CAUSATION AND STATISTICAL INFERENCE: POWER TOOLS FOR THE PRACTICAL

Michael A. Einhorn, Ph.D.*

1.INTRODUCTION

False advertising. Securities fraud. Infringement of intellectual property. Personal injury. Employment discrimination. Commercial damages. Mass torts. Financial recovery is possible, either in actual damages or defendant enrichment. But there is one hurdle beforehand. Go prove liability, and causation of damages.

Economists and statisticians can assist in the analysis, which may be wide-ranging consideration in a court that may consider several different analytic methods.[1] Any such trained expert presumably has some technical chops to apply some popular statistical tests on data averages to analyze the issue; e.g., the Student's t-test, paired t-test, ANOVA (analysis of variance), or linear/nonlinear regression. When comparing the average of two or more groups with the help of hypothesis tests, the assumption of each is that the data is a sample from a normally distributed population (i.e., bell shaped) These are called *parametric tests*.

But there is a catch to the classical (or parametric) approach. While the working assumption of normality can be shown to be quite valid for larger samples (i.e., above

* At mae@mediatechcopy.com, http://www.mediatechcopy.com. The author is an economic consultant and expert witness in the areas of intellectual property, media, entertainment, and product design. He is the author of *Media, Technology, and Copyright: Integrating Law and Economics* (2004) and over seventy related professional articles in intellectual property, economic analysis, and damage valuation. He is also a former professor of economics at Rutgers University. A complete professional biography can be found in the appendix. Helpful comments were received from June Besek, Chris Seidman,Robert Clarida, Dama LeJune. and James Astrachan.

[1] *Matrixx Initiatives, Inc. v. Siracusano*, 131 S. Ct. 1309 [2011]. The case was a pleadings case that consider no actual evidence.

30), litigation frequently often involves small samples that are not so suitably analyzed.[2] Indeed, the consequences of bad assumptions can be downright awful.

The high-tech answer is non-parametric statistics. Designed by big brains at the University of Chicago, Bell Laboratories, and Princeton University, non-parametric statistics do not make such strenuous assumptions about a specific distribution. Consequently, non-parametrics are known in the math shop as *robust; i.e.,* they involve a set of techniques that are valid regardless of the presence of normality

Non-parametric statistics are additionally desirable in litigation for five other reasons

1. Non-parametric analysis are easily adapted when *data on certain causal variables are missing entirely,*[3]
2. Non-parametric analysis is more useful when *influences change over time.*
3. Non-parametric analysis makes use of *extraneous benchmarks that are often passed over* in classical statistics.
4. Non-parametric analysis is *intuitive, inexpensive, and easy to explain to a jury.*
5. Reread Rule 4.

Ranked in the order of increasing statistical power, the three basic non-parametric statistics are the *sign test,* the *median test,* and the *rank sum test.* Each of these tests is now commonly found in textbooks and peer-reviewed professional articles in statistics, economics, biometrics, etc.[4] These techniques are less often found in court. A streetwise expert will consider them carefully.

2. SIGN TEST

A starting non-parametric test is the *sign test.* The sign test (or *paired alternative* test) is the easiest and most intuitive test, but is relatively weak. This means that it is less likely to confirm causation than one of the others.

---

[2]Sometimes available to the data-strapped expert is *meta-analysis,* which aggregates data from several disparate studies into a composite estimate based on the reliability of each Schachtman, infra note xx, §30A:7

[3]The issue of *omitted variable bias* in a linear regression arose in *Bazemore v. Friday,* 751 F. 2d 662, 672 (1984), rev'd, 478 U.S. 385 (1986),

[4]Most recently, Corder, G. W.; Foreman, D. I. (2014). *Nonparametric Statistics: A Step-by-Step Approach*. John Wiley & Sons; Hollander M., Wolfe D.A., Chicken E. (2014). *Nonparametric Statistical Methods*, John Wiley & Sons; Wasserman, Larry (2007). *All of Nonparametric Statistics,* Springer.

Imagine that a leading beverage company uses an infringing work in a national television commercial for its popular product line, Choka Cola Sales increase after the infringement. *Any evidence of causation?*

An expert should first start with some eyeball diagnostics. Company sales should be categorized into two (or more) beverage lines, Choka Cola and other beverages that presumably are less benefitted by the *treatment*, i.e., the advertisement. Data are illustrated in the Appendix to this paper.

Using the eyeballs, the expert views any number of comparative charts, graphs, or histograms and selects those *benchmark product(s)* that have *patterns of growth* that are similar to the contested product *prior to treatment*. The *Six Sigma website* provides a number of useful visual diagnostics that facilitate an expert's review. [5]

After reviewing a number of benchmark alternatives, I identified two other company products — Beach Ade and Awake! — as being suitable historic comparatives for Choka Cola (see Appendix). I will soon have to confirm my instincts.

The sign test should be performed first for those months in the infringement period (e.g., June, 2011 – December, 2011). Comparing pre-treatment data, experts can compute *year-over-year (e.g., June, 2010 – June, 2011; July, 2010 – July, 2011)* percent changes in the product sales in each product line. Year-over-year is often preferable to month-over-month changes, as the latter can be affected by additional *seasonal* patterns that add unwanted noise to the analysis.

Once year-over-year percent changes in the variables are determined, the expert may consider whether the treatment affected growth patterns of Cola and its benchmarks.

For example, in June 2011 to December, 2011, year-over-year changes for the Cola, Ade, and Awake! are as follows.

[5] At http://www.isixsigma.com/tools-templates/graphical-analysis-charts/ (retrieved May 25, 2015)

TABLE 1

YEAR-OVER-YEAR PERCENT CHANGE

2010-2011

	COLA	**ADE**	**AWAKE!**
June	38.6	17.9	19.9
July	40.5	1.1	-10.6
August	23.4	4.3	-18.3
Septembe	62.5	10.0	-4.3
October	53.5	-5.4	-10.5
November	99.6	0.7	6.1
December	53.7	-7.1	16.0

In two separate tests, we can confirm that Cola outpaced Ade (and Awake!) for the seven straight treatment months. Is this significant, or just a matter of luck? In any one month, the random chance of Cola outpacing either Ade is ½. This is a simple toss of a balanced coin. You can't infer that a coin is loaded simply because you flip "heads" once.

More technically, a one month observation would be indiscernible from a random evMore ent and could not invalidate a *null hypothesis*; i.e., the presumption that there is no significant difference between specified populations. Rather, experts must confirm that any result is *statistically significant; i.e.*, the result is not likely to have occurrd randomly.

Back to the table, the chances of Cola outpacing Ade for seven (eight, nine) straight months are the same as seven (eight, nine) straight successful tosses 1/128 (1/256, 1/512). Now we may have something to work with. If you flip heads seven times in a row, you can state with more than 95 percent certainty that the *data are not consistent with the null hypothesis of a random string in a balanced coin.*.[6] Your chance of

[6]For technical reasons, this should not be read or manipulated to mean the "likelihood of the null hypothesis". The confusion is termed the "transpositional fallacy"; e.g., *N. A. Schachtman, at* "Statistical Evidence in Products Liability Litigation," Chapter 30A, §30A:5.2, in Stephanie A. Scharf, Lise T. Spacapan, Traci M. Braun, and Sarah R. Marmor, eds., *Product Liability Litigation: Current Law, Strategies and Best Practices* (PLI 2014)

observing *Type I error* – i.e, the probably of observing a difference between observed and expected value when there actually is none – is quite small

After comparing Cola and Ade in the treatment period, an expert should also "true up" any invalidation of the null hypothesis. Remember that the benchmark variable Ade was originally selected because the expert had come to believe, on preliminary inspection, that the three beverages had similar growth patterns before treatment. Now, the expert must formally *confirm this intuition* by running a number of sign tests over different pre-treatment intervals (presumably of seven months length). If Cola dominates Ade in some manner in one or more of the alternative intervals, the benchmark might not be appropriate.

The expert can next move sequentially to compare Cola and Awake!. Same theory. But the problem here is that we have paired alternatives. Can we compare three (or more) sales in one move. Theoretically yes, but don't try to explain anything to a judge or jury.

Finally, statisticians know that sign tests are *weak tests*, meaning that these tests are more likely to fail than other diagnostics. Nonetheless, a preliminary sign test may set the stage for more powerful diagnostics that may yet confirm some statistical significance. The media and rank sum tests are more powerful, and also allow us to compare more efficiently three or more variables

3. **MOOD'S MEDIAN TEST**

Mood's median test is another simple diagnostic in the non-parametric toolkit. As noted, it is particularly convenient for comparing historic sales patterns simultaneously for Coke, Ade, and Awake!

In the median test, the expert combines data from Table 1 into one sample, *a total of* twenty one data points (i..e, 3 drinks x 7 months). The expert then ranks in descending order the *percent changes* of the entire group. The twenty one percent changes are then assigned to two groups -- HIGH and LOW -- depending on whether the particular percent chance ranks above or below the median. The below chart illustrates results for Table 1, based on a median of 10.0 percent that appears in September, 2011 (Beach Ade)

TABLE 2

MOOD'S MEDIAN TEST

	COLA	**ADE**	**AWAKE!**
June	HIGH	HIGH	HIGH
July	HIGH	LOW	LOW
August	HIGH	LOW	LOW
Septembe	HIGH	MEDIAN	LOW
October	HIGH	LOW	LOW
Novembe	HIGH	LOW	LOW
December	HIGH	LOW	HIGH

The chart tells a good story. Notice that all points for Cola appear in the HIGH group; i.e, each is above the median of 10.0 percent. By contrast, Ade has only HIGH point above the median, while Awake! has only two. If the sample were random, we would expect each drink to have an equal number of HIGH and LOW points.

As with the sign test, the above differences can be tested and examined for statistical significance using online diagnostics. Even if significance cannot be demonstrated, the median test is a useful diagnostic that points the way for a more powerful analysis. It is also easy to explain to a jury

4. **THE RANK SUM TEST**

While the *sign test* and *median test are* simple, they frequently hint of a discrepancy but may lack the full power to establish useful results. An expert can add more clout with the rank sum test, the power hitter of non-parametric statistics. Non-parametric pioneer Frank Wilcoxon originally designed the test to handle a comparison of two variables.

The test was extended in Hyde Park by William Kruskal, W. Allen Wallis, and Milton Friedman to handle three or more variables.

To implement the rank sum test, the percent changes of all product are ranked in one listing; group ranks are then summed. For Table 1, we derive the following

TABLE 3

RANK SUM TEST

	COLA	**ADE**	**AWAKE!**
June	6	9	8
July	5	14	20
August	7	13	21
Septembe	2	11	16
October	4	17	19
November	1	15	12
December	3	18	10
SUM	28	97	106

With a lower rank sum of 28, sales growth at Cola then seems to be ranked near the top more often. Rank sums for Ade (97) and Awake! (106) are less auspicious If the treatment had no effect, the expected sum of ranks would be the same for all products; i.e., 77

The expert must test this difference in rank sums for statistical significance with the high-powered **H** statistic that Kruskal and Wallis derived. With online software, we derive an H = 13.5, which is significant with 95 percent certainty with 2 degrees of freedom.

The rank sum test is more powerful than the sign test or the median test. With regard to the former, the rank sum test allows an explicable comparison of three or more products simultaneously. Compared with the median test, the rank sum test pays more attention to differences in position, being first counts more than slightly above the

median. Because the median test has no such discretion, the rank sum test is more discerning, and more powerful. And juries and judges will still get it.

6.CROSS-VALIDATION

Once a non-random effect is determined for a treatment period, the same diagnostic statistic should be derived in other periods for *cross-validation*. There are many techniques that can be used. Without careful checking, we may have evidence a *Type II error*; we accept an alternative hypothesis when it should have been rejected.

First, check your benchmarks. You selected the benchmarks because you believed that they had similar historic growth patterns to your treatment variable; these patterns were presumably disrupted by the treatment. Now you must prove it. Your appointed diagnostic test should be run on the same variables in the *pre-treatment interval*, and possibly the *post-infringement interval*. If you have chosen proper benchmarks, you should **not** see any evidence of a dominant trend in either outside the treatment period,

Second, check your treatment. For litigation in commercial tort or advertising campaign, effects may be temporary if the treatment stops at some point. It is possible here to examine whether the asymmetric effects that you have found arose from the treatment itself, or from some permanent ongoing shift in circumstances that began in the treatment period. This is best done with a straddle interval that combines pereent changes in each variable from a time before the infringement (e.g., June, 2010) to a corresponding time afterward (June, 2012). If an independent structural change is effective, your diagnostics may still be positive. That said, it is not clear that your treatment variable has the causal effect that you first purported. .

7.ADDITIONAL CONSIDERATIONS

It is not necessary to confine the designated growth benchmarks to only the defendant products. Any variable can serve as a benchmark as long as there is a credible demonstration that the treatment variable has similar growth patterns with the benchmark outside the treatment interval. For example, sales of Choka Cola can be compared with sales from a Cola competitor, or industry sums of sodas or soft drinks, or even a measure of overall economic activity. *However, it is always wise to bring in the*

nearest candidates first, move the analysis outward, and cross-validate results at each stage. .

More statistical power may be possible if the expert can isolate regions, states, or cities where the treatment is preponderant. The expert must ensure that the defendant actually practiced the treatment (e.g., infringing advertisement) in any candidate region. .Foe example, if an advertising campaign were confined to the Southeast, the test is more powerful if so confined. The test would become even more powerful if the expert could consider sales on a state-by-state basis.

After proving causation, an expert may also estimate actual damages or ill-gotten defendant revenues in several ways. Depending on the complexity of the situation, a Court may measure the treatment effect as a simple before-and-after effect — i.e., a simple difference in sales, profits, or earnings that resulted due to the treatment event. To allow for more complex "but for" situations, the "before point" can be trended or otherwise adjusted to reflect the effects of general growth or other variables that may affect growth patterns

7.CONCLUSION

To reiterate from the introduction,[7] attorneys must beware; statistical significance is not the the only reliable indication of causation in a court room. Proper analysis can involve source, content, and subjective evidence.

That said, accurate statistical analysis is useful in the courtroom for three reasons. First, experts may consider whether there is evidence that the defendant's breach has actually caused any damage or enrichment. An integrated approach that merges statistics and further evidence is best.

Second, compared with classical statistical techniques now often found, non-parametric tests have considerable benefits. These tests are more useful with small samples. omitted data, misspecified models, and changing influences.

Third, non-parametric analysis is cheap and easy to explain to juries. Juries will

[7] Supra note 1

understand coin tosses and rank sums as easily as anything else a technician could devise.

---

are too small, critical data can be omitted, models may be incorrectly specified, or observations may be heavily correlated with one another.

REVENUE TOTALS BY MONTH

		COLA	BEACH ADE	AWAKE!
2010	June	$5,036	$9,738	$4,362
	July	5,227	10,093	4,591
	August	5,551	9,110	4,007
	September	5,920	10,239	4,214
	October	5,520	8,039	4,034
	November	5,664	8,709	4,654
	December	9,230	11,299	6,112
2011	January	10,616	11,652	6,992
	February	11,510	11,962	6,278
	March	11,844	10,959	6,410
	April	11,246	13,578	5,899
	May	7,581	13,516	6,001
	June	6,981	11,478	5,230
	July	7,346	10,208	4,106
	August	6,848	9,502	3,272
	September	9,618	11,263	4,033
	October	8,474	7,601	3,609
	November	11,309	8,766	4,940
	December	14,191	10,491	7,091
2012	January	11,276	9,133	6,009
	February	10,951	11,142	6,431
	March	11,001	10,639	6,228
	April	10,376	9,801	5,719
	May	10,226	10,561	6,959
	June	8,703	9,722	5,911
	July	7,022	9,111	4,549
	August	6,510	10,471	3,287
	September	7,610	12,137	3,380

October	7,496	8,862	3,767
November	9,416	8,301	4,101