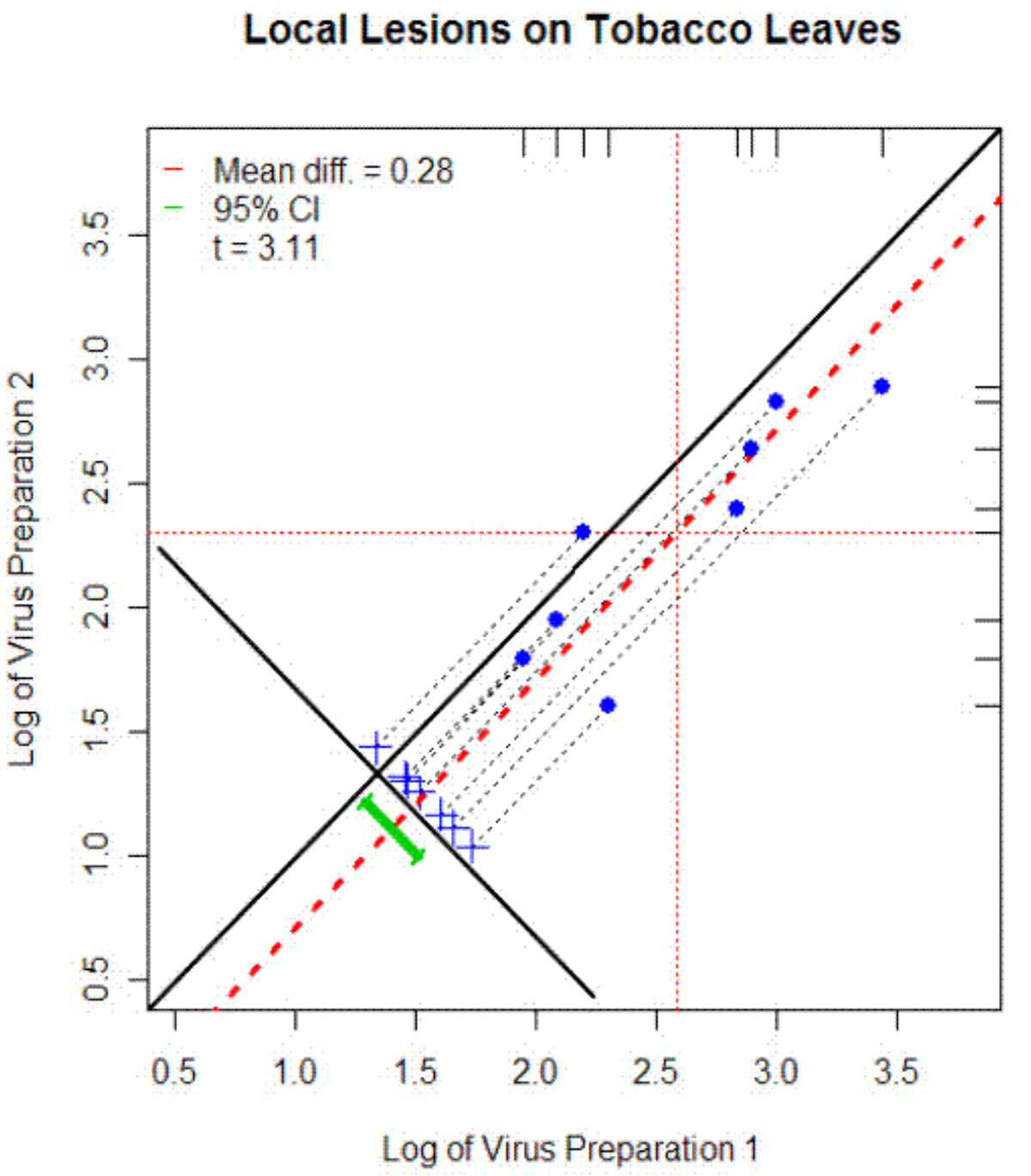


**Figure 6**: Two virus preparations were rubbed on different sides of tobacco leaves. The logarithms of the number of lesions per side for eight leaves are plotted. Youden and Beale

As seen in the legend in Figure 6, when the data are transformed by taking logarithms, the *t*-statistic increases to 3.11 (compared to 2.63 for the untransformed data), and the confidence interval is more clearly separated from zero. These more salient markers of effects are closely related to the finding that the correlation between the sums, or means, of *X*, *Y* values, and their differences has dropped to .18, substantially lower than the raw score counterpart of .77. Finally, the log-based plot itself seems more persuasive in showing simply that the *X*-virus produces more lesions than the *Y*-virus.

Taken together, these results strongly recommend the logarithmic reexpression of the data to show experimental effects; simple raw differences seem less adequate. It may be recalled that neither Youden and Beale, who initially collected these data, nor Snedecor and Cochran who presented them to illustrate the two dependent sample method,

considered transformation methods in the context of this analysis. The broader point of course is that there is no reason generally not to consider transformations or reexpressions in analyses; such steps can both strengthen conclusions and simplify interpretations. It has often been found that logarithmic transformations are helpful in analyses of counted data.

---

## Appendix B: Features to Note for Difference Score Assessment Plots

The R function used here, `granova.ds`, is freely available as part the package `granova` (Graphical ANOVA), for use with the comprehensive statistics software platform R. R is freely available from http://www.r-project.org. Note that R includes many accessories, including several pdf teaching and help files, and has superb graphics capabilities of many kinds. For more information please see the documentation of package `granova` in R.

1. The solid diagonal line has intercept zero, slope one. This is the identity line, as the line indicates $X = Y$. It follows that when differences $D = X–Y$ are below this line then $X$ is larger than $Y$, and vice versa.

2. Each ($X$,$Y$) point corresponds to a filled circle in the scatterplot; the marginal distribution of $X$ is given by the *rug* plot (ticks) along the top, similarly for the marginal distribution of $Y$ along the right side of the plot. The light (red) dotted vertical and horizontal lines correspond to the means of $X$ and $Y$.

3. The perpendicular distance between any point and the heavy black diagonal corresponds to a difference $D = X–Y$; however, the perpendicular Cartesian distance from any point and the main diagonal actually equals $\frac{D}{\sqrt{2}}$. This is because distance measured on the diagonal requires adjustment to correspond to that of the horizontal or vertical metric.

4. Each projection (parallel to the $45^{\circ}$ line, toward lower left) from a filled circle to the perpendicular at the lower left stops at a (blue) + such that the system of `crosses' depicts the marginal distribution of $n$ difference scores $D$; the heavy (red) dashed line, parallel to the identity line, depicts the mean of differences, $\bar{D}$. Note that $\bar{D}$ corresponds to the intersection of marginal means, i.e., to $\bar{X} - \bar{Y}$.

5. The heavy (green) line below the lower left line segment, perpendicular to the identity, depicts a 95% confidence interval for the population mean difference, $\bar{D}$.

6. The upper-left legend shows numerical values for the mean difference, $\bar{D}$, as well as the numerical value of the $t$-statistic for the null hypothesis of no population mean difference.

7. The R function `granova.ds` used in the production of these graphics also produces numerical output: the sample size *n*, the means of *X*, *Y*, and *D*, the standard deviation and effect size of *D*, the correlation between *X*, *Y* and between *X*+*Y*, *D*, the endpoints of the 95% confidence interval, and the *t*-statistic, degrees of freedom and *p*-value for the *t*-statistic.

---

# Appendix C: Datasets

The shoe wear and anorexia data, named *shoes* and *anorexia* respectively, are available in R in the MASS library. The *anorexia* data used here is a subset (family therapy) of the overall dataset.

| Tobacco | Prep 1 | Prep 2 |
|---|---|---|
| 1 | 31 | 18 |
| 2 | 20 | 17 |
| 3 | 18 | 14 |
| 4 | 17 | 11 |
| 5 | 9 | 10 |
| 6 | 8 | 7 |
| 7 | 10 | 5 |
| 8 | 7 | 6 |

Table 1: Two virus preparations were rubbed on different sides of tobacco leaves. The number of lesions per side for eight leaves are recorded. Youden and Beale

| Shoes | A | B |
|---|---|---|
| 1 | 13.2 | 14.0 |
| 2 | 8.2 | 8.8 |
| 3 | 10.9 | 11.2 |
| 4 | 14.3 | 14.2 |
| 5 | 10.7 | 11.8 |
| 6 | 6.6 | 6.4 |
| 7 | 9.5 | 9.8 |
| 8 | 10.8 | 11.3 |
| 9 | 8.8 | 9.3 |
| 10 | 13.3 | 13.6 |

Table 2: Different shoe sole materials for each of eight pairs of shoes test relative wear rates. Box, Hunter and Hunter.

| Anorexia | Wt. Prior to Therapy | Wt. Post Therapy |
|---|---|---|
| 1 | 83.8 | 95.2 |
| 2 | 83.3 | 94.3 |
| 3 | 86.0 | 91.5 |
| 4 | 82.5 | 91.9 |
| 5 | 86.7 | 100.3 |
| 6 | 79.6 | 76.7 |
| 7 | 76.9 | 76.8 |
| 8 | 94.2 | 101.6 |
| 9 | 73.4 | 94.9 |
| 10 | 80.5 | 75.2 |
| 11 | 81.6 | 77.8 |
| 12 | 82.1 | 95.5 |
| 13 | 77.6 | 90.7 |
| 14 | 83.5 | 92.5 |
| 15 | 89.9 | 93.8 |
| 16 | 86.0 | 91.7 |
| 17 | 87.3 | 98.0 |

Table 3: Effect of Family Therapy on weights of anorexic girls. Hand et al.

| β-Endorphine | 12 Hours Prior | 10 Min. Prior |
|---|---|---|
| 1 | 10.0 | 6.5 |
| 2 | 6.5 | 14.0 |
| 3 | 8.0 | 13.5 |
| 4 | 12.0 | 18.0 |
| 5 | 5.0 | 14.5 |
| 6 | 11.5 | 9.0 |
| 7 | 5.0 | 18.0 |
| 8 | 3.5 | 42.0 |
| 9 | 7.5 | 7.5 |
| 10 | 5.8 | 6.0 |
| 11 | 4.7 | 25.0 |
| 12 | 8.0 | 12.0 |
| 13 | 7.0 | 52.0 |
| 14 | 17.0 | 20.0 |
| 15 | 8.8 | 16.0 |
| 18 | 17.0 | 15.0 |
| 17 | 15.0 | 11.5 |
| 18 | 4.4 | 2.5 |
| 19 | 2.0 | 2.0 |

Table 4: Beta-endorphine levels in surgical patients 10 to 12 hours before surgery versus ten minutes before. Hoaglin, Mosteller, Tukey as found in Hand et al.

| Lead | Exposed | Control |
|---|---|---|
| 1 | 38 | 16 |
| 2 | 23 | 18 |
| 3 | 41 | 18 |
| 4 | 18 | 24 |
| 5 | 37 | 19 |
| 6 | 36 | 11 |
| 7 | 23 | 10 |
| 8 | 62 | 15 |
| 9 | 31 | 16 |
| 10 | 34 | 18 |
| 11 | 24 | 18 |
| 12 | 14 | 13 |
| 13 | 21 | 19 |
| 14 | 17 | 10 |
| 15 | 16 | 16 |
| 16 | 20 | 16 |
| 17 | 15 | 24 |
| 18 | 10 | 13 |
| 19 | 45 | 9 |
| 20 | 39 | 14 |
| 21 | 22 | 21 |
| 22 | 35 | 19 |
| 23 | 49 | 7 |
| 24 | 48 | 18 |
| 25 | 44 | 19 |
| 26 | 35 | 12 |
| 27 | 43 | 11 |
| 28 | 39 | 22 |
| 29 | 34 | 25 |
| 30 | 13 | 16 |
| 31 | 73 | 13 |
| 32 | 25 | 11 |
| 33 | 27 | 13 |

Table 5: Blood lead levels of lead workers' children matched with similar control children. Morton et al.

## Acknowledgments

The authors would like to express their appreciation to five individuals who provided helpful comments about earlier drafts of this paper, especially pedagogical aspects, or about the dependent sample graphic: Beth Ballert Potter, Jason Bryer, Thomas Knapp, Deepti Marathe and Paul Zachos.

## References

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

Hand, D. J., Daly, F., McConway, K., Lunn, D., Ostrowski, E., editors (1993). *A Handbook of Small Data Sets.* Number 232, 285. Chapman & Hall, New York.

Hoaglin, D. C., Mosteller, F., Tukey, J. W. (1985). *Exploring data tables, trends, shapes.* Wiley & Sons, New York.

Howell, D. C. (2002). *Statistical methods for psychology.* Duxbury, Pacific Grove, Ca., fifth edition.

Jones, L. V., editor (1996). *The collected works of John W. Tukey: philosophy and principles of data analysis: 1965-1986.* CRC Press, New York.

Maxwell, S. E., Delaney, H. D. (1990). *Designing Experiments and Analyzing Data.* Wadsworth Publishing Company.

Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M., Saah, M. (1982). Lead absorption in children of employees in a lead related industry. *American Journal of Epidemiology,* 115:549-555.

Rosenbaum, P. R. (1989). Exploratory plots for paired data. *American Statistician*, 43:108-110.

Rosenbaum, P. R. (2002). *Observational Studies.* Springer, New York, second edition.

Snedecor, W., Cochran, W. (1980). *Statistical methods.* Iowa State University Press, Ames Iowa, seventh edition.

Youden, W. J., Beale, H. P. (1934). A statistical study of the local lesion method for estimating tobacco mosaic virus. In *Contributions from Boyce Thompson Institute 6*, page 437.

Robert M. Pruzek
Department of Educational and Counseling Psychology
1400 Washington Ave.
State University of New York at Albany
Albany, N.Y. 12222
RMPruzek@yahoo.com

James E. Helmreich
Department of Mathematics
3399 North Rd.
Marist College
Poughkeepsie, NY 12601
James.Helmreich@Marist.edu

---