Causal inference in the context of a graphical approach to data analysis for experiments

R. M. Pruzek, SUNY Albany
Email: rmpruzek@yahoo.com

Department of Educational & Counseling Psychology, Division of Methodology

Abstract: Several scholars have focused in recent years on conceptions of causality and the issues involved when one aims to make inductive inferences about causal effects in applied science (*cf*. Pearl, 2000). Notwithstanding advances in understanding problems associated with causation, however, most specialists still agree that *when feasible, random assignment of units to treatments* provides the strongest basis for support of causal inferences. A particular approach to experimentation is outlined and recommended, one that entails use of prior information in (semi-random) assignments to treatments. Next, a *graphical approach to analysis* is illustrated and discussed in the context of analyzing five real data sets. The approach is particularly relevant to 'dependent sample comparisons.' Each data set to be presented has been published in a textbook, usually introductory. Illustrations will 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 of John Tukey's admonitions about data analysis will be seen to have special force and relevance.

Introduction

Consider four main types of (paired) dependent samples data:

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); and

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

2b. Comparisons of (two) matched individuals, perhaps for two selected treatments, where matching methods were used to form the subsets (pairs).

Category 2a is clearly most important in terms of facilitating efficient, constructive & informative *causal analyses*, notably when randomization has been used in the assignment of units within pairs to treatment groups. Randomization within blocks underpins experimental comparisons and can minimize doubts about *selection bias*. It is chiefly selection bias that tends to confound interpretations of between-group comparisons for observational data. Category 2b is important vis-à-vis analysis of observational data, particularly in applications of propensity score methods; and matching can mitigate selection bias too.

Extension beyond pairs to triplets, quads, etc., is generally straightforward, although this idea is not routinely taught. Ways to elaborate or extend graphics for dependent sample displays shown in the following slides will be noted.

It would not be unreasonable to describe 'idealized' dependent sample comparisons (*cf.* 2a above) as 'gold standards,' as among the generally best kinds of models one could choose for experiments. This is because in principle these designs, or paradigms, can often lead to highly efficient and scientifically informative studies with 'near-optimal' statistical properties. Furthermore, these designs are highly versatile, and can account for many real-world complexities (and lead to interaction discovery), in ways that are not widely taught nor well-understood.

Idealized *experiments based on blocking* entail use of the *most relevant prior information that is available to construct homogeneous blocks of units*. Absent experimental effects, *responses* of the (two) units or individuals within each block on the ultimate outcome measure *should be notably similar to one another*. Each difference between (post-treatment) responses within a block may be taken as evidence of an experimental effect when treatment assignments were random. When effect estimates vary little across blocks it may be reasonable to make strong and generalizable statements of experimental effects even when sample sizes are quite small…as two examples below will illustrate. Because each block yields a separate estimate of the experimental effect each comparison corresponds to an independent replication of the experiment. When such effects are especially similar across blocks then conclusions are especially simple, and generalizability may be warranted.

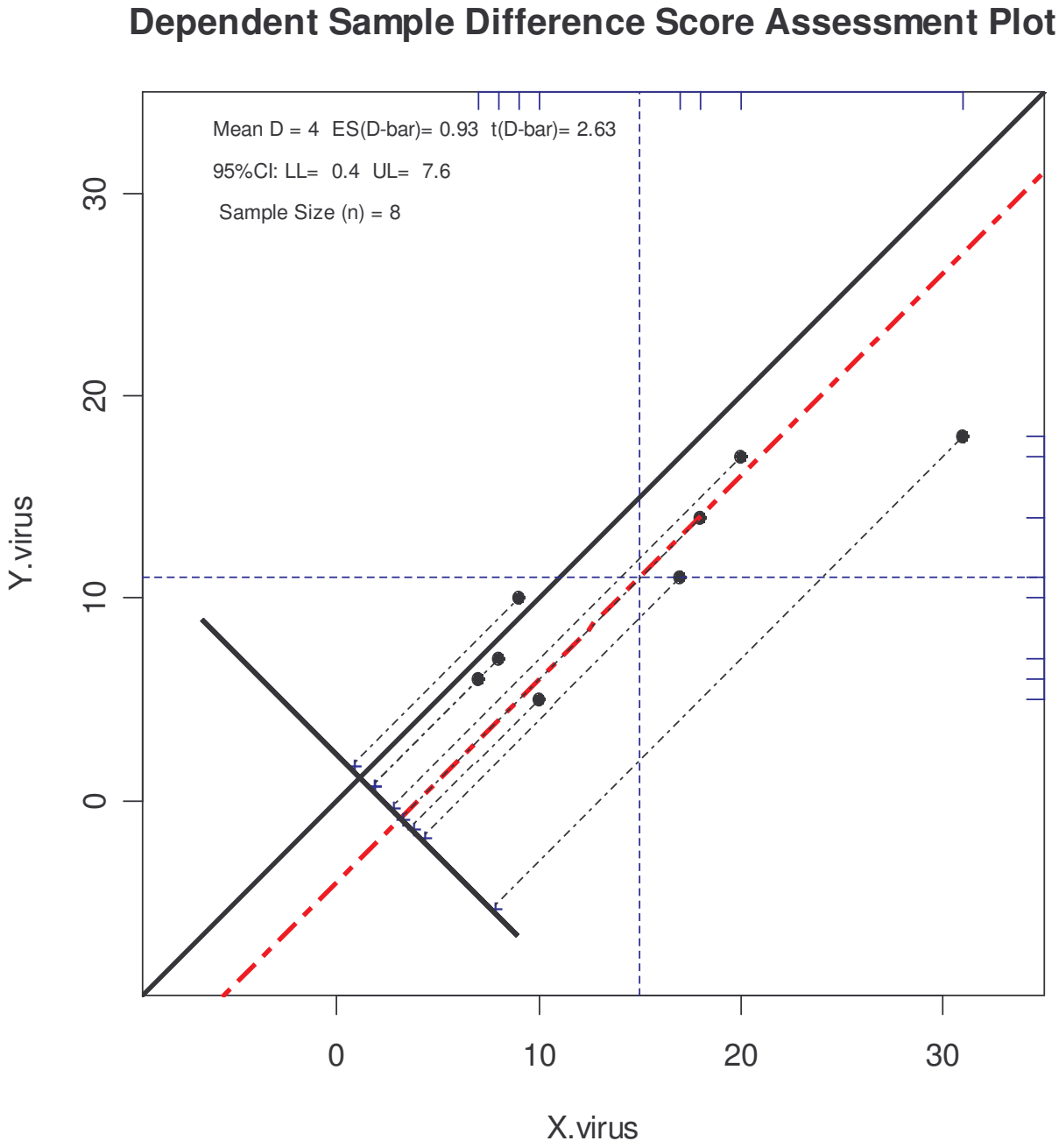
Examples that correspond to categories 1b and 2b will also be provided, and discussed, with special emphasis on the value of graphical presentations.

Some Examples of Dependent Sample data

The following slides illustrate use of a particular graphical method, here called a *Dependent Sample Difference Score Assessment Plot**, to study dependent sample data. The first two examples use real data from *true experiments*; these are the kinds of studies that can yield the strongest scientific inferences, especially when certain easily checked conditions are met. The next pair of slides 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 slide exhibits use of matching to aid the analysis of observational data.

It has been surprising to learn that in these cases, and many other *real data* examples that have been examined graphically, there appear to be interesting patterns, trends, irregularities, etc., that call into question the appropriateness of 'standard' approaches to analysis, where the *mean difference* is either tested for significance, or a confidence interval is generated. The key point 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 insure that interpretation(s) are both as clear and comprehensive as possible.

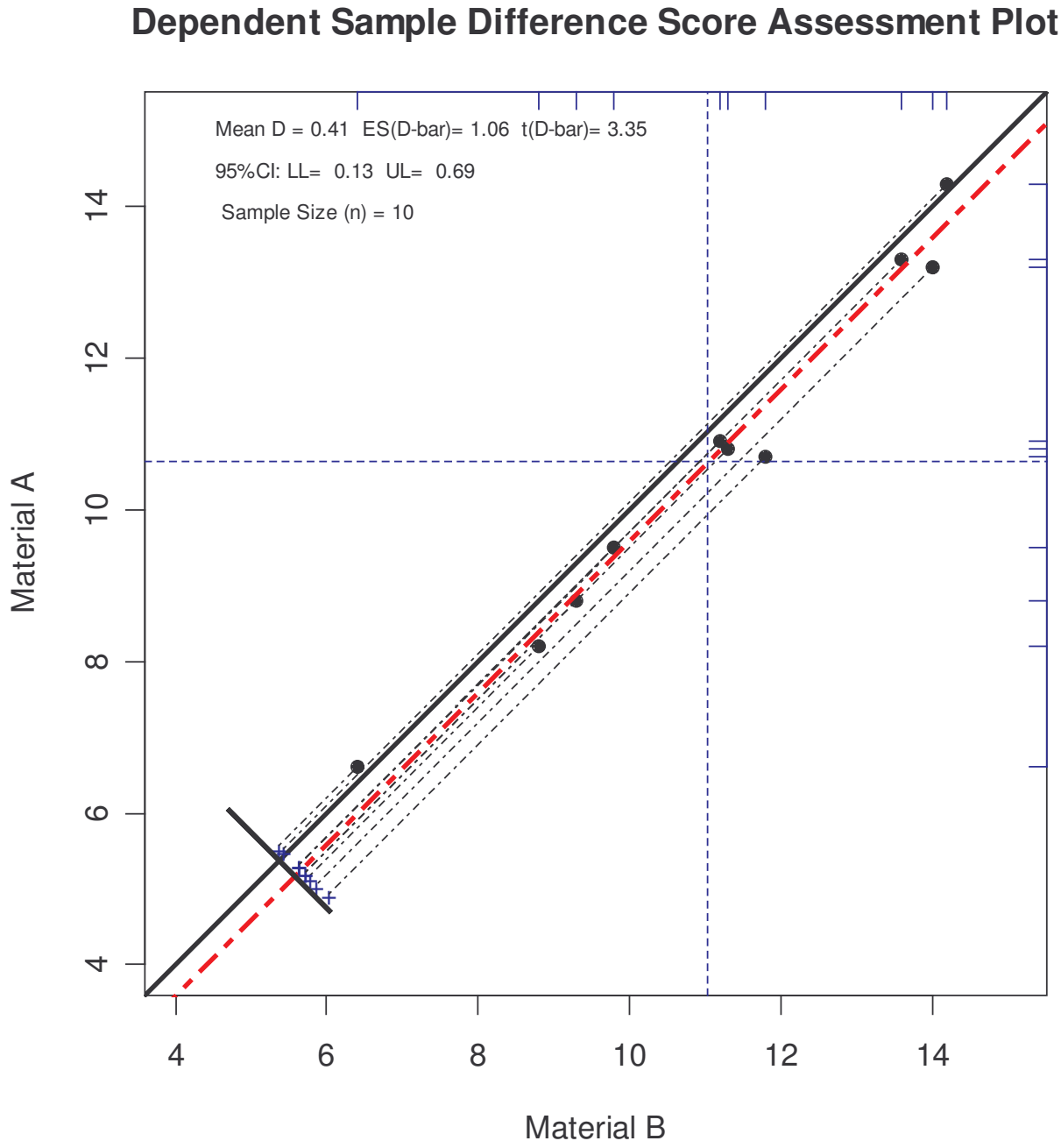
*For a similar graphical method, see Rosenbaum (1989).


Dependent Sample Difference Score Assessment Plot
Mean D = 4 ES(D-bar)= 0.93 t(D-bar)= 2.63
95%CI: LL= 0.4 UL= 7.6
Sample Size (n) = 8
Y.virus
X.virus
0
10
20
30

The data shown above 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 (Snedecor and Cochran, p. 86)." Each point in the figure corresponds to the *numbers of lesions* on the two halves of leaves that had been treated differently. Again, graphical display of experimental data* shows that the standard statistical test yields a statistically 'significant' result, despite the very small sample size; the correlation between X and Y scores is .90.

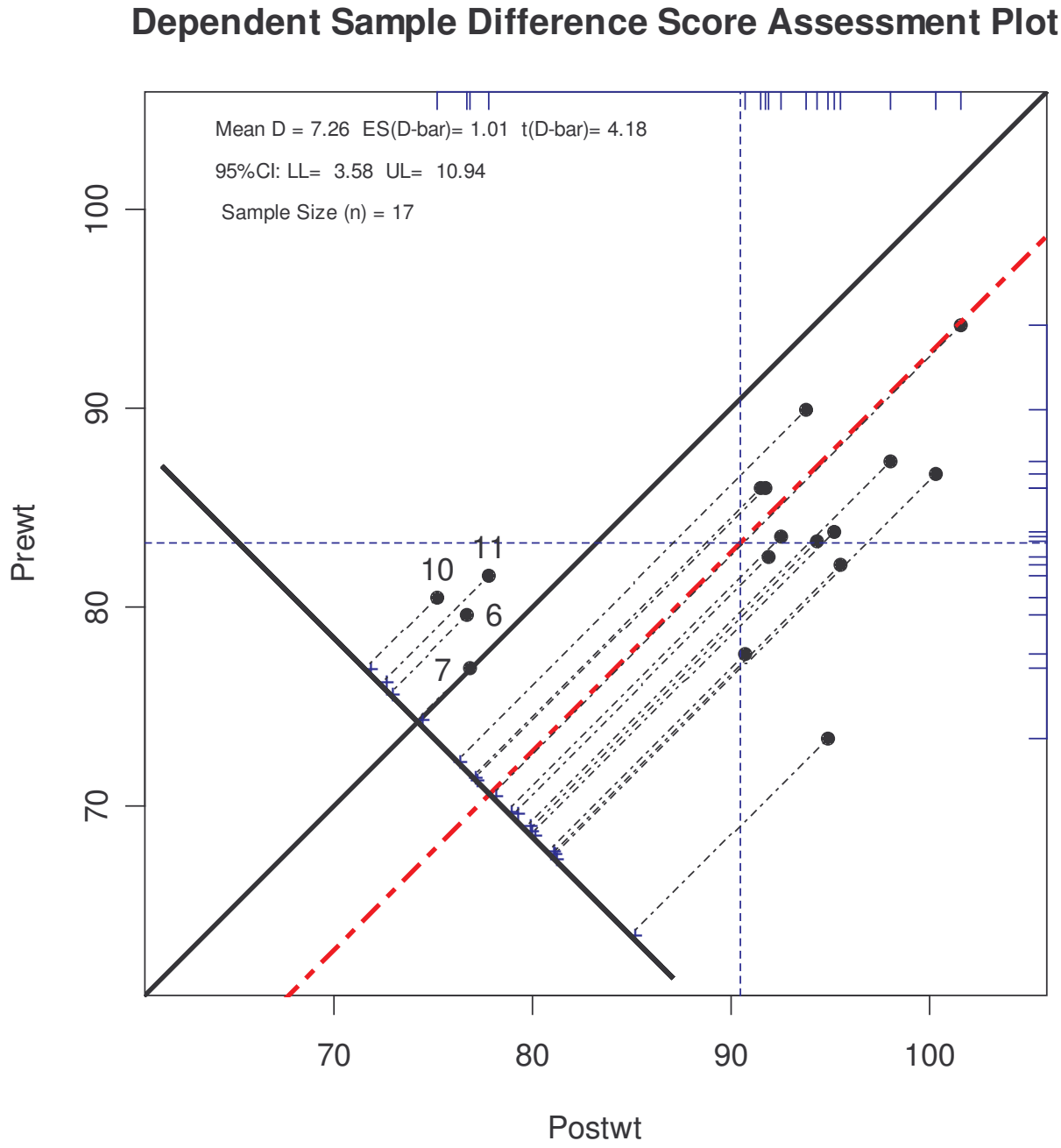
Note that a standard analysis focuses ***only*** *on numerical results and summaries*, and a simple test of significance ignores the trend that can be discerned in the plot. In particular, the plot shows a tendency for the departures of points from the heavy diagonal line to become larger as the mean (X,Y) values become larger, with a correlation of +.77. 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'. Although this is a small data set, this analysis shows that there is 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 that this is the best conclusion, however, see Appendix B.

*See Appendix B to see the relevance of score transformations in this context .


Dependent Sample Difference Score Assessment Plot
Mean D = 0.41 ES(D-bar)= 1.06 t(D-bar)= 3.35
95%CI: LL= 0.13 UL= 0.69
Sample Size (n) = 10
Material A
Material B
4
6
8
10
12
14

The figure above displays the well-known data set on shoe wear initially given by Box, Hunter and Hunter (1978). One sole made of material 'X,' and another made of material 'Y,' were *randomly assigned* to the shoes of ten active boys. The X,Y scores are measures of wear for the two materials. This is a 'true experiment' 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 Y-material wore longer than the X-material. Indeed, 'statistical significance' is noted (see legend) despite the small sample size, this being a consequence of the small variation in the D's. The near uniformity of effects is also largely responsible for an effect size whose magnitude exceeds unity.

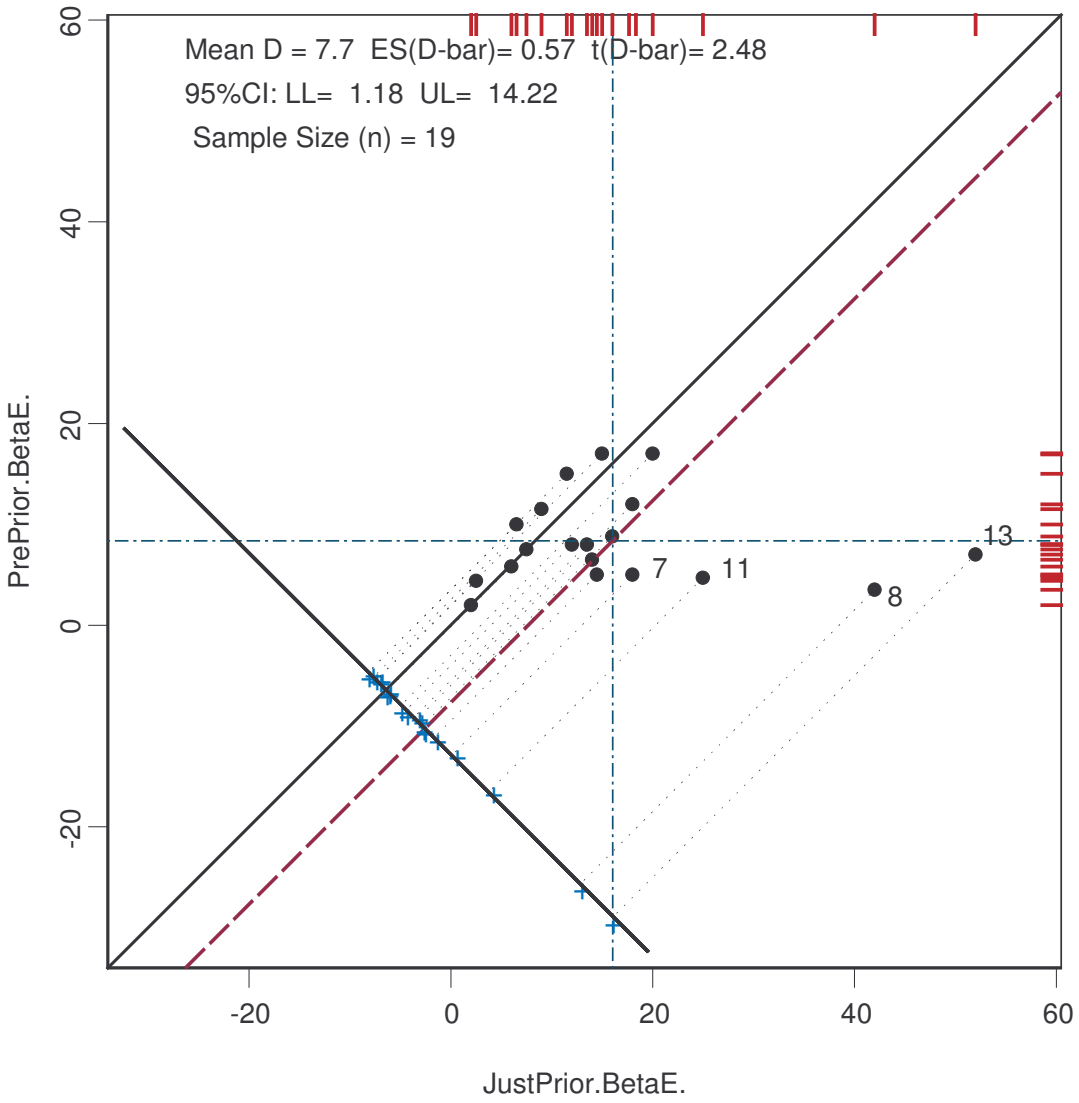
Although the high correlation (.99) between X and Y scores is related to the relatively small variance of the D's, note that a high correlation alone is not sufficient, since as the preceding example showed, the major ellipse associated with the X,Y point swarm may not be a line with slope near to unity. Because random assignments had been used with matched pairs, and the shoe data results are so clear, this kind of design can be see as a *gold standard exemplar* of a true experiment; blocking was highly effective in reducing variation of the D's, and response variable metric is also well chosen.


Dependent Sample Difference Score Assessment Plot
Mean D = 7.26 ES(D-bar)= 1.01 t(D-bar)= 4.18
95%CI: LL= 3.58 UL= 10.94
Sample Size (n) = 17
Prewt
Postwt
70
80
90
100
10
11
6
7

The data set shown above consists of weights in pounds for $n = 17$ girls who were weighed before and after treatment for anorexia. These data were originally published by Hand, *et al*, 1993, and were reprinted in Howell (2001). X scores are weights (in lbs.) after family therapy; Y scores are corresponding weights before therapy. A difference score, D, 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 lbs and the corresponding t-statistic is 4.18. Even the standardized effect size, 1.01, is notable.

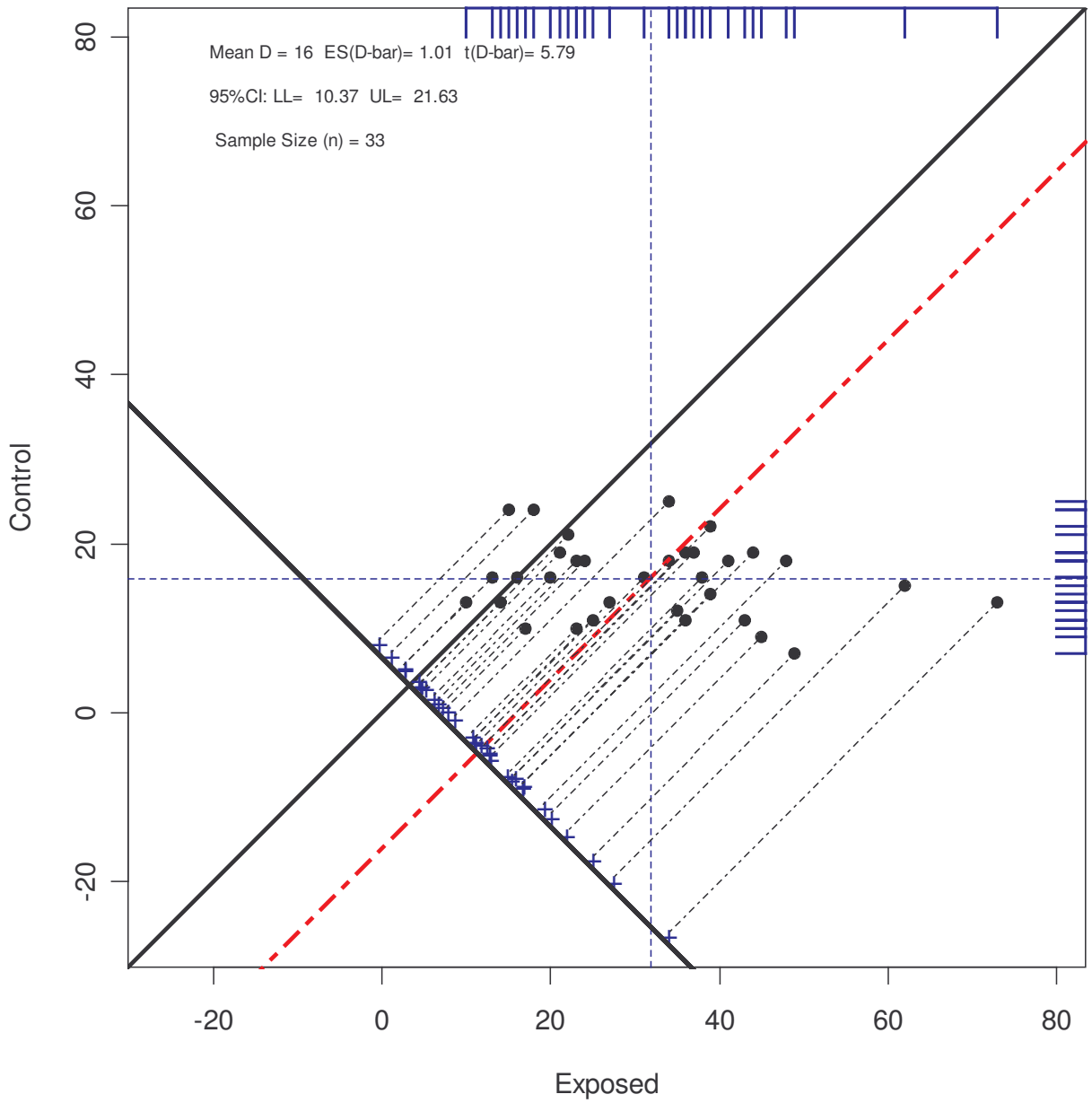
The plot, however, tells a more nuanced story. The *cluster of points* at the extreme left show that these four girls (those w/ *id's* 6,7,10, 11) actually *lost* weight. Indeed, the remaining 13 girls had an average weight gain of 10.4 lbs, and for them the *standardized effect size* was an impressive 2.26 (*t* moves up to 8.22, although post-hoc data selection undermines probabilistic interpretation of this statistic – more on this later). 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.*

Dependent Sample Difference Score Assessment Plot



The preceding figure appears to yield new insights for analysis of time-related data taken from Howell (2002, *pps.* 218-19) concerning blood levels of beta-endorphin for 19 patients prior to surgery. The Y scores show beta-endorphin levels 12 hours before surgery, Xs show levels 10 minutes before surgery; beta-endorphin scores are intended to measure stress. 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 (ids: 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 D's are removed, the remaining data depict a much lower stress effect, non-significant (α =.05), with a lower standardized effect size of .45. In fact, five of these 15 persons 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 (-.06), so that use of repeated measures did not reduce variance of D's in this situation. Some form of 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 the simplistic conclusion that 'surgery significantly increases stress (as evidenced by elevated beta-endorphin levels).'


Mean D = 16 ES(D-bar)= 1.01 t(D-bar)= 5.79
95%CI: LL= 10.37 UL= 21.63
Sample Size (n) = 33
Control
Exposed
-20
0
20
40
60
80

Data shown in the preceding slide are based on an observational study by Morten, *et. al* (1982, *Amer. Jour. Epidemiology*, p. 549 ff). Children of parents who had worked in a factory where lead was used in making batteries were matched by *age* and *neighborhood* with children whose parents *did not* work in lead-related industries. Whole blood was assessed for lead content yielding measurments in MicroLiters/dl; results shown compare the Exposed w/ Control Children. Conventional dependent sample analysis shows that the Effect Size was about 1 standard deviation unit, the (95%) C.I. is far from zero, and the t-statistic for the mean of the difference scores was 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.

Examination of the graphic shows more, however. Note the *wide dispersion of lead measurements* for Exposed children in comparison with their Control counterparts. One interpretation of this result is that it is the *particular characteristics* of parents' experiences at work or home with their children that 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.

Although it is not certain that Control & Exposed children did not differ in other ways (than age and neighborhood of residence), Rosenbaum (2002) uses a sensitivity analysis to show that the hidden bias would have to be quite extreme to explain away differences this large. This example can be seen as related to propensity score analysis.

Some further points

Although the concept of matching is virtually never extended beyond pairs in textbook treatments (no example could be found in any introductory text; and seemingly only one experimental design text notes this possibility), the *two dependent sample paradigm* is easily extended to deal with more complex situations.

For example, suppose a matched sample procedure were used, but instead of forming pairs, triples or quads were constructed in the context of comparing three or four treatments. In such cases, *planned comparison contrasts* may be used to generate *two or more* ***pairs of difference scores*** *for the system of contrasts*. In the case of a three group experiment one might use coefficients, say $\mathbf{c}_1 = [1, -1, 0]$ to generate a difference score for the first in relation to the second treatment; then a second contrast, $\mathbf{c}_2 = [\ ½, ½\ , -1]$, yields difference scores based on comparison of the *average* of the first two treatments, and the third. For each contrast, an Assessment Plot has potential to provide visual evidence of treatment effects, going beyond standard summaries for planned comparison contrasts. Effect size computation, formal hypothesis tests, and confidence intervals are easily generated using one or more pooled variance estimates, borrowing strength as it were, in order to obtain greatest efficiency for inferential applications as taught by Fisher as early as 1935; but a search for patterns, trends and anomalies can also accompany 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 x 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, which is 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. Note that 'robustification' may be beside the point also, since it usually entails heavy emphasis on data summary.

Still, 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 to the investigator who collected the data than revised questions that follow from careful analysis of sample data.

It seems essential to plot one's data, and that generally requires software to facilitate effective visualization. The [Splus or R] function used here is available from rmpruzek@yahoo.com upon request; more importantly, the entire **R** package is free. You may download the comprehensive statistics software package called **R**, that is 'not unlike Splus,' 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.

Discussion

Given departures of points from the identity diagonal, the analyst usually seeks evidence of minor variation among effects across pairs, or blocks, because the strongest generalizations about experimental effects are supported when there are differences depart from zero and are similar to one another. But experiments often yield evidence of discernable, and perhaps distinctive, differences in experimental effects across blocks (pairs). When there are patterns, trends, clusters and outliers (most easily seen in plots) this can be seen 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 details of what data have to say, not summarizing too quickly, trying to avoid the temptation to focus 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 insure that numerical summaries, inferential statistics, and graphics would be used in service of understanding data rather than becoming ends in themselves. How can we do more to insure that his teaching is not forgotten? Numerous articles, chapters, and talks by John Tukey provide elaboration of this central point; see one of the several Collected works of John Tukey (e.g., Jones, 1996), or for a psychology-related reference, Tukey (1969).

However much I have read John Tukey's writing over the years, much of the strongest evidence I have found to reinforce a central message of John Tukey has come from passing several real data sets through software designed to expose details of dependent sample experimental data. Indeed, with the help of my research assistant*, I find it is fairly rare, even for the most well-designed experimental studies, to see data so clean that standard summary statistics are 'wholly adequate' for real data analysis.

I would further argue, *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 rarely to concentrate on data-driven questions, nor to use data to refine or modify initial hypotheses or research questions. Authors almost inevitably focus on ***methods, qua methods***, even when it would be a very worthwhile 'digression' to focus on data!

Indeed, there seems to be a general deficiency in introductory statistics books as related to teaching of many central ideas of what we might call Tukey-based data analysis. In our review of over 20 introductory statistics books, only a few authors were found to provide more than perfunctory discussion of the notion that research questions must often be modified, refined, or elaborated in the light of data. And none provided X,Y plots for dependent sample data, without which patterns, trends and clusters in data g*enerally will not be found.*

*Thanks to Katerina Passa for her fine help.

References

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

Conover, W.J. & Iman, R.L. (1981) Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician*, *35*, *1*, 124-132.

Howell, D. C.(2002) *Statistical methods for psychology.* (5th Ed.) Pacific Grove, Ca.: Duxbury.

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press. 2000.

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

Snedecor, W. & Cochran, W. (1980) *Statistical methods* (7th Edition) Ames Iowa: Iowa State University Press.

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

Tukey, J. W. (1969) Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83–91.

Appendix A

Features to note for *Difference Score Assessment Plots**:

1. Solid diagonal line has intercept zero, slope one. Since 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; the marginal distribution of X is given by the 'rug' plot (blue ticks) along the top, similarly for marginal distribution of Y along right side of plot. Dashed vertical and horizontal lines correspond to $\bar{X}$ and $\bar{Y}$.

3. The perpendicular distance between any point and heavy black diagonal *corresponds to* a difference D = X – Y; however, the perpendicular Cartesian distance from any point and the main diagonal actually equals D/sqrt(2). This is because distance measured on the diagonal must be adjusted to correspond to that of horizontal or vertical metric.

4. Each projection (parallel to the 45° line, toward lower left) from a diamond to the perpendicular at the lower left stops first at a (blue) '**+**' such that the system of (jittered) **+**'s depicts the marginal distribution of *n* D's; the heavy (red) dashed line depicts $\bar{D}$. Note that $\bar{D}$ corresponds to intersection of marginal means, *i.e.*, to $\bar{X}$, $\bar{Y}$.

5# The second marginal distribution that derives from the D's (the red points, at extreme lower left) is a uniform distribution whose mean is the same as that of the D's, and whose standard deviation is defined so that the *t*-statistic for the uniform (ranked) counterpart of the D distribution yields a one-sample t-statistic that is the same as the parametric *t* (*cf*. Conover and Iman, 1981). (An optimization method is used to accomplish the latter task.)

6. The upper-left legend shows the numerical value of $\bar{D}$, as well as standardized effect size (ES) (computed as $\bar{D} / s(D)$ where s(D) denotes standard deviation of D's); also, first line of this legend shows *t*-statistic associated with ES.

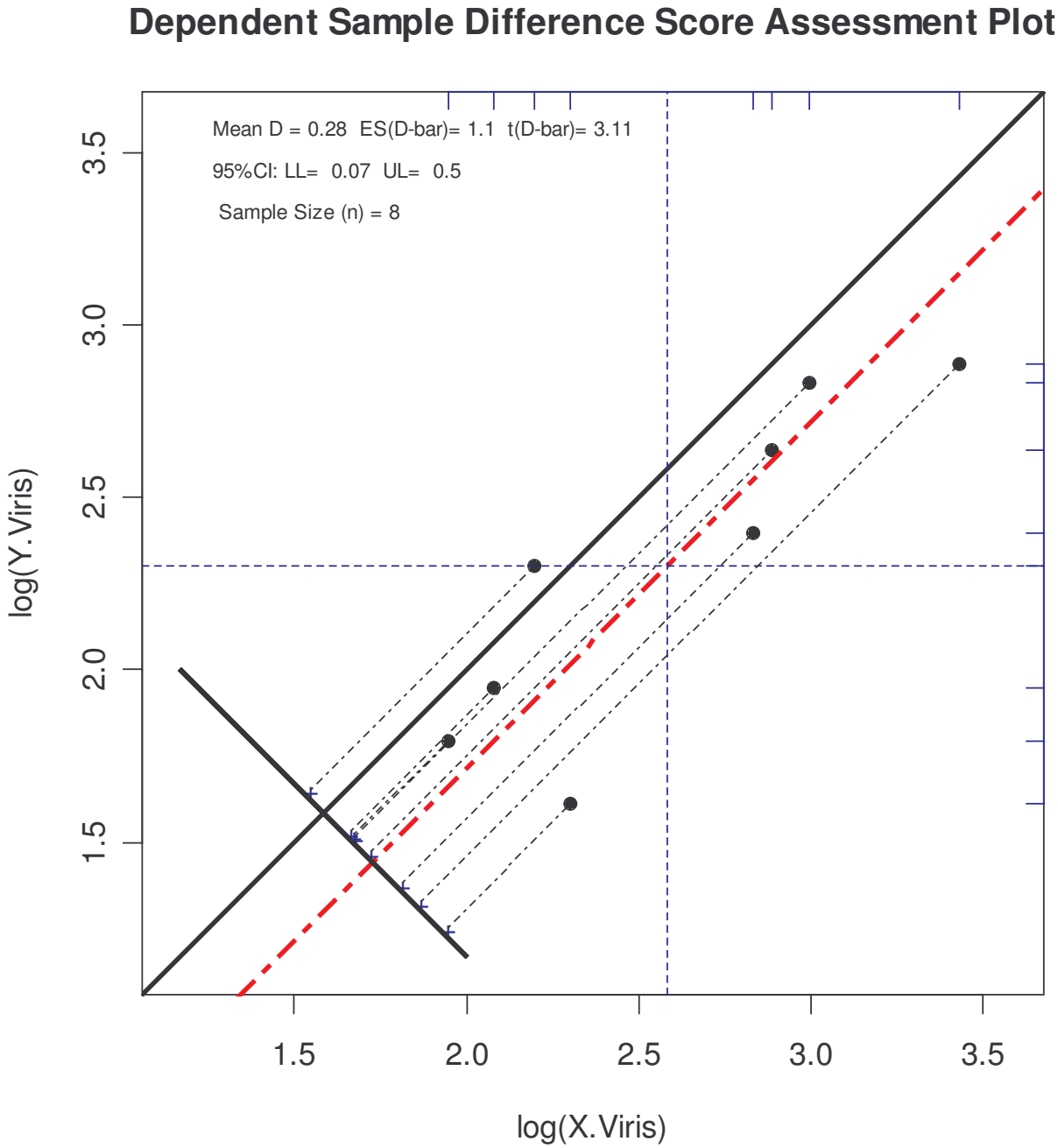
7. The second line of upper-left legend gives 95% confidence limits for the mean population difference. Sample size (*n*), *i.e.,* the number of points, is also printed in the legend.

8. When *n* exceeds 49, the Wilcoxon Z statistic (based on ranks of D's) is printed. (Note that when Z's magnitude exceeds that of *t* then 'outlier' D's are likely to be evident in plot.)

* For a similar plot, see Paul Rosenbaum (1989). #The second marginal is not implemented in this presentation.

Appendix B: 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, that is for the first of the preceding examples? In particular, suppose that instead of computing, for each pair, the *difference* between X and Y scores, that the ***ratio*** of X to Y were computed. 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 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 ***logs of ratios***, whence it is recalled that log(X/Y) = log(X) – log(Y). It follows that logs, and differences between logs, new Ds, may be used as a basis for analysis; this will now be demonstrated for these data.


Dependent Sample Difference Score Assessment Plot
Mean D = 0.28 ES(D-bar)= 1.1 t(D-bar)= 3.11
95%CI: LL= 0.07 UL= 0.5
Sample Size (n) = 8
log(Y.Viris)
log(X.Viris)
1.5
2.0
2.5
3.0
3.5

As seen in the legend in this final figure, when logs are used in the analysis, the ES rises to 1.1, compared to the previous value of .93; further, the *t*-statistic increases to 3.11 (compared to 2.63 before), 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 its 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 a particular transformation (or reexpression), *viz.,* logs, 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 log transforms are helpful in analyses of 'counted' data.