JMP Academic Case Study 010

Subliminal Messages

Pre-Test and Post-Test

Produced by

Dr. DeWayne Derryberry, Idaho State University
Department of Mathematics

Subliminal Messages
Pre Test and Post Test

Key Ideas

Paired Versus Unpaired Data, Regression to the Mean, Two-Sample t-Test, Placebo Effect, Experiments, Sample Size and Power, and Effect Size (Cohen's d).

Background

An experiment was completed to assess whether there is evidence that subliminal messages (messages we are exposed to but may not be aware of) could help raise scores on a math skills assessment. Eighteen students who had failed a math skills test were randomly assigned to receive daily either positive subliminal messages ("Each day I am getting better at math") or neutral subliminal messages ("People are walking on the street"). The students were participating in a summer program designed to raise their math skills. At the end of the program the students were re-assessed.

The Task

Determine whether the subliminal messages were effective, and if so, by how much.

The Data **subliminal.jmp**

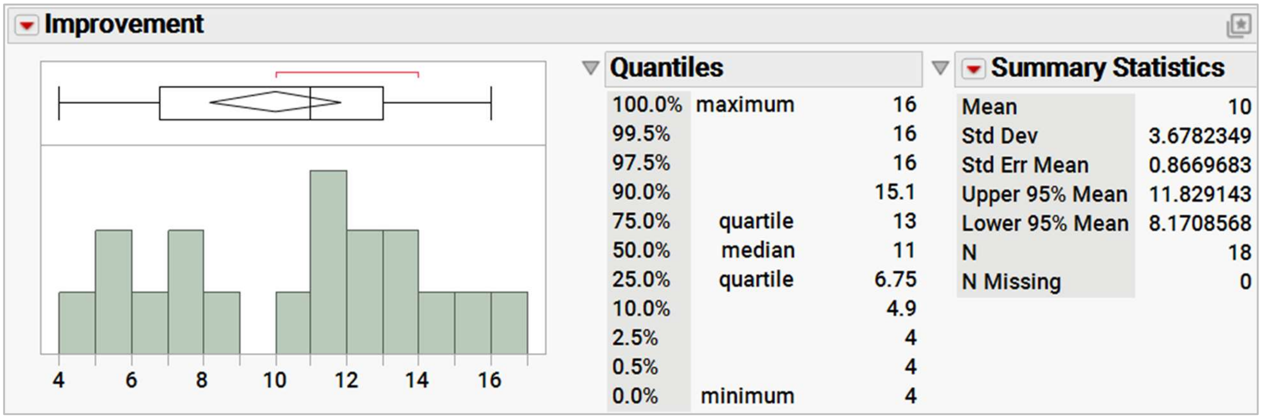
The variables are initial performance, final performance, and improvement for all 18 subjects.

Message	Whether the student received positive or neutral subliminal messages
Before	Math score upon entry into the program
After	Math score after the program
Improvement	The improvement in scores after the program (After – Before)

Analysis

Exhibit 1 shows the results of the study. Everyone showed improvement (every improvement score is positive), but did the positive message group show greater improvement?

Exhibit 1 Distribution of Improvement



Quantiles

100.0%	maximum	16
99.5%		16
97.5%		16
90.0%		15.1
75.0%	quartile	13
50.0%	median	11
25.0%	quartile	6.75
10.0%		4.9
2.5%		4
0.5%		4
0.0%	minimum	4

Summary Statistics

Mean	10
Std Dev	3.6782349
Std Err Mean	0.8669683
Upper 95% Mean	11.829143
Lower 95% Mean	8.1708568
N	18
N Missing	0

(Analyze > Distribution; select Improvement as Y, Columns and click OK. For a horizontal layout select Stack under the top red triangle.)

Improvement in the Neutral: The Neutral Group

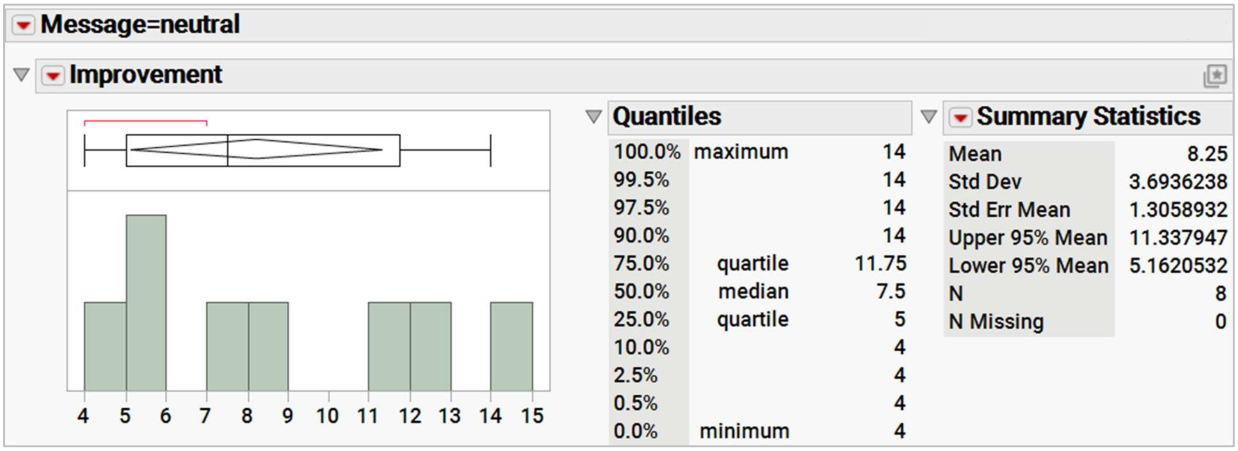
A preliminary question is whether the neutral group improved. This seems like an odd question. After all, everyone was participating in a summer-long workshop to improve math skills. However, there are many reasons why improvement might (or might not) occur, and even if there were no tasks in the program other than the subliminal messages, to aid math skills improvement.

There are three reasons why the neutral group is likely to improve in math score over the summer:

1. The program. Presumably, there were lots of activities aimed at raising math skills.
2. The placebo effect. Often, subjects will improve in an experimental setting even when any treatment they receive is only intended to appear as an effective treatment.
3. Regression to the mean. These students were chosen because they performed poorly on a previous assessment. In fact, if you take any group and segregate the very poor and very good performers on a test, the poor performers may not do well on a re-test, but will improve on average on the re-test. Similarly, those who did very well the first time will, on average, do more poorly on the second try. Performance is part luck and part skill. Those who performed very well on the first test were partly lucky, and luck is hard to replicate. Likewise, those who performed very poorly on the first try were (on average) unlucky, which is hard to replicate (and something you would not want to replicate!).

Everyone in the neutral group did, in fact, improve.

Exhibit 2 Distribution of Improvement for Neutral Group



Message=neutral

Improvement

Quantiles

100.0%	maximum	14
99.5%		14
97.5%		14
90.0%		14
75.0%	quartile	11.75
50.0%	median	7.5
25.0%	quartile	5
10.0%		4
2.5%		4
0.5%		4
0.0%	minimum	4

Summary Statistics

Mean	8.25
Std Dev	3.6936238
Std Err Mean	1.3058932
Upper 95% Mean	11.337947
Lower 95% Mean	5.1620532
N	8
N Missing	0

(Analyze > Distribution; select Improvement as Y, Columns and Message as By, and click OK.)

Although this study will allow us to determine whether positive subliminal messages have a positive impact on math skills, it cannot tell us how big a role these other factors played in everyone’s overall improvement. In particular, there is no way of knowing if the other aspects of the program were useful. This is especially significant, because in these kinds of programs where participants are selected due to a poor performance on a previous assessment, regression to the mean will lead to some improvement in any case.

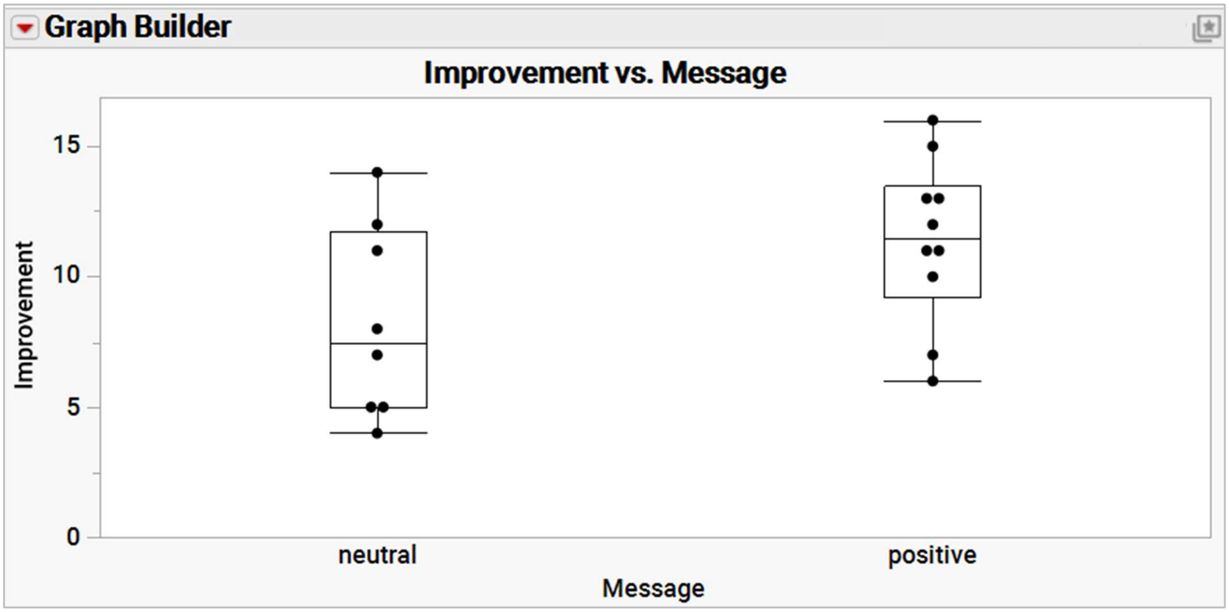
You might think about this the next time someone claims a workshop or program improved their skills. If they enrolled in the program because of a poor initial performance, some improvement is expected just due to dumb luck.

Were Subliminal Messages Effective?

The main question is whether positive subliminal messages are effective. Before any analysis we should recall that the sample sizes are very small. This means that only the most pronounced effects will produce a statistically significant result. This experiment may even be viewed as a pilot study. If the p-value is at all small and the effects appear to be big, this is a result that should be reported and follow-up studies should be done.

Do the two groups display similar improvement in scores? It appears that the group receiving the positive messages shows more improvement.

Exhibit 3 Distribution of Improvement for Neutral and Positive Messages



(Graph > Graph Builder; Drag and drop Improvement in Y and Message in X. Click and drag the box plot icon from the icon pallet at the top onto the graph, and click the Done button.)

A more formal assessment of the effectiveness of subliminal messages is to test whether the two groups have the same average improvement. Since we have two groups, we'll conduct a two-sample t-test.

Our hypotheses are:

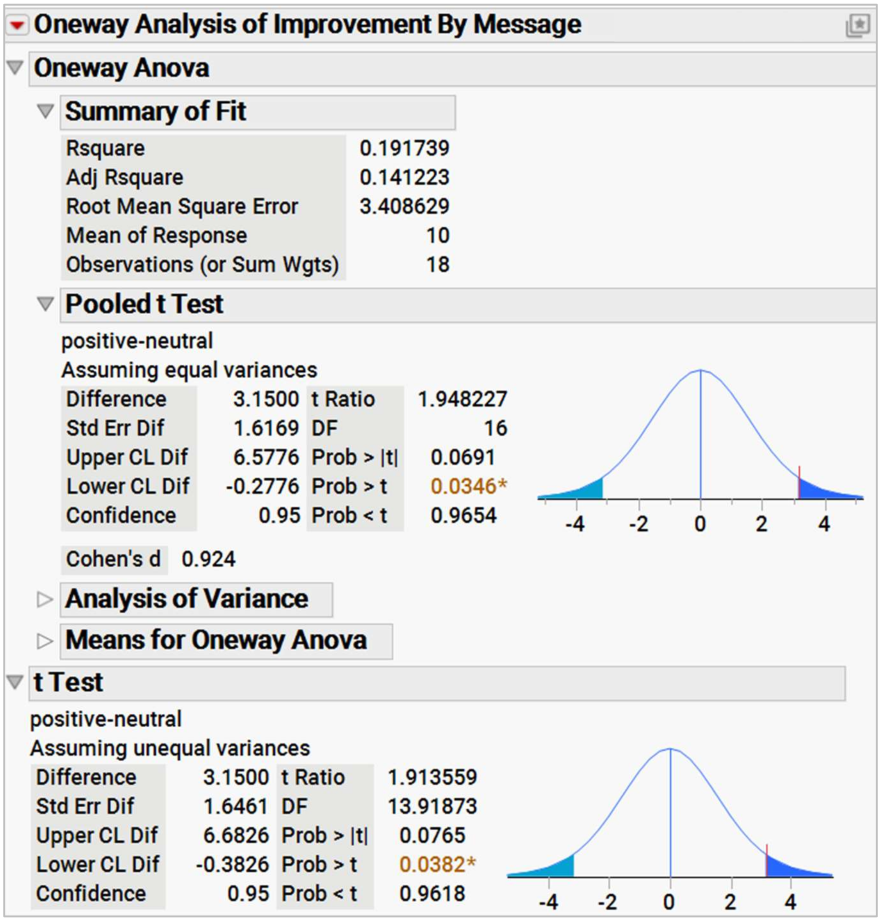
Ho: $\mu_{\text{subliminal}} = \mu_{\text{neutral}}$

Ha: $\mu_{\text{subliminal}} > \mu_{\text{neutral}}$

Since we're interested in whether students receiving positive subliminal messages show more improvement than students receiving neutral messages, we'll conduct a one-tailed test.

Both the t-test with equal variance (middle in Exhibit 4) and the test with unequal variance (bottom in Exhibit 4) produce small p-values (0.0346 and 0.0382 respectively). The improvement scores for students receiving positive subliminal messages are significantly higher.

Exhibit 4   Two Sample t-Tests for Improvement



Oneway Analysis of Improvement By Message

Oneway Anova

Summary of Fit

Rsquare	0.191739
Adj Rsquare	0.141223
Root Mean Square Error	3.408629
Mean of Response	10
Observations (or Sum Wgts)	18

Pooled t Test

positive-neutral
Assuming equal variances

Difference	3.1500	t Ratio	1.948227
Std Err Dif	1.6169	DF	16
Upper CL Dif	6.5776	Prob > |t|	0.0691
Lower CL Dif	-0.2776	Prob > t	0.0346*
Confidence	0.95	Prob < t	0.9654

Cohen's d   0.924

Analysis of Variance

Means for Oneway Anova

t Test

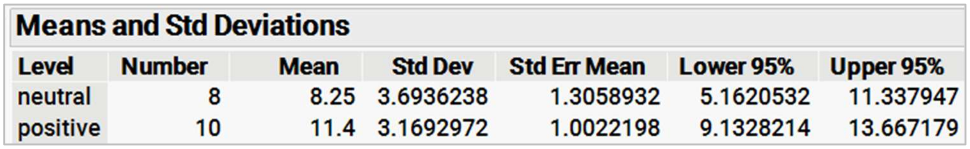
positive-neutral
Assuming unequal variances

Difference	3.1500	t Ratio	1.913559
Std Err Dif	1.6461	DF	13.91873
Upper CL Dif	6.6826	Prob > |t|	0.0765
Lower CL Dif	-0.3826	Prob > t	0.0382*
Confidence	0.95	Prob < t	0.9618

(Analyze > Fit Y by X; Select Improvement as Y, Response and Message as X, Factor, and click OK. From the red triangle select Means/ANOVA/Pooled t for the t-test for equal variances and select t Test for the t-test for unequal variances.)

Exhibit 5 shows further evidence that the improvement scores for the positive group are higher than for the neutral group. The average improvement for the positive group is 3.15 points higher than for the neutral group, and the confidence intervals for the two groups do not overlap.

Exhibit 5   Comparing Means for the Two Groups



Means and Std Deviations

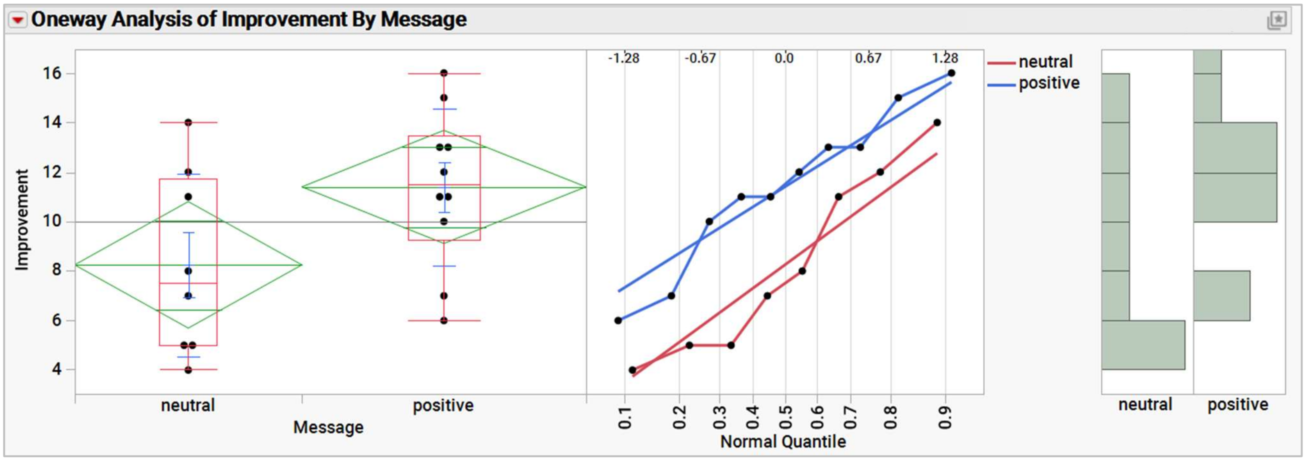
Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
neutral	8	8.25	3.6936238	1.3058932	5.1620532	11.337947
positive	10	11.4	3.1692972	1.0022198	9.1328214	13.667179

(From the Oneway output, select Means and Std Dev from the red triangle.)

When sample sizes are small, as they are here, the test may have low power. That is, it is possible for there is be a difference in the population, but the inferential procedure (in this case a two sample t-test) cannot identify a statistically significant difference (will not produce a small p-value).

In our example, despite the small sample sizes, our results are significant. However, when sample sizes are small, there is also a concern about whether the data meets the assumptions of a t-test. For small data sets (n = 8, 10) it is impossible to detect all but the most egregious violations of normality using histograms, boxplots or other displays. For example, can we detect any problems from the boxplots, normal quantile plots, or histograms in Exhibit 6?

Exhibit 6 Assessing Normality



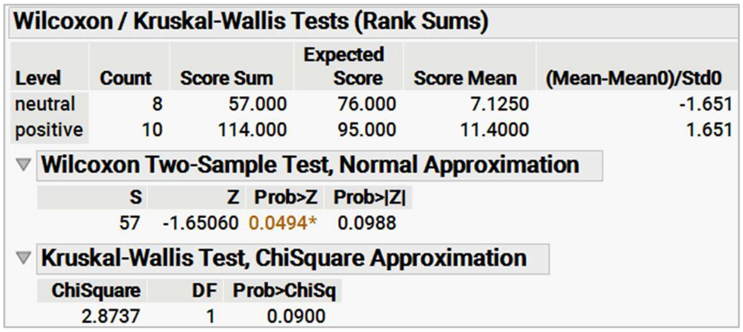
(From the initial Oneway output, under the red triangle; select Normal Quantile Plot > Plot Actual by Quantile, Display Options > Box Plots, and Display Options > Histograms.)

In this case, about all that can be concluded from the displays is that there are no extreme outliers in the data. It is possible that, based on past experience with the testing instrument and familiarity with the scores, a psychologist might know whether the data is approximately normal. Although this goes beyond the information we were given, well-designed and widely adopted psychological assessment tools usually produce well-behaved data. So, using the t-test here is a bit of an act of faith with regard to the assessment tool being used.

If there is still a concern about the distribution of the scores, the nonparametric alternative to the two-sample t-test, the rank sums test, is the usual choice. This test makes fewer assumptions, so it has less power.

The nonparametric test (Exhibit 7) gives similar, if slightly weaker, evidence of a difference in improvement (one-sided p-value = 0.0494, one half of the reported p-value).

Exhibit 7 Nonparametric Test for the Effectiveness of Positive Subliminal Messages



Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
neutral	8	57.000	76.000	7.1250	-1.651
positive	10	114.000	95.000	11.4000	1.651

Wilcoxon Two-Sample Test, Normal Approximation

S	Z	Prob>Z	Prob>|Z|
57	-1.65060	0.0494*	0.0988

Kruskal-Wallis Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
2.8737	1	0.0900

(From the initial Oneway output, under the red triangle; select Nonparametric > Wilcoxon/Kruskal-Walllis Tests.)

While both the t-test and the nonparametric test provide evidence that the improvement scores for the positive group are higher than for the neutral group, a more practical question might be, “is this a meaningful difference?” We observed a difference of 3.15 points. The hypothesis test alone does not tell us whether this difference is large or small. In fact, for tests with very small samples, a large difference may not yield significant results. Likewise, with very large samples test results may be significant even with very small differences.

Effect Size and Cohen’s d

Because of this limitation, a t-test is sometimes supplemented with a measure of effect size known as “Cohen’s d,” or simply “d.” This is a standardized measure of how much the means differ, stated in terms of the standard deviation.

$$d = \frac{|\overline{X_1} - \overline{X_2}|}{S_p}$$

This value is not computed directly in JMP. However, the sample means are provided in the Means for Oneway Anova table (Exhibit 5) and S_p, or the pooled estimate of the standard deviation, is listed as the Root Mean Square Error in the Summary of Fit table (top, in Exhibit 5).
The resulting value for the effect size is:

$$d = \frac{|\overline{X_1} - \overline{X_2}|}{S_p} = \frac{11.4 - 8.25}{3.409} = 0.9$$

To determine if this is considered to be a small, medium or large effect, we refer to the guidelines for interpreting effect sizes in Exhibit 8 (from Cohen, 1992).

Exhibit 8 Cohen Guidelines for Interpreting Effect Sizes

Effect Size	Computed d
Small Effect	d = 0.20
Medium Effect	d = 0.50
Large Effect	d = 0.80

Based on this, we can conclude that not only is the difference between the means for the neutral and positive group significant (based on the hypothesis tests), the effect size is large.

Although subliminal messages sound like quackery, there does appear to be a benefit. The difference itself, 3.15 units of improvement (Exhibit 4), sounds practically significant as well, given the range of scores under consideration. Put another way, given the before and after scores we are seeing, any method that raises scores an average of more than three points is having an impact. Given the small sample size, this is an impressive result.

Were the Groups Different Before?

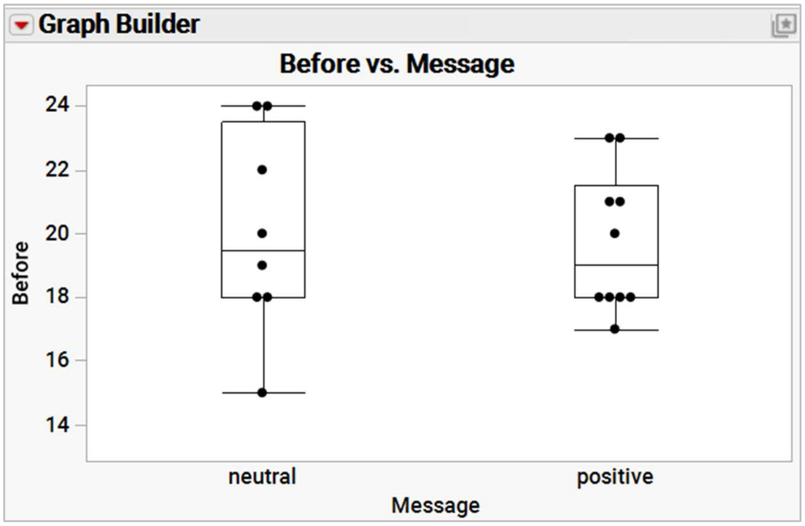
The purpose of random assignment is to begin the study with two groups that are alike, so that the role of lurking/confounding variables is minimized. In an observational study, when two groups are compared and different final outcomes are observed, the differences in the final outcomes could be due to many things, because the two groups being compared differed in many ways to begin with.

For example, suppose we observe those who do and do not own pets and observe better heart health in those who own pets. We cannot conclude pet ownership leads to better heart health because pet owners and non-pet owners differ in many ways. After all, to be a responsible pet owner, one must, to begin with, be in good enough health to take care of a pet properly. This already indicates a different health profile for these two groups.

Taking a pre-existing group and randomly assigning them to two groups reduces this problem. In any really large group, if we randomly assign subjects to two groups, those groups are likely to be alike in about every way relevant to the experiment. Proper random assignment, as a process, always works (in the sense that chance is the only factor responsible for observed differences before any treatment). However, for small groups there can sometimes be large chance differences before treatment that undermine the interpretability of the data.

Do the two groups display similar (lack of) math skills going into the experiment? Before scores for the two groups certainly appear to be similar (Exhibit 9).

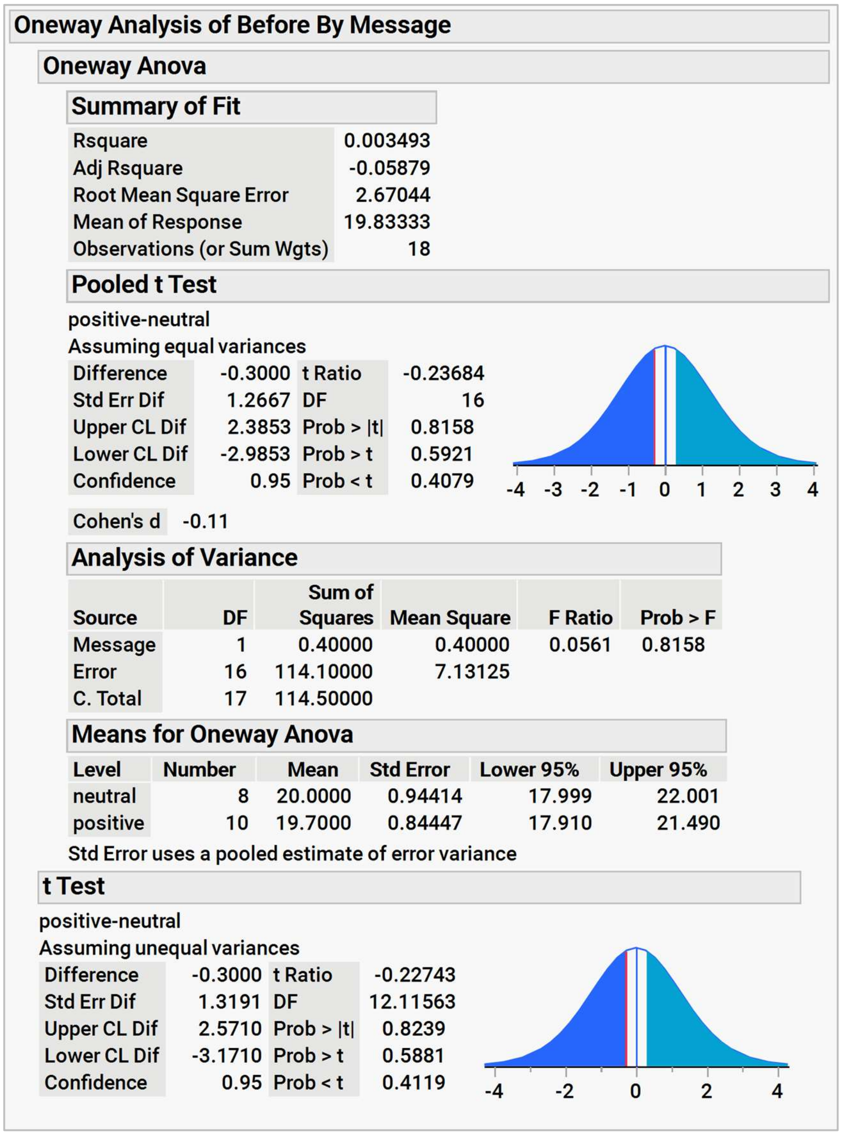
Exhibit 9 Distribution of Pre-Program Scores by Type of Message



(Graph > Graph Builder; Drag and drop Before in Y and Message in X. Click the box plot icon from the icon pallet at the top onto the graph, and click the Done button.)

The t-tests confirm that the two groups are not significantly different in math skills before the experiment (Exhibit 10).

Exhibit 10 Comparison for Before Scores for Positive and Neutral Groups



Oneway Analysis of Before By Message

Oneway Anova

Summary of Fit

Rsquare	0.003493
Adj Rsquare	-0.05879
Root Mean Square Error	2.67044
Mean of Response	19.83333
Observations (or Sum Wgts)	18

Pooled t Test

positive-neutral

Assuming equal variances

Difference	-0.3000	t Ratio	-0.23684
Std Err Dif	1.2667	DF	16
Upper CL Dif	2.3853	Prob > |t|	0.8158
Lower CL Dif	-2.9853	Prob > t	0.5921
Confidence	0.95	Prob < t	0.4079

Cohen's d	-0.11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Message	1	0.40000	0.40000	0.0561	0.8158
Error	16	114.10000	7.13125		
C. Total	17	114.50000			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
neutral	8	20.0000	0.94414	17.999	22.001
positive	10	19.7000	0.84447	17.910	21.490

Std Error uses a pooled estimate of error variance

t Test

positive-neutral

Assuming unequal variances

Difference	-0.3000	t Ratio	-0.22743
Std Err Dif	1.3191	DF	12.11563
Upper CL Dif	2.5710	Prob > |t|	0.8239
Lower CL Dif	-3.1710	Prob > t	0.5881
Confidence	0.95	Prob < t	0.4119

This gives us some degree of confidence that the groups are similar in math ability going in. However, this hypothesis test addresses the question, “Is the difference observed between the groups before treatment large enough that we believe chance isn’t responsible for this difference?” Since these groups were formed using random assignment, we know by definition that the groups are different only because of random chance. What we are really interested in knowing is whether the groups are different enough before treatment to cloud the interpretation of the differences observed after treatment.

We’ll again rely on Cohen’s d to evaluate the observed difference between the groups before the treatment. The calculated effect size is:

$$d = \frac{|\overline{X_1} - \overline{X_2}|}{S_p} = \frac{20.0 - 19.7}{2.67} = 0.11$$

Based on the guidelines for interpreting effect sizes, we can conclude that the effect size is small. This provides additional confidence that potential chance differences between the groups before the study will not affect the interpretability of our results.

Summary

Statistical Insights

With regard to generalizability, the study is limited. Others would need to use sound judgment to know to what extent this result would apply to another population, but with regard to causality this study is on solid footing. Because it was a randomized experiment, unless some further information is given to show a flaw in the procedure, the type of message is almost certainly the reason for the difference between the groups.

Some might argue that there is always a chance that the apparent difference is not cause-effect at all, but just due to putting, for example, all the hard workers in one group and all the slackers in another. Cohen's d shows that the difference between the two groups entering the study is very small, and is not likely to have an influence on the final results.

JMP® Features and Hints

In this case we used the Distribution platform and the Graph Builder to provide numeric and visual summaries of the data. Two t-tests (equal and unequal variances) and a nonparametric rank sums test were conducted from the Fit Y by X platform, and additional options for summarizing and visualizing data were selected from the red triangle in the Oneway output window.

Cohen's d was calculated by hand using the sample means and the pooled estimate of the standard deviation (reported as RMSE in the Summary of Fit Table) that were produced by running the equal variance t-test.

Exercises

A randomized comparative experiment was conducted to investigate the effect of calcium on blood pressure in African-American men. A treatment group of 10 men received a calcium supplement for 12 weeks, and a control group of 11 men received a placebo during the same period. All subjects had their seated systolic blood pressure tested before and after the 12-week period, and the decrease in blood pressure (Begin – End) was recorded.

From DASL (The Data and Story Library): lib.stat.cmu.edu/DASL/Stories/CalciumandBloodPressure.html.

The data is in subliminal Exercise.jmp. For this data set:

1. Was calcium effective in reducing blood pressure, and if so, by how much?
2. How do the results from a t-test and a nonparametric test compare?
3. Someone says, "I think calcium may be effective, but the sample size of this study prevented us from detecting it." Comment on this statement. In what way is Cohen's d useful here?
4. Calculate Cohen's d. Does this give any indication of the potential effect of calcium supplements on blood pressures?
5. Compare the blood pressures for the two groups before the experiment. Are there significant differences? Calculate Cohen's d. Should the researchers be worried about the differences between the two groups before the study?