A paradigm to support causal inferences when resources are limited.

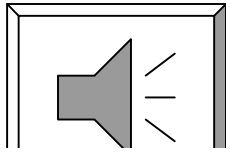
R.M. Pruzek

Introduction: Suppose you want to do a study that will be seen by persons you respect as most interesting and full of promise for further work. You would be hard-pressed to do better in many situations than to perform a well-designed ***true experiment***, especially if the results turn out to be 'significant.' True experiments are usually characterized as studies in which persons (or entities) are randomly assigned to treatments at the outset of study; responses are recorded and compared following the treatments. Unfortunately, it is not often taught that simple random assignment to treatments, while helpful, is generally inferior, often hugely so, to another approach that also entails random assignment, but not the most simple kind. The first several slides below show examples of the alternative paradigm. Those that follow show variations, extensions and improvements.

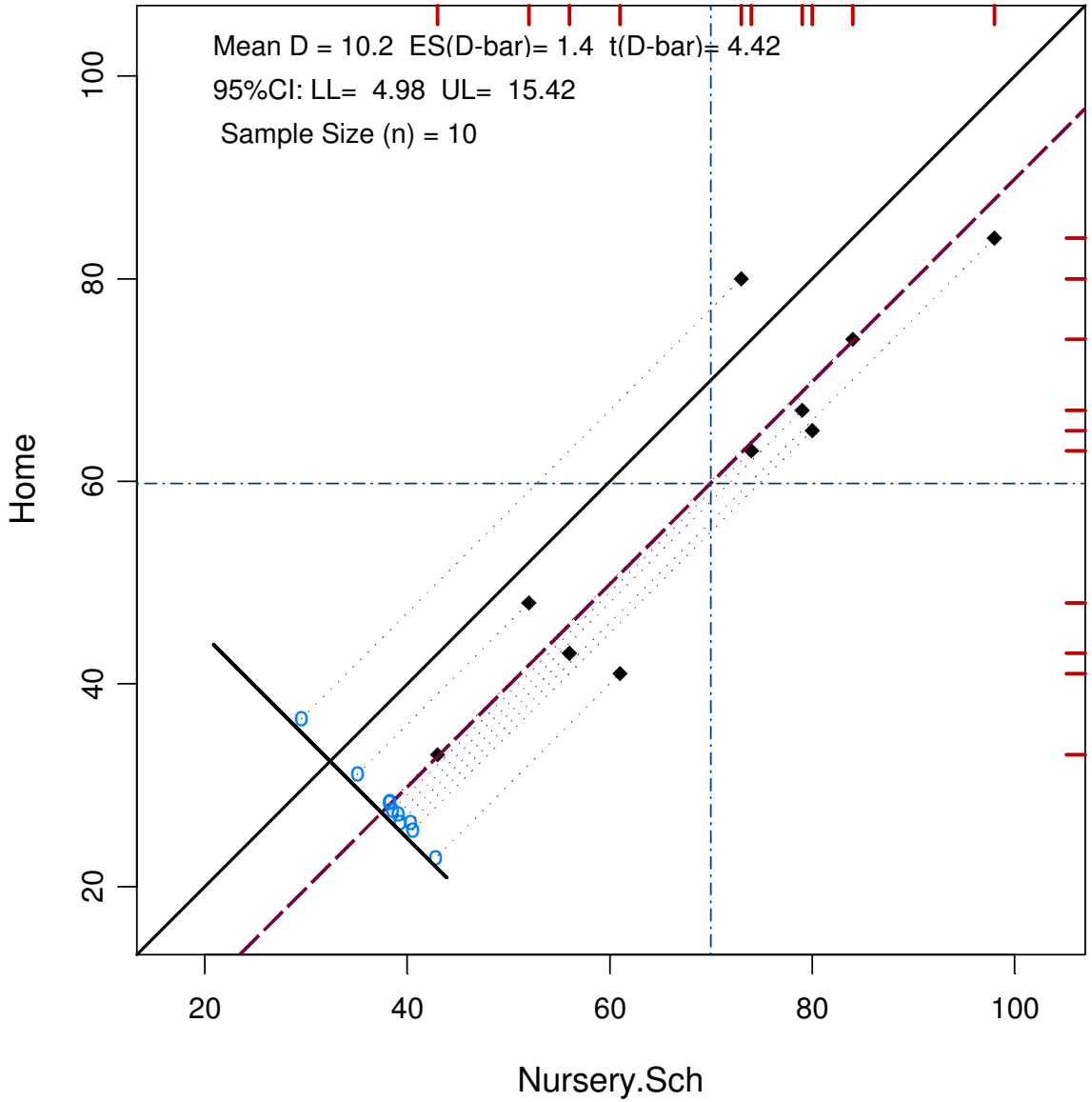
Some preliminaries:

For the following pairs of slides, I first show ‘raw’ data in the form of a modern plot, and then present corresponding narrative. The graphics are annotated; they should be studied carefully to discern details of what data have to say. The central idea is to use specialized plots to present certain kinds of data so as to provide more information than is usually given for these kinds of data. An Appendix containing technical details is provided at the end.

The second slide in each pair contains a brief narrative to describe what the data seem to show. In the early slides primary attention is given to more or less ‘conventional’ statistical findings; in latter slides, where the data become somewhat more complex, the story becomes less conventional, but more realistic. The ultimate aim is to discuss and account for issues that often arise in analyses of real data. That all data below are ‘real’ is worthy of special note.



Social awareness data, two environments for 10 identical twin pairs

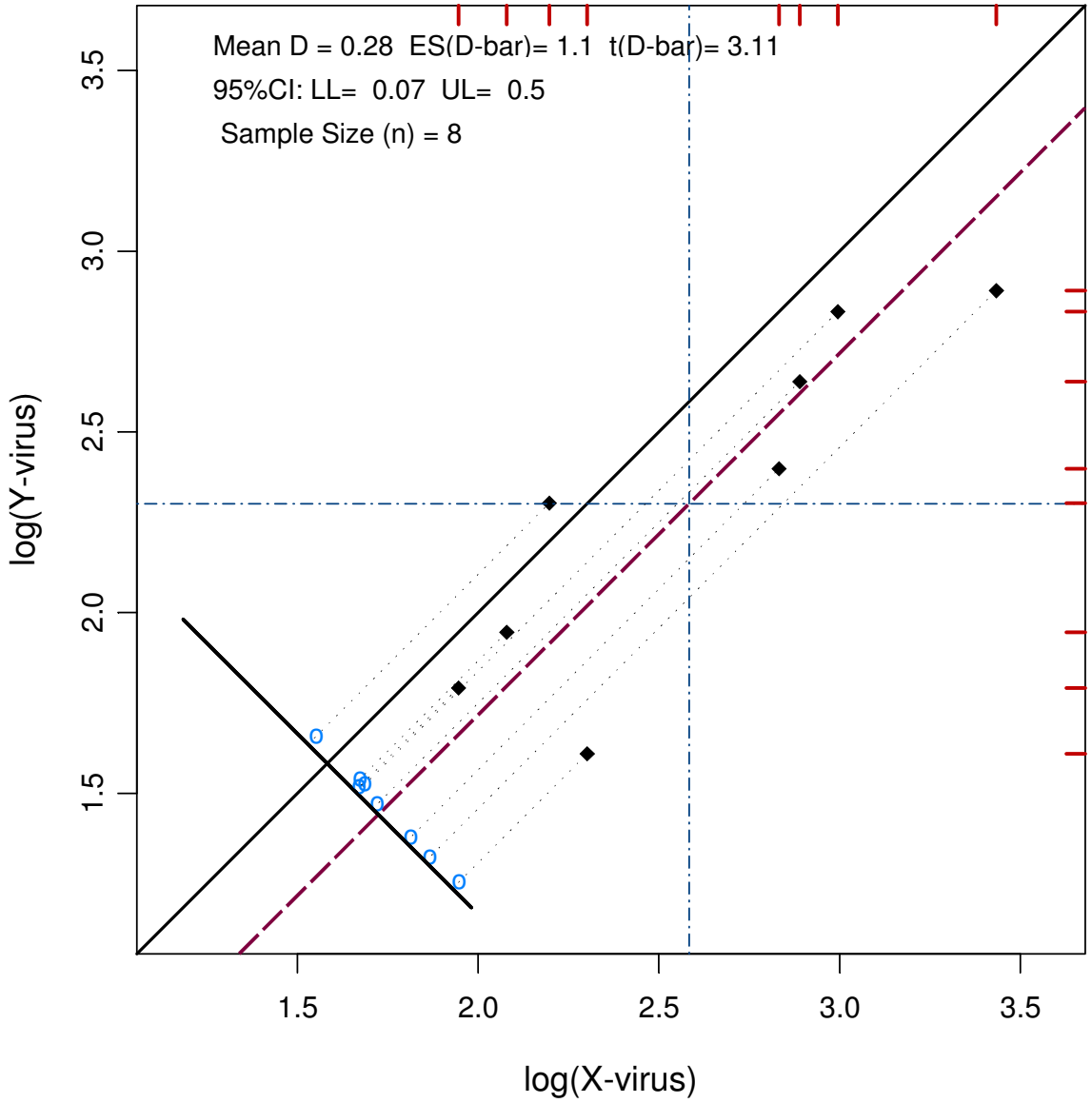


The figure above shows data for ten pairs of identical twins, age four (Siegel, 1956), randomly selected for an experiment to investigate how nursery school affects *social awareness* at this age. For each pair, one twin was randomly assigned to nursery school while the other stayed home. At the end of the time period, all 20 children took the same test and their scores were recorded.

As can be seen from the legend, a formal parametric test shows significance ($t = 4.40$) with a standard alpha; of course, the 95% confidence interval does not span zero. The small variance of the D scores (and $r_{xy} = .91$) is responsible for the large effect (ES $=1.4 = \bar{D} / S_D]$) despite the small sample size.

However, the graphic shows that for one pair the social awareness score was *lower* for the twin who attended nursery school than for the one who stayed home. Although more details would be good to have for all twin pairs, it might be most interesting to learn more about the 'outlier' twin pair.

Youden & Beale tobacco leaf virus data, from a true experiment

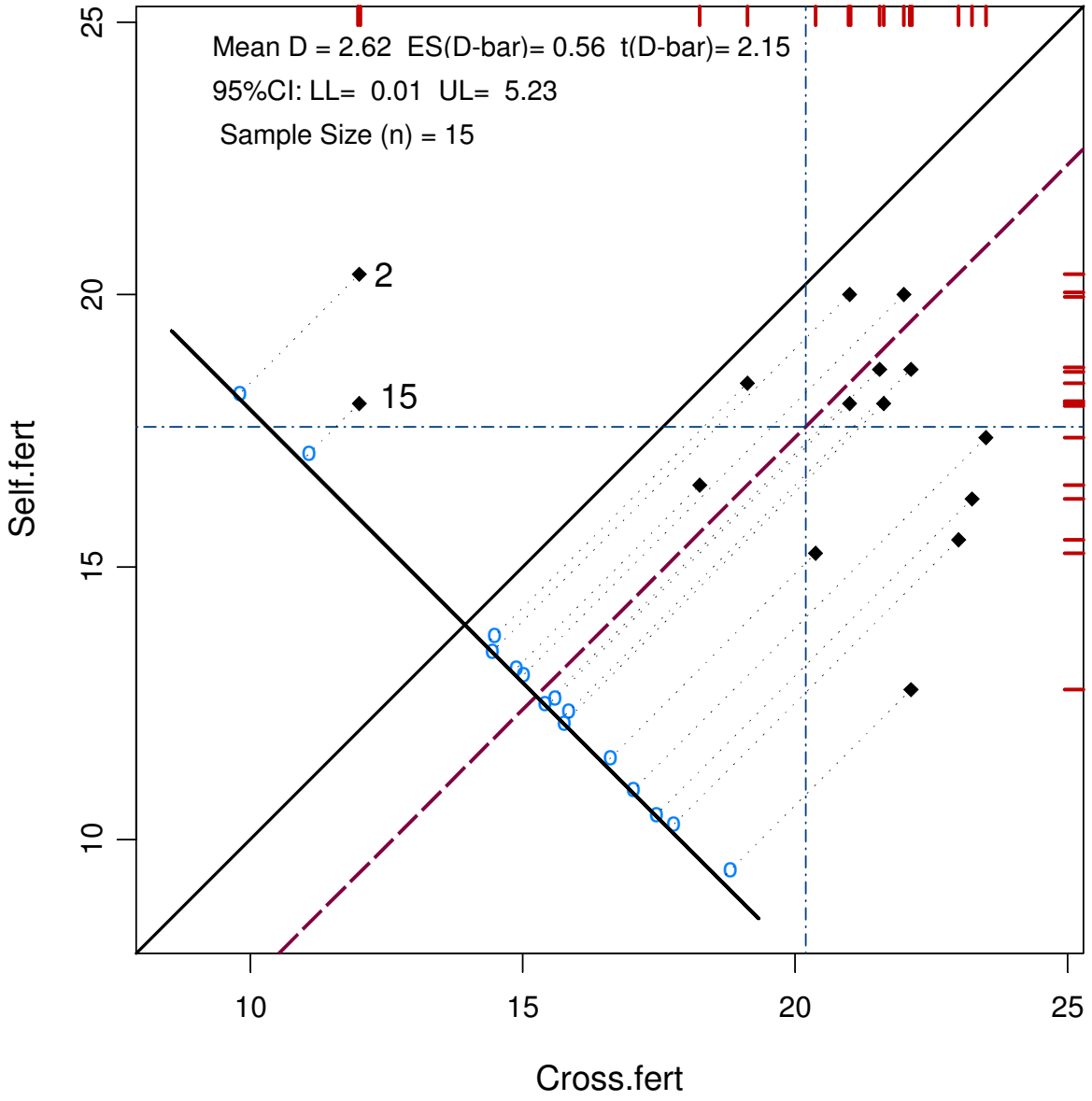


The data above were taken from Snedecor & Cochran (1980) and correspond to a *true matched pairs experiment*. The data came originally from Youden & Beale in 1934 who "wished to find out if two preparations of a virus would produce different effects on tobacco plants. Half a leaf of a tobacco plant was rubbed with cheesecloth soaked in one preparation of virus extract, and the second half was rubbed … with the second extract (Snedecor and Cochran, p. 86)."

Each point in the figure corresponds to the log* of *numbers of lesions* on the two halves of leaves that had been treated differently. The preceding graphical display of these experimental data shows that there was variation across tobacco leaves, and the test statistic, $t = 3.11$, can be interpreted straightforwardly. This is (strong) evidence to suggest that the 'X-virus effect' is stronger than that of Y, and this should generalize to a 'hypothetical population.' That the test result is 'significant' despite the very small sample size is again a consequence of the small variation in difference scores (D = X - Y), and is also related to finding a high correlation ($r_{xy} = .87$) between X & Y scores.

**Unfortunately, logarithms were not used previously; will discuss if time permits.*

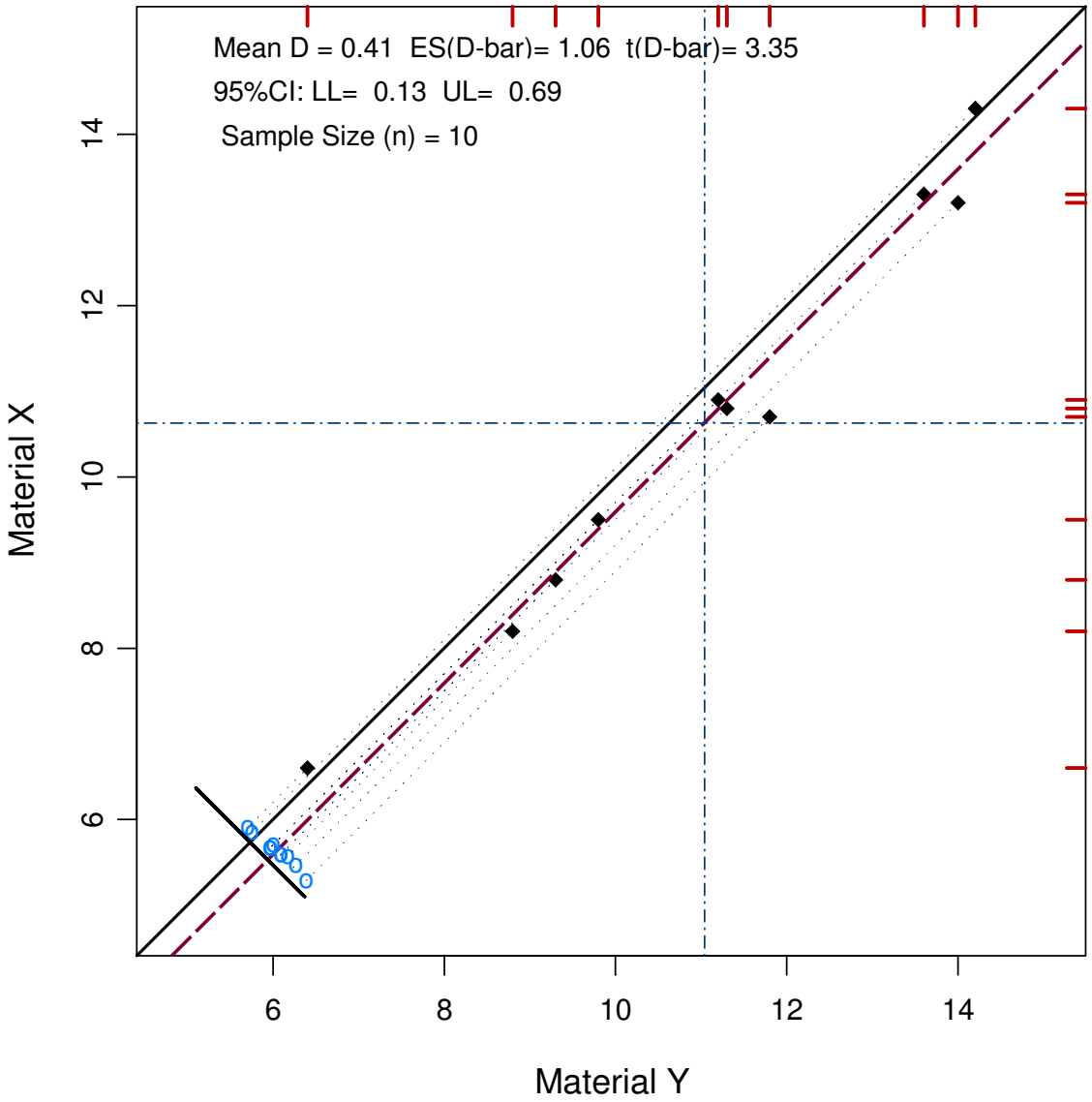
Dependent sample difference plot, Darwin Fertilization data



The data shown above are measurements recorded by Charles Darwin in 1878. The X,Y pairs correspond to heights in inches of cross-fertilized and self-fertilized plants, Zea mays, where each pair had been grown in the same pot.

Note that the *t*-statistic again reaches significance by conventional standards despite the small sample size ($t = 2.15$), showing an advantage of cross-fertilization. Effect size is .56, but the correlation between the two measures is actually negative, $r_{xy} = -.33$. Perhaps the most notable result, shown clearly in the graphic, is that two pots showed self-fertilization to work *much better* than cross-fertilization. It would be interesting to hear Darwin's commentary about this result – which he almost surely was aware of, since he was noted for his keen observational skills. As is common, the statistical analyst who presented these data said nothing about the outliers. Still, it is a key point for our purposes to note that real data often entail such complications, and their study may well lead to useful findings, or qualifications. More about such things later.

Box, Hunter & Hunter shoe wear data, for ten boys



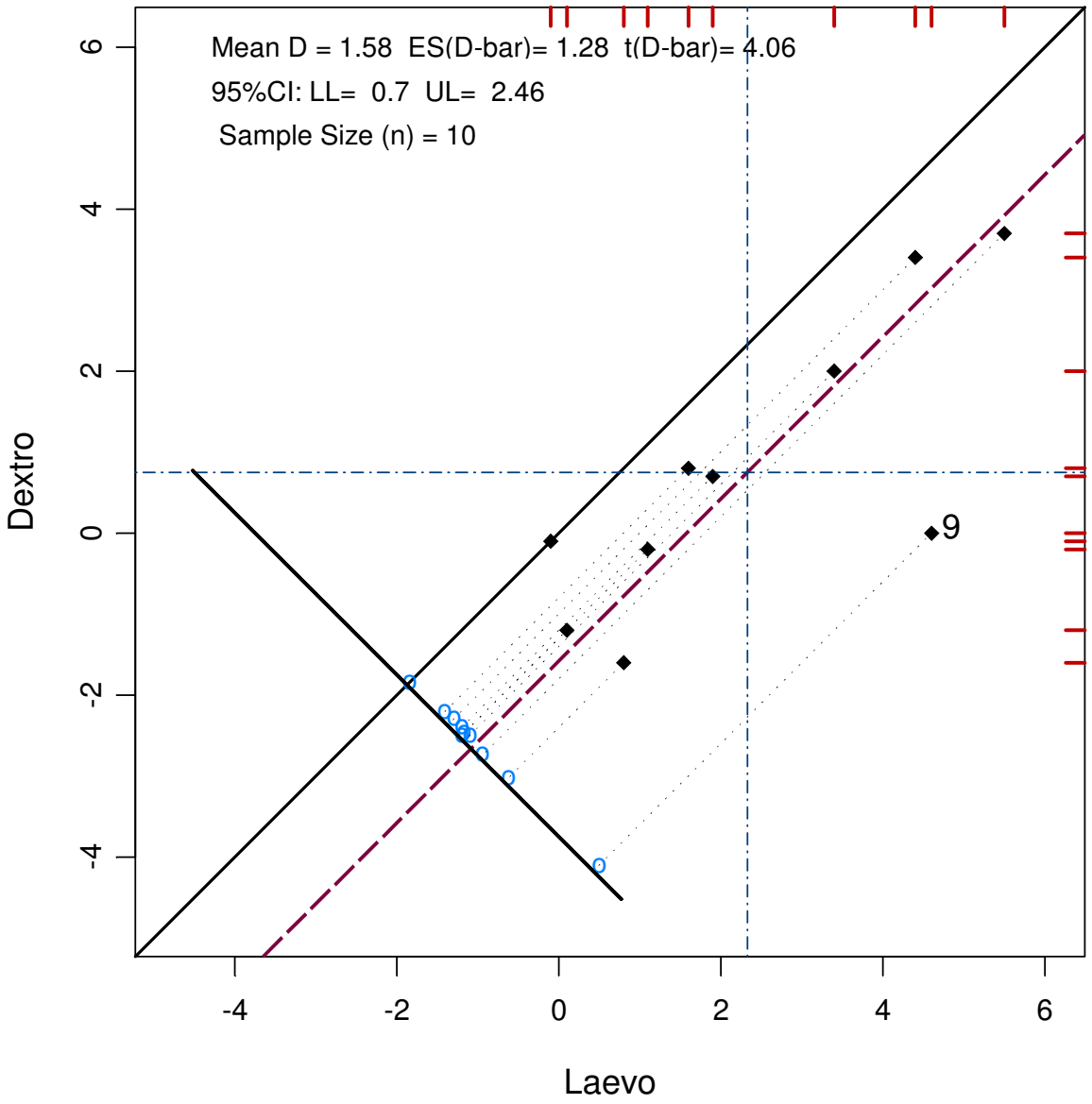
The figure above displays the well-known data set on shoe wear initially given in Box, Hunter and Hunter (1978). One sole made of 'X-material,' and another made of 'Y-material,' were randomly assigned to the two shoes of ten boys. The X &Y scores are measures of wear for the two materials. This true experiment is again a matched pairs design and it too yields a relatively strong cause-effect inference, despite the small sample size. This is a consequence of the very small variation in the D's. The near uniformity of effects is also largely responsible for an effect size whose magnitude exceeds unity.

In cases like this the numerical summary is particularly effective in summarizing the data, and the plot shows clearly why these data provide a basis for generalization. Although data from the behavioral sciences rarely admit to the precision of physical measurement seen here, and therefore rarely show such small variation in effects, the basic methods still provide a sound model for planning of an experiment. The key is to find 'homogeneous pairs' at the outset.

Interim summary: All* foregoing graphics pertain to results from ***true matched pairs experiments***. Each point in each plot corresponds to a particular replication a larger experiment; the set of difference scores summarizes effects for *n* experimental replications. In each case *random assignment* within pairs was used (sometimes tacitly); this provided a basis for making relatively strong inferences of *cause and effect* at the end of the experiment. *Absent treatment effects, each point should lie close to the X=Y diagonal; departures from this diagonal signify effects; systematic departures will often generalize.* The most efficient experiments are ones where the ***members of pairs are highly similar at the outset***. Still, as outliers reminded us, even the best of such experiments can result in findings that complicate or qualify results. In general, ***interactions*** may be expected, a point worthy of discussion. [*First example recently found to be artificial.]

In the next set of slides, data for ***repeated measures studies,*** those with a time component, are presented. Although **rm** studies are in principle weaker, they can still be quite informative.

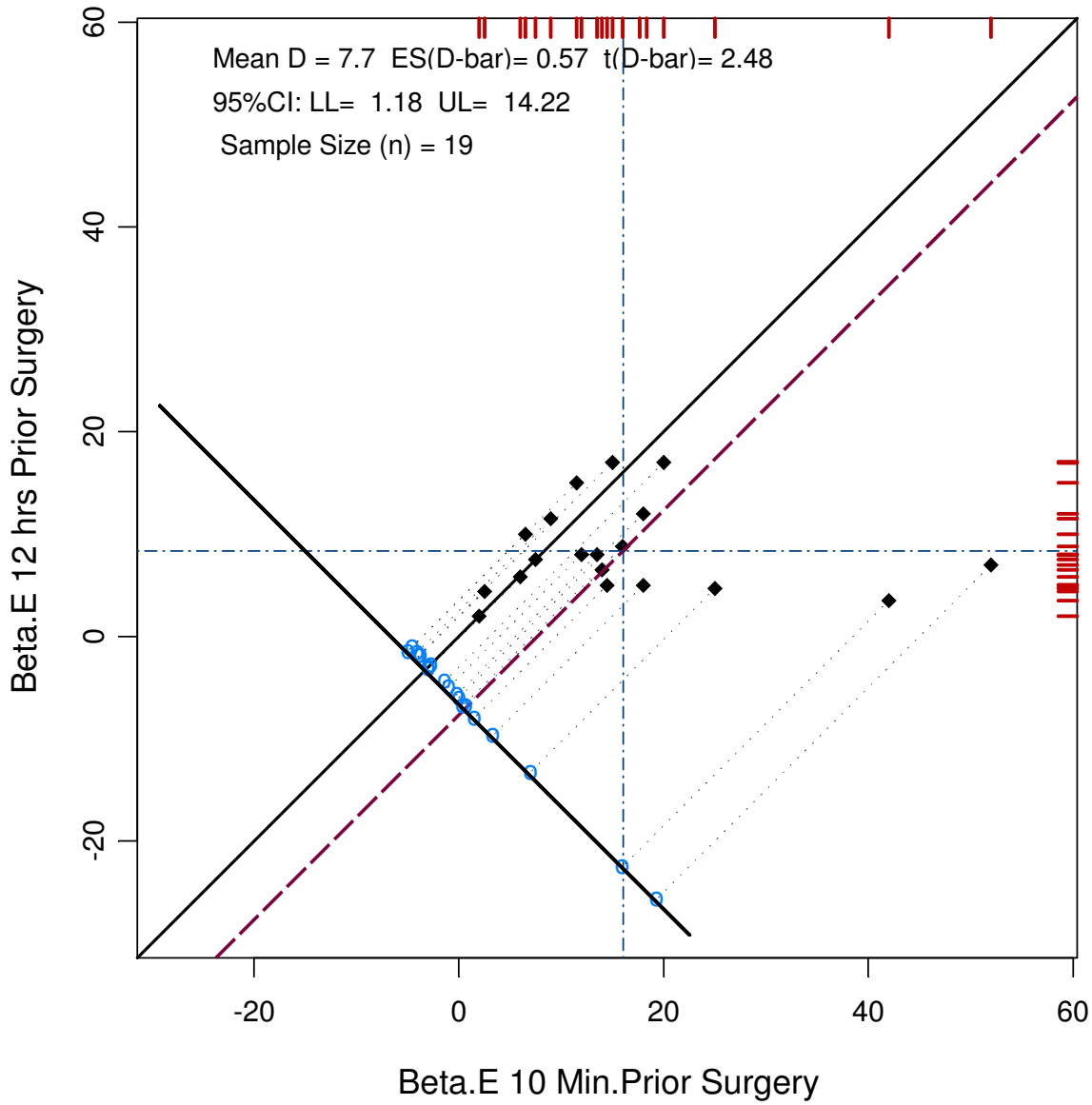
Cushney, Peebles data, hours of 'excess' sleep per week, two drugs compared



The preceding data (for ten patients) are from Cushney and Peebles (1905) *J. of Phisiology*. Initially, the average number of hours they slept was determined. In Part 1 it was decided by a flip of a coin which one of the two drugs, Laevo or Dextro, would be given first. The average (over a week) number of *excess* hours of sleep (over their usual average) was recorded. In Part 2 (after a 'wash out' period) the other drug was given; the average (for a week) number of excess hours of sleep (over their usual average) was then recorded.

The drug Laevo showed the larger effect, with a significant t (4.05). Curiously, when the point (no. 9) showing largest drug effect is removed, the t-statistic becomes larger ($t = 5.66$). You should be able to answer this question! Study details are unavailable, which is just as well since these drugs are no longer of interest. Still, the methodology of that study remains worthy of examination.

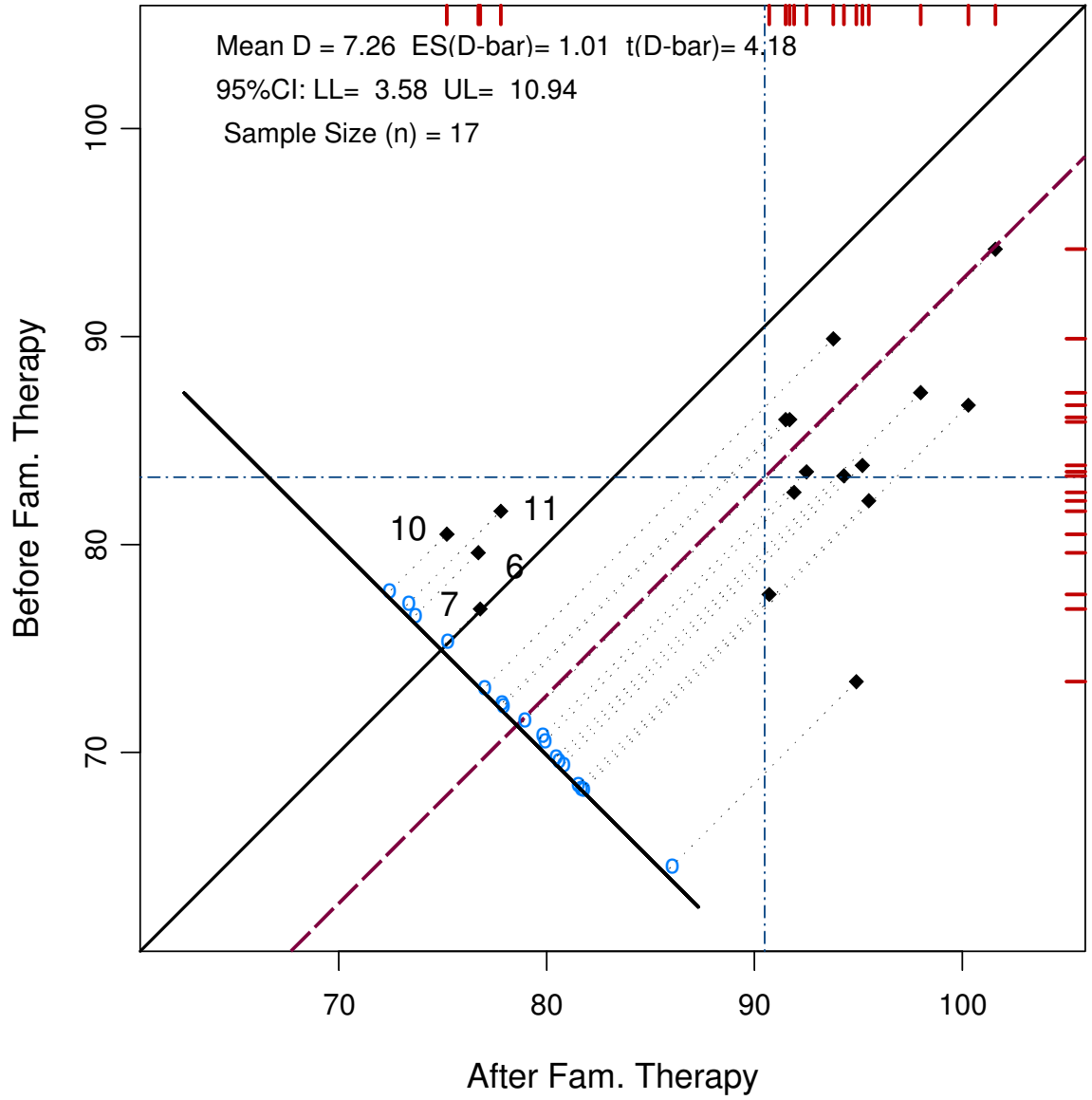
Beta-endorphin scores (to reflect stress) of patients prior to surgery



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, X's show levels 10 minutes before surgery; beta-endorphin levels in the blood were taken to measure stress. These are repeated measures data, *not* those of a 'true experiment.' Conventional analysis yields a *t*-statistic equal to 2.48, suggesting that stress levels rise 'significantly' just prior to surgery. The effect size of .57 is moderate-to-large by conventional standards.

The plot, however, tells a rather different story: Three or four patients had beta-endorphin levels much higher just prior to surgery, compared with earlier scores. But if data for these four highest D's are removed, the remaining data depict a much lower stress effect, non-significant (α =.05), and a 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. What is your interpretation?

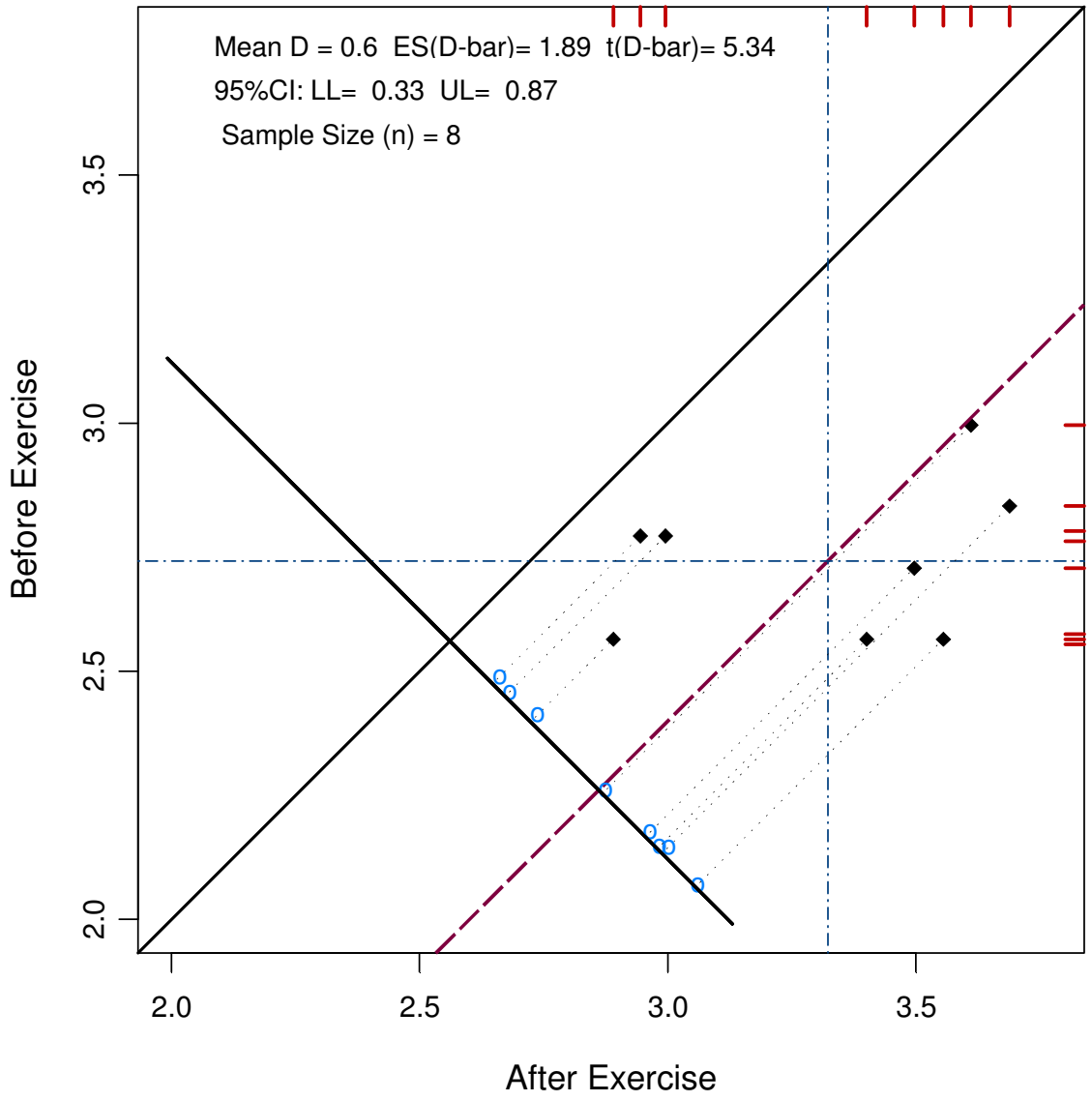
Weights of n=17 girls before and after family therapy for anorexia



The data set shown above consists of weights in pounds for $n = 17$ girls who were weighed before and after therapy as a treatment for anorexia. These data were originally published by Hand, *et al*, 1994. Y scores are weights (in lbs.) before therapy; X scores are corresponding weights after family therapy. Note that most girls gained weight, with an average gain of 7.26 lbs; the overall effect is statistically significant effect (t= 4.18) and even the standardized effect size, 1.01, is notable. (Again, this is a repeated measures study, with intervening treatment.)

The plot, however, tells a more nuanced story. The *cluster of points* at the extreme left shows that 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. Surely, one would like to know more about the girls who lost weight over the course of therapy. What is your interpretation of these results in light of this plot?

Effects of exercise on lactate for eight men; Devore & Peck, p. 435



Several men participated in a training camp in which blood lactate levels were measured before and after exercise, viz., three games of racquetball. The results show (logarithms of) lactate levels for eight men that were reported in the research article following the data collection, *Research Quarterly for Exercise and Sport*, (1991): 109-114. As you can see, the summary statistics show a 'highly significant' statistical effect, and of course a confidence interval that does not span zero.

The plot, however, shows *two subsets of men*, of size three and five respectively, for which the results seem notably different. Note that logs were taken because this transform of the data led to a larger *t*-statistic, and sharpened the picture. It appears that different men, perhaps with differing levels of fitness, responded quite differently from one another, although the sample size may be too small to infer much here.

Discussion: 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, triplets were constructed in the context of comparing three treatments. In such a case, *planned comparison contrasts* may be used to generate *two pairs of difference scores* for the *system of contrasts*. For example, one could use coefficients such as $\mathbf{c}_1 = [1, -1, 0]$ to generate a difference score for the first in relation to the second treatment, then $\mathbf{c}_2 = [\ ½, ½\ , -1]$ to yield difference scores based on comparison of the *average* of the first two treatments, and the third. The idea extends easily to several treatment comparisons via contrasts.

If they are appropriate, effect sizes, formal hypothesis tests, and confidence intervals are easily generated in these contexts. Prior knowledge serves as a substitute, as it were, for data and can lead to highly efficient use of resources when advantage is taken of homogeneous pairs, triplets, etc. It is always advantageous to use resources efficiently and the foregoing examples show that there may be notable potential for 'maximal gain from minimal resources' if a graphically based dependent sample paradigm is used effectively.

But note well: If graphics do expose clusters, patterns or outliers, then interpretation of conventional summary statistics will generally be incomplete, and can be misleading. Initial questions or hypotheses may demand modification or major revision, depending on details of what the data have to say. For example, if notable clusters of points are found in such plots, this can be taken as evidence of interactions, and one will want to learn what distinguishes the clusters of points from one another. Further studies may be helpful to track down just how experimental effects are dependent on how units are defined or selected. Review and extend these examples to generalize results.

Appendix

Features to note about dependent sample difference score plots:

1. The heavy diagonal line has *intercept zero, slope one*. Since $X = Y$ for line, it follows that difference points, $D = X - Y$, below this line indicate X is larger than Y, and vice versa. (For a similar plot, see Rosenbaum (1989, *American Statistician*).

2. Each X, Y point corresponds to a bold diamond; the marginal distribution of X is given by the 'rug' plot (red ticks) along the top, similarly for marginal distribution of Y along right side of plot. Dashed horizontal and vertical lines correspond to $\bar{X}$, $\bar{Y}$.

3. The perpendicular distance between any point and heavy black diagonal *corresponds to* a difference $D = X - Y$; however, the perpendicular projection showing Cartesian distance between any point and the diagonal actually equals D/sqrt(2). This is because distance measured on the diagonal requires adjusted to correspond to horizontal or vertical metric.

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

5. The upper-left legend provides the numerical value of $\bar{D}$ & 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 includes the t-statistic associated with $\bar{D}$.

6. The second line of upper-left legend provides 95% confidence limits for mean population difference. Sample size, the number of D's, is denoted as n.

Exercise: Using either data from the one page handout (p. 429 ff of Devore and Peck, 1994), or from any other source you find interesting, use the `dep.samp.aplt` function (below) to analyze dependent sample data for two groups. ***Be sure to write up your results to show comprehensively what you have learned from the data analysis***.