Honest Hypothesis Testing: A Parable

Michael Stutzer

University of Colorado

To appear in *Advances in Financial Education* (2016)

Honest Hypothesis Testing: A Parable

Abstract

Typical empirical methods training for finance Ph.D students emphasizes theory and application of statistical tools, augmented by study of empirical applications published in the academic journals. But newly minted faculty/authors will also have to cope with opinions of their administrative and academic journal evaluators, something not as heavily stressed in the heavily methods-oriented curriculum. We put a humorous slant on this by providing a parable in which a seemingly simple, standard hypothesis test is subject to the crucible of peer opinion both within and outside the professor's domicile. The parable is used to teach that the interpretation of seemingly simple test statistics is not so simple, and is intended to help instructors augment traditional statistics pedagogy.

Keywords: Hypothesis testing, Bayesian methods, Ph.D econometrics training

INTRODUCTION

An assistant professor of finance is asked by the department chair to analyze some data from a survey of the department's graduates. The department chair was most interested in the response to the survey question: “*Do you wish that you had studied finance somewhere else*”? Of the $N=400$ respondents, 220 answered “no”, while the other 180 did wish they had gone to school somewhere else. The department chair expressed relief that the fraction of “satisfied customers” was higher than $\frac{1}{2}$, so released the data for use by the assistant professor (hereinafter referred to as “The AP”), who planned to use it and other findings in a paper intended for a finance education journal. Of course The AP assumed that the referee would also demand statistical

significance of that finding and any others. Denoting the unobserved population fraction of the satisfied customers by θ , The AP decided to conduct the standard two-way hypothesis test:

$$H_0 : \theta = 1/2 \text{ vs. } H_A : \theta \neq 1/2 \quad (1)$$

Denoting the corresponding sample proportion by $\bar{\theta}$, under the null hypothesis the test statistic t is

$$t = \frac{\bar{\theta} - 1/2}{\sqrt{1/4N}} = \frac{(\frac{220}{400} - .5)}{\sqrt{1/(4*400)}} = \frac{.05}{.025} = 2 \quad (2)$$

standard deviations away from the standard normal distribution's mean of zero, resulting in the two-sided p-value of .0455, and hence statistically significant at the $\alpha = 5\%$ level. Augmented by other findings, the paper was then presented at a finance education conference. The AP expected the conference discussant to spend the designated 5 minutes flashing numerous slides listing many questions, e.g. What pains were taken to ensure that the sample was random and not affected by selection bias?; How was the exact wording of the question determined?; Did you examine conditioning variables that might affect the result, e.g. how many years it took the student to graduate?

The author was surprised to find that the discussant's sole recommendation was the following:

“The author should show a Bayes Posterior Odds Ratio for the null hypothesis.”

Only vaguely recalling a topic like that in an advanced Econometrics class, the author went to see an Old Geezer professor who was still willing to work on occasion. The Old Geezer quickly located a relevant article titled “Testing Precise Hypotheses”, by the statisticians Berger and Delampady (1987). The Old Geezer said, “This kind of thing rarely appears in finance

journals, so I'm not surprised that you didn't include it. I can explain this to you if you like.” He then produced the following derivation:

Using Bayes' Theorem, he wrote

$$Prob[H_0 | \bar{\theta} = .55] = \frac{Prob[\bar{\theta} = .55 | H_0]PriorProb[H_0]}{Prob[\bar{\theta} = .55 | H_0]PriorProb[H_0] + Prob[\bar{\theta} = .55 | H_A]PriorProb[H_A]} \quad (3)$$

The $Prob[\bar{\theta} = .55 | H_0]$ is the likelihood of the observed proportion, evaluated under the null hypothesis. So in this case, it is computed by evaluating the standard normal *density* evaluated at $t = 2$. Use the Excel function “NORM.S.DIST(2,FALSE)” to find that it is .054. The Old Geezer insisted that despite the department chair's implied belief that $PriorProb[H_0] = 0$, the conference discussant would want The AP to specify a prior belief that wouldn't result in a foregone conclusion. They agreed that the discussant would not object to setting $PriorProb[H_0] = 1/2$, so that $PriorProb[H_A] = 1/2$. But this still left unknown the term $Prob[\bar{\theta} = .55 | H_A]$. Fortunately, Berger and Delampady (op.cit.) proposed a way to calculate that. The **average likelihood** for H_A averaged over some density $g(\theta)$ is defined to be

$$m_g(\bar{\theta} = .55) \equiv \int Prob[\bar{\theta} = .55 | \theta]g(\theta)d\theta. \quad (4)$$

Substitute all this into (3), take the reciprocal of both sides, and simplify to obtain:

$$\begin{aligned}
& \frac{1}{\text{Prob}[H_0 | \bar{\theta} = .55]} \\
&= \left[1 + \frac{\text{PriorProb}[H_A]}{\text{PriorProb}[H_0]} \frac{m_g(\bar{\theta} = .55)}{\text{Prob}[\bar{\theta} = .55 | \theta = 1/2]} \right] \\
&= \left[1 + \frac{1}{\frac{\text{Prob}[\bar{\theta} = .55 | \theta = 1/2]}{m_g(\bar{\theta} = .55)}} \right] \\
&\equiv 1 + \frac{1}{B} = \frac{B+1}{B}
\end{aligned} \tag{5}$$

where the **Bayes Factor** B is defined to be:

$$B \equiv \frac{\text{Prob}[\bar{\theta} = .55 | \theta = 1/2]}{m_g(\bar{\theta} = .55)} = \frac{.054}{m_g(\bar{\theta} = .55)} \tag{6}$$

Taking the reciprocal of (5) yields the simple representation for the Bayes posterior probability of the null hypothesis:

$$\text{Prob}[H_0 | \bar{\theta} = .55] = \frac{B}{B+1} \tag{7}$$

So the **Bayes Posterior Odds Ratio** is

$$\text{Odds Ratio} \equiv \frac{\text{Prob}[H_0 | \bar{\theta} = .55]}{\text{Prob}[H_A | \bar{\theta} = .55]} = \frac{\frac{B}{B+1}}{1 - \frac{B}{B+1}} = B. \tag{8}$$

The AP thereby learned that by specifying equal prior probabilities for the null and alternative hypotheses, the discussant's demand to see the Bayes Posterior Odds Ratio would be satisfied by producing the Bayes Factor B .

To compute (4), The Old Geezer proposed specifying a noncommittal uniform, i.e. constant,

density $g(\theta) = 1$ for θ strictly between 0 and 1. He placed $t(\theta) = \frac{.55 - \theta}{\sqrt{1/4 * 400}}$ for each θ

between .01 and .99 (spaced $d\theta = .01$ apart) in a spreadsheet column of 99 cells. In the next

column, he approximated the integral in (4) by using the *NORM.S.DIST*($t(\theta)$, *FALSE*) function

in Excel to evaluate it for $t(\theta)$ placed in the cell to its left, multiplied that by $d\theta = .01$, and summed the resulting column to approximate the integral in (4).

The AP left the Old Geezer's office, eager to duplicate the computations. This resulted in $m_g(\bar{\theta} = .55) \approx .025$. Substituting into (6) found $B = 2.16$. That is, the odds in favor of the null were better than 2:1. As a result, the posterior probability (7) of the null hypothesis was

$$Prob[H_0 | \bar{\theta} = .55] = \frac{B}{B+1} = 2.18 / 3.18 \approx 69\%. \quad (9)$$

“Drat”, thought The AP, “The department chair would have been pleased to read that I statistically significantly rejected the null hypothesis. But the conference discussant's preferred procedure would require reporting something that favors the null hypothesis by better than 2:1 odds.”

In desperation, The AP brought the results to The Old Geezer, along with some soul-searching questions:

First Question: “Isn't the p-value the probability that the null hypothesis is true ?”

First Answer: “No. In fact, the p-value is computed by *assuming* that the null hypothesis is true, so it can't provide evidence about that. In fact, you just told me that 69% is the probability that the null is true, not the far, far, lower p-value and critical value $\alpha = 5\%$.”

Second Question: “Well, if we repeatedly randomly sampled $N = 400$ graduates from the same population, each time re-testing the null hypothesis, isn't the p-value the fraction of times that we would falsely reject the null hypothesis ?”

Second Answer: “No, it can't be that either: each time you re-tested you would wind up with a different p-value, none of which answer that question. You have a Ph.D and are paid more than me, so go do a Monte Carlo simulation to determine that fraction if you think it is so important.”

Third Question: “Well, then what the Hell *is* the p-value ?”

Third Answer: “Neyman and Pearson weren't shlubs, and they despised Ronald Fisher's p-value. Maybe you should stop worshipping it. If you want to know what the critical value $\alpha = 5\%$ means, they would assert that if you repeatedly randomly sampled $N = 400$ graduates with replacement from the same population, each time computing a $1 - \alpha$ confidence interval from the estimate $\bar{\theta}$ made with that sample, then about $1 - \alpha$ of those confidence intervals would contain the unknown value of θ . If you ever do that Monte Carlo study I suggested, you could check that claim, too. In the meantime, I'll show you a simple approximation you can use to approximate the associated posterior probability of the null hypothesis, when you assign 50-50 prior odds to it.”

The Old Geezer limped back to the whiteboard to write that when $g(\theta)$ is a uniform density of width c much bigger than the standard error of estimate $\sigma_{\bar{\theta}}$, the approximate Bayes Factor B

and associated posterior probability of the null $\frac{B}{B+1}$ are:

(Startz, 2014)

$$B \stackrel{c \gg 2\sigma_{\bar{\theta}}}{\approx} c \text{NORM.S.DIST}(t, \text{FALSE}) / \sigma_{\bar{\theta}}$$

$$\text{Prob}[H_0 | \bar{\theta}] = \frac{B}{B+1} \stackrel{c \gg 2\sigma_{\bar{\theta}}}{\approx} \frac{\text{NORM.S.DIST}(t, \text{FALSE})}{\text{NORM.S.DIST}(t, \text{FALSE}) + \frac{\sigma_{\bar{\theta}}}{c}} \quad (10)$$

The AP took formulae (10) back to the computer. Substituting the test statistic $t = 2$, the standard error of estimate $\sigma_{\bar{\theta}} = .025$, and the uniform prior's width $c = 1$ into (10) gives $B \approx 2.16$ implying 69% probability for the null. These formulae, found in Startz (op.cit.), produced results very close to what The Old Geezer found by approximating the integral on his spreadsheet. Using (10), The AP quickly discovered that setting $c = .25$ instead of $c = 1$ would yield a posterior probability of the null equal to only 35% , i.e. the alternative hypothesis would be favored by nearly 2:1 odds. That would surely please the department chair. But what if the discussant or a like-minded journal referee suspected that the corresponding $[\cdot375, \cdot625]$ prior uniform density had been “c-hacked” to produce rejection of the null?

Split the Sample to Test More Proportions

Not wanting to waste any more time on esoteric statistical philosophy and associated Excel tomfoolery, The AP took a closer look at the rest of the survey, finding that responses were gathered from an equal number of men and women. The AP discovered that 114 of the 200 men were what the department chair deemed to be satisfied customers, while only 106 of the 200 women were. The AP quickly ran the test:

$$H_0 : \theta_{men} = 1/2 \text{ vs. } H_A : \theta_{men} \neq 1/2 \quad (11)$$

The standard test statistic is $t = \overline{\theta}_{men} - 1/2 = \frac{114}{200} - 1/2 = .07$, which under the null is

approximately normally distributed with mean 1/2 and standard deviation $\sqrt{\frac{1}{4 * 200}} = .0354$.

The test statistic is thus 1.98 standard deviations from its hypothesized mean, i.e. statistically significant at the $\alpha = 5\%$ level (p-value = .048). Then The AP ran the same test on the women,

finding $t = \overline{\theta_{\text{women}}} - 1/2 = \frac{106}{200} - 1/2 = .03$, which under the null is approximately normally

distributed with mean $1/2$ and standard deviation $\sqrt{\frac{1}{4 * 200}} = .0354$, i.e. too variable for

statistical significance (p-value = .396). The AP went to the Old Geezer with the good news:

“It is the men who are the statistically significantly satisfied customers, not the women. This in accord with the well-accepted Interplanetary Theory of Gender Differences: *Men are from Mars while Women are from Venus* (Gray, 2012). The Venusians communicate more effectively with each other than the Martians do. So our female graduates must have more frequently shared their horror stories of pointless, pedantic PowerPoint slides flashed in class. This resulted in somewhat fewer satisfied customers than the taciturn men had. Moreover, the result is more statistically significant than our overall finding (p-value = .034 vs. .046).”

Eager to have a paper under review, The AP submitted this to a journal that had previously published related studies, without problematic Bayesian Hypothesis Tests. A surprisingly short but encouraging referee report arrived a few months later:

“As one of its co-creators, I am impressed with your interest in testing *The Interplanetary Theory of Gender Differences*. I only have one substantive request: please also use your data to conduct a *Difference of Two Proportions test* for the men and women.”

The AP thought, “More make-work. I already showed that the null was rejected for the men, but not for the women, in accord with the well-accepted (not to mention best-selling) Interplanetary Theory of Gender Differences. Now I have to perform another test. Well, it’ll be worth it when it is accepted for publication.”

After a night's celebration, The AP found a good applied statistics text -- *Applied Statistics in Business and Economics* (Doane and Seward, 2012) -- and found a description on pp.410-414 to ensure that the Difference of Two Proportions test was run properly. The following was quickly computed:

$$H_0 : \theta_{men} = \theta_{women} \text{ vs. } H_A : \theta_{men} \neq \theta_{women} \quad (12)$$

$$t = \frac{.57 - .53}{\sqrt{\frac{114+106}{400} * (1 - \frac{114+106}{400}) * (\frac{1}{200} + \frac{1}{200})}} = .804 \quad (13)$$

$$p - value = .421 \quad (14)$$

The horror! The Difference of Two Proportions test adds the separate variances from the two separate tests. The resulting standard error is large enough to make the 4% difference in proportions statistically *insignificant*.

Multiple Comparisons

In desperation, The AP went to The Old Geezer with the bad news:

“I can't believe this crap. I did a perfectly good test that showed men were more likely than not to be satisfied customers, while women weren't. The referee liked it, and asked for only one more test. But that test says that men and women weren't actually different, which flies in the face of the well-accepted Interplanetary Theory of Gender Differences. What do I do now ?”

The Old Geezer suggested that The AP look for other schools' surveys of finance graduates, to see if they offered any support for the Interplanetary Theory of Gender Differences.

The AP tasked a research assistant with calls to other finance departments, and eventually located four other departments that had conducted similar surveys, that also indicated the gender of respondents. Even better, the gender differences and sample sizes were both larger.

The AP ran the Difference of Two Proportions tests for each of the four additional samples, finding p-values of .03, .08, .04, and .02, respectively. The AP's reply to the referee read as follows:

“I recognized that gender-specific findings from a single survey could be idiosyncratic. So I went to considerable effort to locate four other finance departments that had conducted similar surveys. Fortunately, they had much larger sample sizes. We applied your recommended Difference of Two Proportions Test to each, and found the gender difference to be statistically significant ($\alpha = .05$) in 3 of the 5 schools (their respective p-values were .421, .08, .04, .03, and .02). This provides some support for the Interplanetary Theory of Gender Differences.”

The referee's reply was swift.

*“I commend your effort to find additional test data, and have no reason to doubt that it was randomly sampled independently from your own school's data. But I must question your interpretation of the significance level, $\alpha = 5\%$. $1 - \alpha = 95\%$ was the probability that you wouldn't falsely reject a true null about your own school. But now you have 5 separate null hypotheses. The probability that you won't falsely reject **any** of them when they are all true is $(1 - \alpha)^5 = .774$. So it is **as-if** you had raised your α from 5% to 22.6% . But this can be fixed, by controlling what is called the **family-wise error rate**. You must use a different level α_B that satisfies*

$$(1 - \alpha_B)^5 = 1 - \alpha \quad (15)$$

*Now take the log of both sides, and because reasonable α are pretty close to zero, you can use the approximation $\log(1 - x) \approx -x$ to find the **Bonferroni corrected** $\alpha_B = \alpha / 5 = 1\%$. That's the good news. The bad news is that none of your p -values are below 1%. Under the circumstances, I must recommend rejection of your paper."*

The AP was stunned by the sudden reversal of fortune, thinking:

"Perhaps The Old Geezer has a rejoinder to this. He knows this stuff, but I can't go back there until I do that Monte Carlo simulation he wanted. I don't have time to go to Monte Carlo. Even if I did, I don't have enough travel funds to cover the cost. Doesn't The Old Geezer care that I need to publish or perish?"

Facing the Music

The AP decided that academic Hara-kiri was the honorable way to end this, so made an appointment to see the department chair. After The AP described the fruitless process, the department chair said:

"I appreciate the Bayesian approach to testing, but was only interested in knowing whether *more than* half the graduates were satisfied customers. So why did you formulate the test with a point null that exactly half were? You should test:"

$$H_0 : \theta \leq 1/2 \text{ vs. } H_A : \theta > 1/2 \quad (16)$$

The AP agreed to do that. Looking back at equation (3), The AP saw that it was now essential to specify both $Prob[\bar{\theta} = .55 | \theta \leq \frac{1}{2}]$ and $Prob[\bar{\theta} = .55 | \theta > \frac{1}{2}]$. Again it made sense to assume

a uniform prior distribution over all proportions θ between zero and one, in which case the Bayes Factor B would no longer be given by (6), but would become

$$B = \frac{\text{Prob}[\bar{\theta} = .55 | \theta \leq \frac{1}{2}]}{\text{Prob}[\bar{\theta} = .55 | \theta > \frac{1}{2}]} = \frac{\int_0^{1/2} \text{NORM.S.DIST}\left(\frac{.55 - \theta}{\sqrt{\theta(1 - \theta) / 400}}, \text{FALSE}\right) d\theta}{\int_{1/2}^1 \text{NORM.S.DIST}\left(\frac{.55 - \theta}{\sqrt{\theta(1 - \theta) / 400}}, \text{FALSE}\right) d\theta} \quad (17)$$

$$\approx \frac{.00087}{.02391} = .036$$

The integrals were approximated by the same spreadsheet technique The Old Geezer used.

Using (7), the posterior probability of the null in (16) is

$$\text{Prob}[H_0 | \bar{\theta} = .55] = \frac{B}{B + 1} = .036 / 1.036 \approx 3.5\% \quad (18)$$

The AP emailed the result to the conference discussant who started this journey:

“Look at the difference between (18) and (9)! All I did was change from a point null at $\frac{1}{2}$ to an interval null less than $\frac{1}{2}$, and the test result is radically reversed. The posterior probability that the null is true decreased from 69% to only 3.5%, which is pretty darn close to the much-maligned p-value in my original test ($p = 4.6\%$)! I plan to add the department chair as a co-author when I submit this for publication. Sure I did all the hard work, but nobody cares how many co-authors there are, so it can only help my career.”

Relevance For Academic Finance Research

That ends this parable on an optimistic note, but I hope it doesn't end your advocacy for meaning and truth in empirical work. The problems highlighted can and have appeared in academic finance research. Let's consider the conflicting results between classical hypothesis tests and Bayesian Posterior Odds ratios that were highlighted in Section 1. French [p.59, 1980] examined day-of-the-week stock market returns, and used classical hypothesis tests to

conclude that “the hypothesis that Monday's expected return was positive can be rejected during any five-year period at a 5 percent significance level.” Yet Connolly (1991, pp.77-78) calculated Bayesian Posterior Odds ratios for two-tailed tests of the hypothesis (1), i.e. that the fraction of positive Monday returns was $\frac{1}{2}$, over many of the same years examined by French, and concluded that “Posterior Odds favor the hypothesis of an equal percentage of positive and negative returns except in a few years.” This is the phenomenon illustrated in Section 1 herein. Connolly (op.cit.) did not report posterior odds ratios for the hypothesis (16), but had he done so, he might have found results analogous to those reported in Section 4 herein.

Now let's consider the multiple comparison problems highlighted in Section 3. You may think that when only one data set is used, the analysis can't be affected by the multiple comparisons problem that plagued The AP's analysis of the four additional surveys' findings. But suppose you obtained a data set containing dozens of possible variables that might help explain the variation in a single independent variable. You run the usual ad-hoc (usually linear) multiple regression, which uncovers one or more variables with p-value lower than 5%. Perhaps “the literature” has an “explanation” for the non-zero coefficient(s) on the variable(s), although that (social science) theory likely will not have the inherent testability (i.e. easily falsified) and prior plausibility of a respected physical theory (e.g. Einstein's Relativity Theory). If the social science literature does not provide a suitable explanation, you attempt to devise one. Perhaps your subsequent working paper described this process in reverse order, starting with a section titled “Hypothesis Development” that includes this explanation along with alternatives to be tested in an empirical section that follows. In that section, you further explored the nature of the relationship with more modeling and testing, in an exploratory process contingent on the results of hypothesis tests reported earlier in the paper. This *modus operandi* is so common that

few give it a second thought. But statisticians have identified this “garden of forked paths” as a source of multiple comparison problems (Gelman, 2014). The AP did this when turning to the gender difference in the data, only after having seen the problematic posterior odds ratio, but did not reduce the critical value α in light of the multiple comparison.

In finance, this process has been writ large by countless publications advocating candidates for additional factors in multiple linear factor models, each of which purports to explain the cross section of stock returns better than other linear factor models (e.g. the CAPM). A recent paper carefully analyzed the multiple comparison problems resulting from this decades-long mining of ubiquitous returns databases, and concluded:

“The estimation of our model suggests that a newly discovered factor needs to clear a much higher hurdle, with a t-ratio greater than 3.0. Echoing a recent disturbing conclusion in the medical literature, we argue that most claimed research findings in financial economics are likely false.” (Harvey, Liu, and Zhu, 2016)

The ubiquitous “robustness tests” that are reported in regression-based articles that appear in academic finance journals are the typical response to those concerns. But the usual process only displays a small fraction of the plausible alternative specifications of analogous independent and dependent variables, and functional forms that could have been mandated in the reviewing cycles. Three prominent business school faculty have proposed a promising alternative. Simmons, Nelson, and Simonsohn (2015) propose using computer power to simultaneously examine thousands of such alternative specifications. They devised graphical tools to display the array of results, and utilized a clever bootstrap procedure to test whether or not an author's

highlighted effect is outside the range of results that would be expected using thousands of alternative specifications that by construction, contain no such effect.

I hope that you will implement and teach such procedures, and that the academic finance journals will eventually insist on them. Until then, articles in journals of high reputation should not receive a free pass. As noted by Velickovic (2015):

“A critical approach must exist regardless of the journal in which the paper is published.”

In addition to raising awareness during empirical methods training, faculty can provide other concrete recommendations for more honest hypothesis testing when devising exercises for students. For example, an instructor could describe the nature of a multivariate data set before providing it. Then before providing the actual data, the instructor could make students commit to performing one (and only one) multiple linear regression and one specific hypothesis test for a single, pre-determined coefficient $H_0 : \beta = \beta_0$ vs. $H_A : \beta \neq 0$, stating a plausible motivation for testing that specific hypothesis. At a minimum, insist that the students choose an interval of length c centered at β_0 , with c equal to 10 times the reported standard error $\sigma_{\bar{\beta}}$ of the estimated coefficient, to enable use of (10). Then train the students to plug the t -statistic for the null, c , and $\sigma_{\bar{\beta}}$ into (10) to compute the Bayes Factor and corresponding posterior probability of the null hypothesis. In addition to reporting the usual t -statistic, counsel the students to report that this posterior probability arose both from the data and the prior uniform density over the range $[\beta_0 - c/2, \beta_0 + c/2]$.

References

- Berger, J. O. and Delampady, M. “Testing Precise Hypotheses,” *Statistical Science*, 2 (1987), 317-352.
- Connolly, R. A. “A Posterior Odds Analysis of the Weekend Effect,” *Journal of Econometrics*, 49 (1991), 51-104.
- Doane, D. and Seward, L. *Applied Statistics in Business and Economics* (New York, 2012), 4th Edition, McGraw Hill/Irwin.
- French, K. R. “Stock Returns and the Weekend Effect,” *Journal of Financial Economics* 8 (1980), 55-69.
- Gelman, A. “The Statistical Crisis in Science,” *The American Scientist* 102 (2014), 460-465.
- Gray, J. *Men are from Mars, Women are from Venus: The Classic Guide to Understanding the Opposite Sex* (New York, 2012), Harper.
- Harvey, C., Liu, Y. and Zhu, H. “....and the Cross Section of Expected Returns,” *Review of Financial Studies*, 29 (2016), 5–68.
- Simmons, J. P., Nelson, L. D. and Simonsohn, U. “Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications,” Wharton School Working Paper, Nov. 2015.
- Startz, R. “Choosing the More Likely Hypothesis,” *Foundations and Trends in Econometrics*, 7 (2014), 119-189.
- Velickovic, V. “What Everyone Should Know About Statistical Correlation”, *American Scientist*, 103 (2015), 26-29.