

Some guidelines for small sample research: dependent samples to the rescue

R.M.Pruzek 2005

My principal goal on the following pages is to provide evidence through examples to show how a dependent sample paradigm can be helpful or constructive in research based on relatively small samples. The focus is on dependent sample comparisons. Some of what follows bears a similarity to what can be found in conventional introductory statistics books wherein the two dependent sample paradigm is nearly always covered. Still, this prototype is rarely highlighted, often appears late in the text, rarely with real data that are comprehensively discussed, and almost never with revealing graphics! Take special note of how dependency derives from how data are collected, first for an experimental study, then for its observational counterpart. In the case of experiments, the aim is to make blocks as homogeneous as possible with respect to the ultimate response variable (which of course has not yet been observed, so knowledge or experience must ground method(s) used to construct blocks). In the case of observational studies aim to match as closely as possible on key variables likely to be closely related to the response, again using whatever means seem reasonable or sensible to accomplish the feat.

While attention is restricted (mostly) to comparisons of two treatments, or a treatment with a control, it should be clear that the general concepts of blocking or matching can work well for comparisons of three treatments, or two treatments with a control, etc. Given three groups, A, B and C, one can always (for some manipulation of these labels) compare A with B ignoring C; also, the average of A & B with C. So for three groups, two pairs result, each of which yields a plots not unlike those shown below.

The investigator should look for evidence about whether treatments work differently for different blocks, i.e. kinds of entities or individuals. Graphical methods seem ideally suited for this purpose. While the investigator should always be alert to complications in findings s/he should recognize that with small samples s/he is likely to know more about individual units than in large sample settings and this can help explain complications when n's are small.

While small samples generally yield weaker inferences than large samples, small sample inferences need not always be so weak as to be useless. Sometimes increasing n means to muddy or obfuscate principal comparisons. A key to improvement, at least in intervention research or treatment comparisons, is to focus on comparisons that are most warranted, to capitalize on virtues afforded by close study of the few cases available in order to make findings most meaningful or interpretable. Knowledge can be a substitute for data. There are virtues & limitations to both small sample and large sample studies; when resources or circumstances are such as to force one into small sample situations then use of concepts and methods for dependent sample comparisons may help make the most of the situation at hand.

What follows is a small bootstrapping experiment to demonstrate that small samples can provide confident inferences about effects for two dependent sample comparisons

two pairs, (z2,z3) & (z1, z2) were used for these purposes; I focus on (z2, z3) initially

<pre>Cbind(z1, z2, z3) <u>z1</u> <u>z2</u> <u>z3</u> 1 5.0 8.0 8.0 2 4.6 8.3 9.2 3 2.4 3.6 4.5 4 7.7 8.4 9.3 5 6.4 8.9 6.8 6 7.8 5.9 7.5 7 8.3 7.8 10.3 8 5.7 6.4 4.2 9 3.2 1.8 3.1 10 2.7 5.1 6.2 11 8.0 8.6 7.3 12 4.9 5.3 7.6 13 5.1 4.8 6.5 14 5.4 5.0 3.6 15 6.9 9.5 7.9 16 3.9 5.2 10.1 17 3.5 4.5 6.9 18 7.2 6.5 8.8 19 2.8 5.6 9.3 20 3.6 5.4 9.5 21 5.6 7.3 9.7 22 2.0 5.4 8.8 23 1.3 3.5 8.3 24 4.9 9.4 7.8 25 6.1 7.5 6.6</pre> <p><i>These are simulated data, made to be correlated, i.e., interdependent.</i></p>	<pre>> rnd2(my.summary(cbind(z1, z2, z3))) <u>z1</u> <u>z2</u> <u>z3</u> means 5.00 6.30 7.50 ← note differences in the means s.d.s 2.00 2.00 2.00 In fact, only the <z2,z3> pair are studied here. skewns -.03 -.15 -.68 The interested student should try similar low 1.32 1.80 3.09 things with, say the <z1,z2> pair; and finally, high 8.27 9.53 10.25 the <z1,z3> pair to see what happens w/ small r. -----</pre> <pre>> rnd2(cor(cbind(z1, z2, z3))) <u>z1</u> <u>z2</u> <u>z3</u> z1 1.00 .67 .15 Dependencies are central; see the bold correlations .67 1.00 .42 z3 .15 .42 1.00 -----</pre> <p><i>NB:</i> The next three pages provide simple demonstrations showing, 1st using simulated data, how dependent sample analyses might proceed. Recently developed graphical tools are used to show the data points, as well as the difference scores for a mainstream method, <i>a paired sample comparison</i>. Note that the usual <i>t</i>-statistic for this comparison is printed in the legend in the upper left corner of the first plot below; so is the corresponding standardized effect size and related statistics. More importantly, this kind of graphic shows ALL the information in the data for this comparison. Such an analysis is especially relevant when it is desired to compare two treatments (and generalizations to 3, 4 etc. are straightforward) in situations where blocks have been formed at the outset, then comparisons are made of treatment outcomes within each block. In such cases each point in the dependent sample assessment plot pertains to a single block, and an experimental effect can be discerned according to how far each point (block result) is from the identity line, the locus of points where X = Y. Bootstrap methods are used below to strengthen confidence in the inference, and especially to provide useful inferences about the standardized ES.</p> <p><i>(The z1,z2 comparison was deleted to save space/time at the last minute; it did not add much that the real data analyses do not show, especially the Snedecor and Cochran data.)</i></p> <p><i>Note that all analyses have been done using the freeware R; see note below on p. 4.</i></p>
--	---

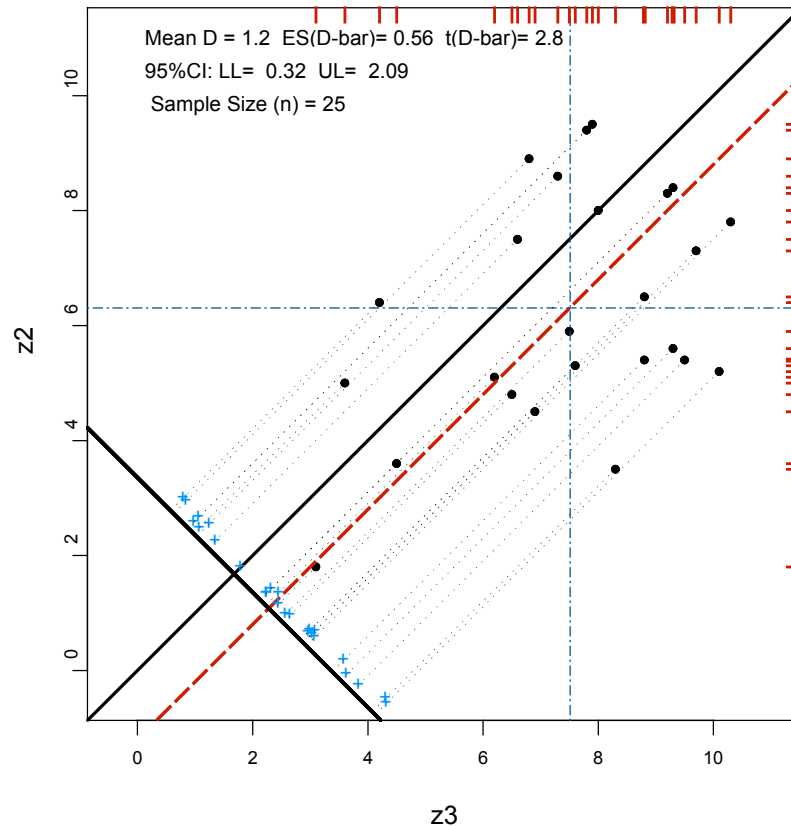
```
>dep.samp.aplt(cbind(z3,z2))
```

Summary Stats	
n	25.00
mean (D)	1.20
SD (D)	2.15
ES (D)	.56
LL 95%CI	.31
UL 95%CI	2.09
r (x, y)	.42
r (x+y, d)	.00
t (D-bar)	2.80
Wilcx.Z	2.41

The statistics above are for $n=25$ points for a two dependent sample comparison w/ a mean difference of 1.2, a standardized effect size (ES) of .56, a confidence interval that does **not span zero** (and the correlation between the two variates is only .42). The conventional t statistic is 2.80, clearly significant despite the small sample size. (A two **independent** sample t for this n and these means is NOT significant; the t is 1.59 under independence.)

Even better results occur when there is a stronger dependency between X & Y , as in $\langle z1, z2 \rangle$, or the Snedecor-Cochran data below, so that n 's notably smaller than 25 may be sufficient to detect non-null effects, i.e. to make inferences that effects generalize to the relevant population.

Dependent Sample Difference Score Assessment Plot



Note that the figure is just a scatterplot between two variables, $X=z3$ and $Y=z2$, where the heavy (black) diagonal line corresponds to a identity: $X = Y$; the key assumption is that X and Y are commensurable. Points below the identity diagonal show $X > Y$, and vice-versa. The blue crosses on the line segment at the lower left show the distribution of $X-Y$ differences; the heavy dashed (red) line depicts the mean of the differences. The (blue) dashed horizontal and vertical lines correspond to the means of Y and X respectively; marginal distributions for X & Y are at top and on the right margins. Finally, the key summary statistics are provided in the legend at upper left.

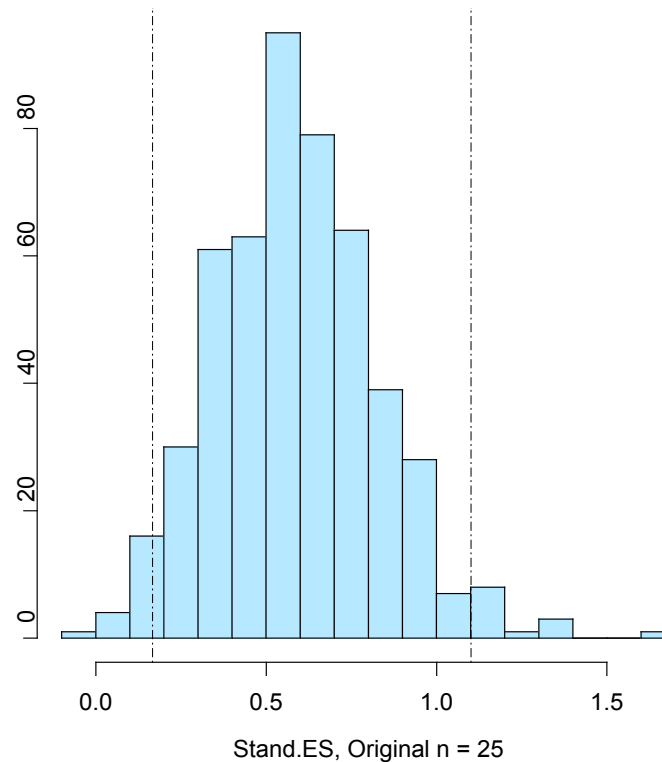
```
>boot.twodep(cbind(z2,z3),500)
  diff.mns std.diff Stand.ES
  1.22      2.1      0.59
95%CI.lwES 95%CI.upES t.diff
0.17      1.1      2.94
Emp.powr(prob. Rej. H0[false])
0.78
```

**Examine these values
in relation to those above.**

These *bootstrap results* (obtained by sampling the n rows of the input data **500 times [w/ replacement]**, and computing effect sizes, and more, for each sample) **provide a basis for inference**, including a 95% confidence interval for population Standardized Effect Size, where the **distribution of ES values across $B = 500$ bootstrap samples is plotted to the right.**

Of course the R software is free, can be downloaded from r-project.org [CRAN]; and these functions **dep.samp.aplt & boot.twodep** can be obtained without cost from rpruzek@uamail.albany.edu

Bootstrap distrib., Standardized ES's w/ .025 & .975 Confidence limits, B= 500



The above example, as noted, was not based on real data; but these two functions work equally well on real data, and can be especially helpful for small samples, possibly (much) smaller than 25. See the analysis that follows. The next data set is real, as it was reported in Snedecor & Cochran's Statistical Methods (6th Edit.). It is their illustration of a two dependent sample comparison; data were collected by Youden and Beale. There are only $n = 8$ pairs in this **randomized block design** (logs of counts of lesions were used in the analysis), still the following results help show that an inference of a significant effect is supported when (non-parametric) bootstrapping methods are used. (The graphic below helps confirm that there are no outliers, patterns, or clusters in the X, Y data system whose existence could tend to complicate such an analysis, and especially its interpretation; in general, dependent sample comparisons such these confer generalizable results when the ES confidence interval does *not* span zero, and when anomalies are not found in the X, Y point plot.)

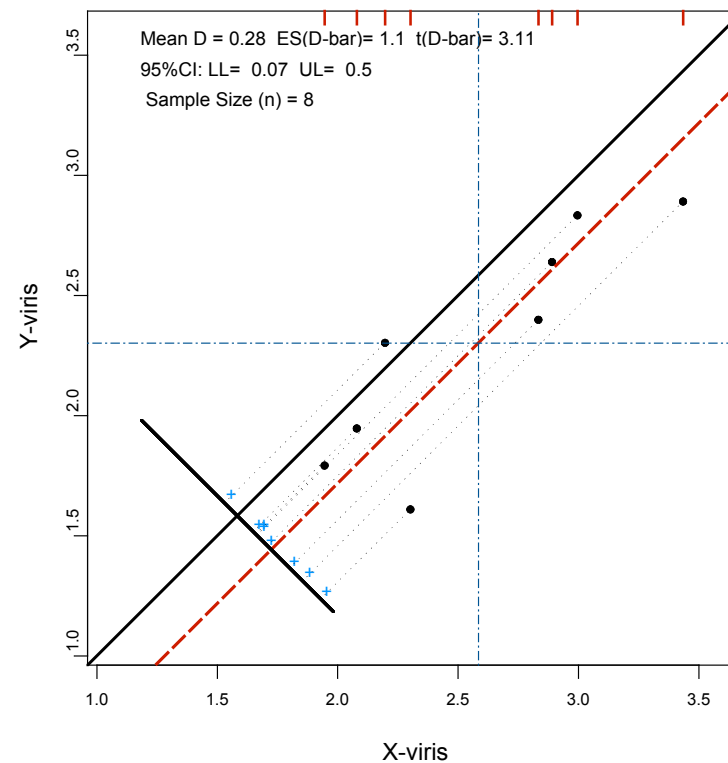
```
dep.samp.aplt(log(snd.coch.x2))
```

Summary Stats

n	8.000
mean (D)	0.284
SD (D)	0.258
ES (D)	1.100
LL 95%CI	0.070
UL 95%CI	0.500
r(x, y)	0.870 ←
r(x+y, d)	0.180
t(D-bar)	3.113
Wilcx.Z	2.310

X and Y on right correspond to counts of nos. of lesions for two kinds of tobacco virus; the data are real; the *X-virus can is seen to be stronger than Y for the population of comparable tobacco leaves.*

Dependent Sample Difference Score Assessment Plot



diff.mns std.diff **Stand.ES**

0.29 0.24 **1.26**

95%CI.lwES 95%CI.upES

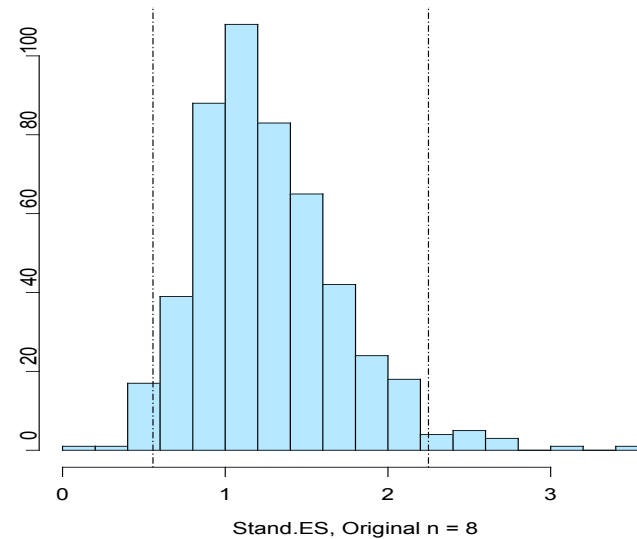
0.56 2.25

t.diff Emp.powr

3.53 0.92

The bootstrap distribution on the right, w/ the confidence interval identified, shows the range of values that has statistical support for the population *standardized effect size* in this comparison of two kinds of viruses for tobacco leaves. *The inference is strong despite fact that n = 8!*

Bootstrap distrib., Standardized ES's w/ .025 & .975 Confidence limits, B= 500



Next, we examine *observational data*, based on Table 3.1 in Rosenbaum's *Observational Studies* (2002). Note that instead of blocking, which is a type of a priori matching, this matching is done after the 'treatment' (here Exposed) data have been collected. Observational study results are generally weaker than those of true experiments, and sample sizes generally should therefore be larger for observational studies. Still, the results are notable in the following case despite the fact that there are only n = 33 pairs. The data are explained in the left column, below.

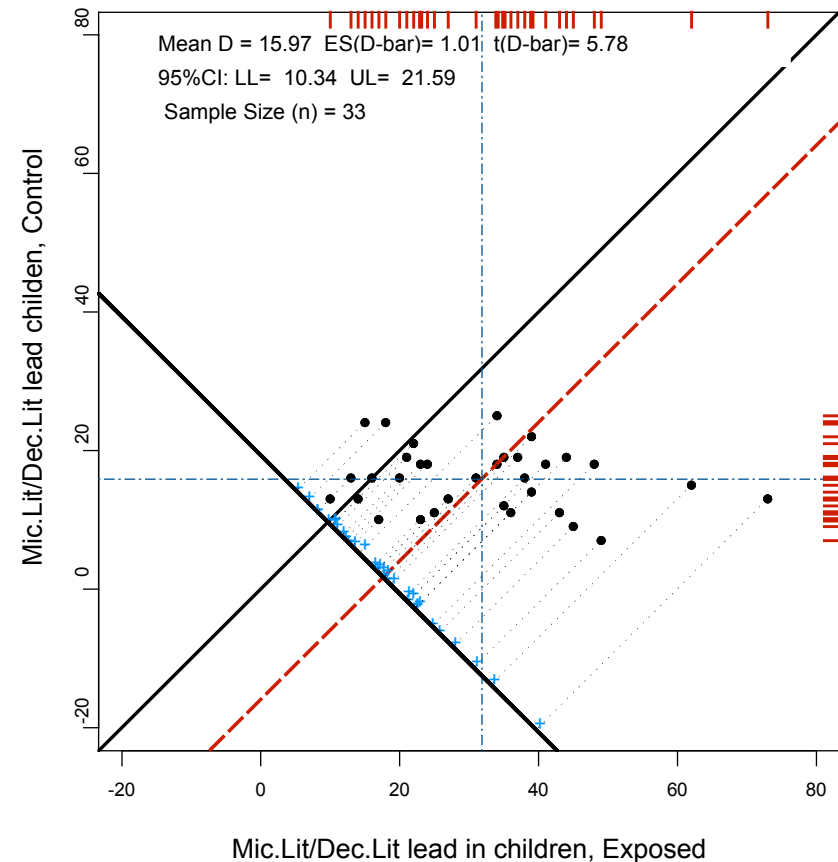
```
>dep.samp.aplt((lead.x2.rosnbm82),
pm=1.2,lf=.4,tt=F)
```

Summary Stats

n	33.00
mean (D)	15.97
SD (D)	15.86
ES (D)	1.01
LL 95%CI	10.34
UL 95%CI	21.59
r (x, y)	-0.18
r (x+y, d)	0.82
t (D-bar)	5.78
Wilcx.Z	4.41

Data on right are based on an observational study by Morten, et. al (1982, *Amer. Jour. Epidemiology*, p. 549 ff). **Children of parents who worked in a factory where lead was used in making batteries were matched by age and neighborhood w/ children whose parents did not work in lead-related industries.** Whole blood was assessed for lead content yielding MicroLiters/dl; results shown compare the Exposed w/ Control Children. Note the *narrow dispersion of lead for Control children compared w/ wide dispersion for Exposed.* It is not certain that control & exposed children did not differ in other ways (than age and residence) so Rosenbaum (2002) uses a sensitivity analysis to show that the hidden bias would have to be quite extreme to explain away differences this large. ***It seems reasonable therefore to conclude that children exposed to lead via parents' employment do generally have more lead in their blood. Consider: is this a causal inference? ?***

Dep. Samp. Assessment Plot, Matched Pairs Children, n = 3

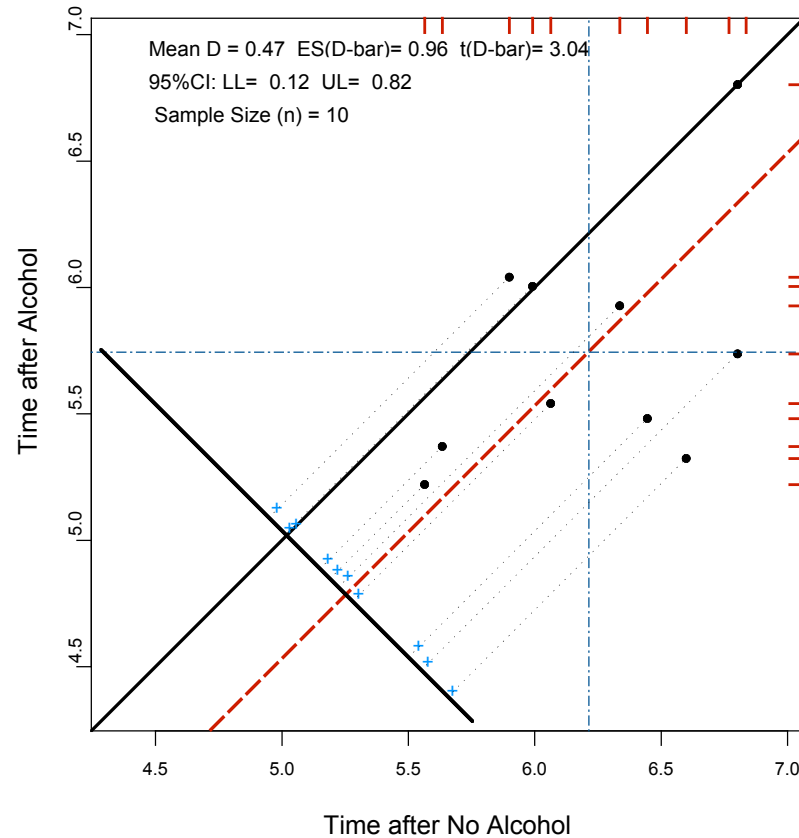


dep.samp.aplt(log(pilots.perf.alc))

Summary Stats	
n	10.000
mean (D)	0.470
SD (D)	0.488
ES (D)	0.960
LL 95%CI	0.120
UL 95%CI	0.820
r (x, y)	0.440
r (x+y, d)	-0.050
t (D-bar)	3.044
Wilcx.Z	2.190

Data for 10 pilots who performed tasks at a simulated altitude of 25,000 feet taken from Utts and Heckard (2004). [*Mind on statistics*] before and, and three days later, after consuming alcohol. Q: Does useful performance time increase with alcohol use compared to when sober? *Note: longer time shows degradation of performance.* Use of logs in analysis helps sharpen focus, and increase power, a common finding when outcomes are units of time, or frequency counts. See next page for bootstrapping results for this small data set.

Dependent Sample Difference Score Assessment Plot



```
boot.twodep(log(pilots.perf.alc[,2:1]))
```

diff.mns std.diff Stand.ES

0.47 0.46 1.02

95%CI.lwES 95%CI.upES t.diff

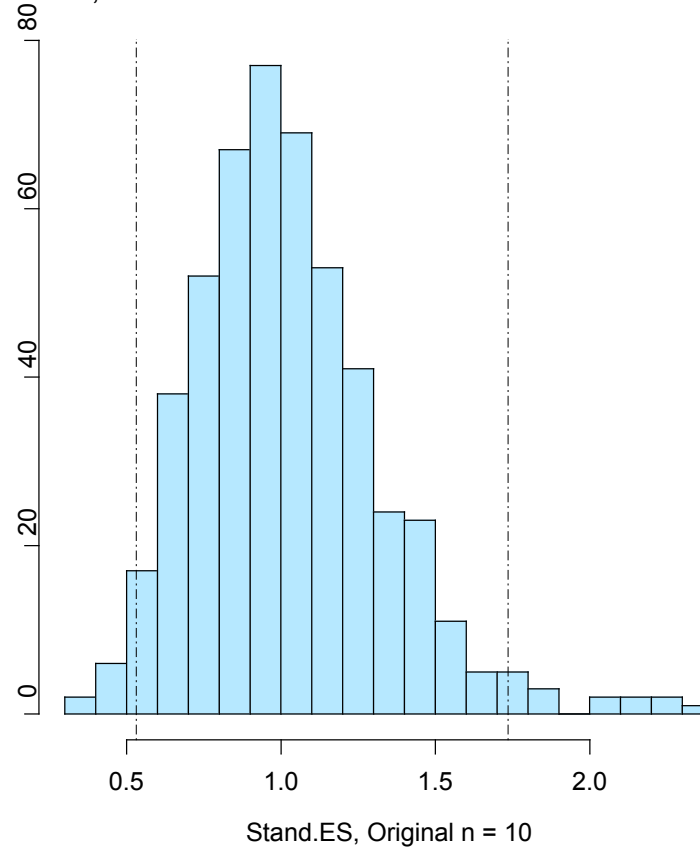
0.53 1.74 3.24

Emp.pwr

0.93

These bootstrapping results help confirm (without normality or any other parametric assumptions) that the 'notable' statistical finding generalizes to the putative population. Note that even though there are only 10 pilots in this sample, the confidence interval for the standardized effect size is far from zero (which would indicate no effect).

Bootstrap distrib., Standardized ES's w/ .025 & .975 Confidence limits, B= 500



```

boot.twodep <-function(xx, B = 500)
{ #generates bootstrap results for standardized effect sizes for dependent sample data
  #xx is taken to be an input matrix with n rows and two columns, say X and Y.
  xbts <- matrix(0, B, 4)
  for(i in 1:B) {
    xs <- xx[sample(1:nrow(xx), nrow(xx), repl = T), ]
    xb2 <- colMeans(xs)
    xbd <- diff(xb2)
    xbs <- stdev(xs[, 1] - xs[, 2])
    xbts[i, ] <- c(xbd, xbs, xbd/xbs, sqrt(nrow(xx)) * (xbd/xbs))
    xbts
  }
  smxbs <- colMeans(xbts)
  tdif <- sqrt(nrow(xx)) * smxbs[3]
  clim <- quantile(xbts[, 3], probs = c(0.025, 0.975))
  emp.pwr <- length(xbts[, 4][xbts[, 4] > 2])/B
  #using '2' as cutoff for t.effect.size; this is approximate, but relevant for t.statistic
  std.es <- stdev(xbts[, 3])
  smxbs <- as.array(c(smxbs[1:3], clim, tdif, emp.pwr))
  dimnames(smxbs) <- list(c("diff.mns", "std.diff", "Stand.ES", "95%CI.lwES", "95%CI.upES", "t.diff",
"Emp.pwr"))
  smxbs <- round(smxbs, 2)
  truehist(xbts[, 3], 22, prob = F, col = 13, xlab = paste("Stand.ES, Original n =", nrow(xx)))
  title(paste("Bootstrap distrib., Standardized ES's w/ .025 & .975 Confidence limits, B=", B), cex = 0.8)
  abline(v = clim, lty = 3, lwd = 0.8)
  smxbs
}

```