2020

Clear-Sighted Statistics: Module 19: Wrapping Up

Edward Volchok

CUNY Queensborough Community College

Clear-Sighted Statistics: An OER Textbook

Module 19: Wrapping Up

> “The first principle is that you must not fool yourself—and you are the easiest person to fool.”[1]
>
> -- Richard P. Feynman
> Theoretical Physicist
> *“Cargo Cult Science”*
> 1974

> “One of the most frustrating aspects of the journal business is the null hypothesis. It just will not go away.… It is impossible to drag authors away from their *p* values, and the more zeros after the decimal point, the harder people cling to them. It is almost as if all the statistics courses in the world stopped after introducing Type I error….Perhaps *p* values are like mosquitoes. They have an evolutionary niche somewhere and no amount of scratching, swatting, or spraying will dislodge them….investigators must learn to argue for the [practical] significance of their results without reference to inferential statistics.”[2]
>
> -- John P. Campbell
> Editor, *Journal of Applied Psychology*
> “Some Remarks From the Outgoing Editor”
> 1982

> “Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas [personality and social psychology] is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worse things that ever happened in the history of psychology.”[3]
>
> -- R. Chris Fraley
> Psychologist, University of Illinois at Urbana-Champaign
> Cited in *The Cult of Statistical Significance*

I. Introduction

In this, our final, module, we will:

1) Recap the key lessons of *Clear-Sighted Statistics*.

2) Discuss what we did not cover and what would be covered in more advanced courses.

3) Discuss the growing criticism of NHST as we report on how the science of statistics is advancing and how these developments will *probably* change the ways inferential statistics will be conducted in the future.

II. Statistics in a "Post-Truth" Society and the Need for Informed Skepticism

Throughout *Clear-Sighted Statistics*, we have differentiated descriptive statistics or exploratory data analysis from inferential statistics. We noted that both descriptive and inferential statistics are prone to errors and distortions. Knaves willfully distort data for self-interested ends. Fools misapply statistical techniques and arrive at dubious conclusions because they do not understand what they are doing. The damage caused by knaves and fools plagues us and may even lead to destructive cynicism. But, when properly performed, statistical analysis can help us understand more clearly what the data means and establish a more solid foundation for our findings and our decisions.

A key lesson you should learn from *Clear-Sighted Statistics* is that whenever you conduct a statistical analysis, you should do so with transparency and honesty. Transparency means that we should tell our audience how the data were acquired, what techniques were employed, and what questions remain unanswered. Honesty means that we should clearly state the limitations of our approach and our findings.

The late John Wilder Tukey, who taught at Princeton University and worked at Bell Laboratories, was one of history's greatest descriptive statistics experts. He was aware of how descriptive statistics in general and data visualization (charts) in particular can mislead us. In Module 4, we showed how knaves and fools distort data with charts. The following quote by Dr. Tukey's highlights his concern about the misuse of charts: "Visualization is often used for evil—twisting insignificant data changes and making them look meaningful. Don't do that crap if you want to be my friend. Present results clearly and honestly. If something isn't working—those reviewing results need to know."[4] Dead men like Professor Tukey do not need our friendship, but a free, democratic society needs its citizens to be able to distinguish fact from fiction. To be able to do this, citizens need statistical literacy because, as we have shown, numbers can trick us. Unscrupulous people will try to fool the innumerate.

Inferential statistics is also prone to distortions by fools and knaves. Inferential statistics estimates unknown population parameters on the basis of sample statistics. Whenever we draw a sample, we risk *random sampling error*, that is when the sample statistics do not equal the population parameter, which is usually unknown. We cannot escape sampling error. It is a natural consequence of drawing samples from populations. Sampling error is not the result of human error. Confidence intervals and NHST can help us measure the risk of sampling error. Poorly conducted studies, however, are also vulnerable to a host of *systematic errors*, which result from human error. Systematic errors are often a more serious problem than sampling errors.

Whenever we deal with inferential statistics, we must remember that these analyses are based on probability. When we estimate parameters using confidence intervals at a 95 percent confidence level, we are saying that if we conducted repeated surveys, the parameter will be within the confidence interval for 95 percent of the surveys. Facts based on samples are probabilistic. We do not have 100 percent certainty. This is why we never consider the null hypothesis to be true when we fail to reject it. Similarly, when we reject the null hypothesis, we are not declaring the alternate hypothesis true. When rejecting the null hypothesis, all we are saying is that the data do not support the null hypothesis. Anything stronger is folly that divulges the author's hubris.

While often misinterpreted, p-values merely use probability theory to measure how compatible the data are with the null hypothesis. Rejecting or failing to reject the null hypothesis says nothing about the size of the effect. In addition, statistical significance does not imply any practical, real-world significance. We could fail to reject the null hypothesis even though the effect has practical, real-world importance. As we have shown, we could also reject the null hypothesis when the effect has no practical implication. It is only with multiple replications of a study's results that a research hypothesis might begin to rise to the level of a theory.

To repeat, **facts are not 100 percent certain**. This should lead us to two attitudes: Humility and skepticism. As for humility, we must not overstate the importance of our findings. And, we must remain skeptical and open to new information that might change our findings. One of the best descriptions of scientific skepticism comes from data scientist Cathy O'Neil, author of *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democra*cy. In in her pamphlet, *On Being a Data Skeptic*, Dr. O'Neil writes:

> A skeptic is someone who maintains a consistently inquisitive attitude toward facts, opinions, or (especially) beliefs stated as facts. A skeptic asks questions when confronted with a claim that has been taken for granted. That's not to say a skeptic brow-beats someone for their beliefs, but rather that they set up reasonable experiments to test those beliefs.[5]

Following the advice of the late Richard P. Feynman, delivered in his famous commencement address to the 1974 graduating class of the California Institute of Technology, we must be aware of the risks of fooling ourselves. For decades, statisticians have been debating whether inferential statistics as it is taught and conducted, is blinding us to the real implications of our data. We will turn to the short-comings of NHST in Section IV of this module.

III. Where We Go From Here: What is Covered in More Advanced Courses

Statistics is a multifaceted discipline fundamental to so many fields that its breadth cannot be adequately covered in an introductory course. Most disciplines in the social sciences and many in the natural sciences use statistical techniques even though they often employ widely different approaches. The broad application of statistics was noted by Tukey, who famously said, "The best thing about being a statistician is that you get to play in everyone's backyard."[6]

Clear-Sighted Statistics was written to introduce students to the basic statistical methods used in business and the social sciences. More advanced statistics courses delve deeper into inferential statistics as well as other topics. Let's discuss some of these new areas that would be introduced before we discuss some of the more advanced concerns about inferential statistics that are currently being discussed by practitioners.

A) New Topics

New topics include: 1) Time series analysis and forecasting, 2) Decision Theory, 3) Statistical Process Control, and 4) Meta-Analysis.

Time series analysis involves analyzing data that have been collected over time. You will recall that this type of data is called *longitudinal* data. A goal of time series analysis is forecasting the future, which is a serious issue for decision-makers in business and government.

Time series analysis is based on four types of trends:

1. **Secular Trends:** The long-term non-periodic variation in the longitudinal data. The timescale used is a key determinant on whether longitudinal data are perceived as a secular trend. Examples of secular trends include: 1) The aging of the population in advanced post-industrial countries, 2) Expansion of digital technologies, 3) The reliance on fossil fuels like coal, oil, and natural gas, and 4) Trends in global warming.

2. **Cyclical Variations:** These are oscillating movements in time series data. The business cycle with its swings between boom and bust are a classic example of cyclical variations.

3. **Seasonal Variations:** These are repeated patterns of changes in time series data within a year. Ice cream sales on the Coney Island boardwalk, for example, have a distinct seasonal variation. The number of people employed at ski resorts or sales of Christmas trees are also longitudinal data with strong seasonal variations.

4. **Irregular or Random Variations:** These are variations in the time series data that do not follow a predictable model and are, therefore, not predictable. An example of irregular or random variation would be the impact on the American economy of impeaching the President of the United States.

With time series data, we often calculate *moving* or *rolling averages* in an attempt to smooth random fluctuations in the data so that trends might be easier to detect. With a moving average, individual observations are adjusted by the mean of that observation and the observations that precede and follow it. Moving averages are often used in financial analysis. Moving averages are considered a *trend-following* or *lagging* indicator because they are based on historical data.

Decision Theory deals with a branch of statistical science that evaluates decision-makers' choices based on the possible outcomes that might occur in an uncertain future. With statistical decision theory, statistical information informs the decision-maker of the uncertainties—the probabilities—involved in a decision. In Module 7, Basic Concepts of Probability, we briefly discussed Pascal's Wager as a rudimentary example of decision theory. Decision theory is a major topic in graduate-level management curriculum.

Statistical Process Control, SPC, is a collection of techniques used to improve the quality of manufacturing processes. W. Edwards Deming was a leading innovator in SPC. One of the best known SPC techniques, *Six Sigma* (6σ), was developed by William B. Smith at Motorola in the 1980s. *Six Sigma* is a data-driven process that seeks to reduce manufacturing errors to no more than 3.4 out of a million randomly selected production units. The name, *Six Sigma*, comes from the fact that the error goal of one in 3.4 million would be six standard deviations from the mean.[7]

Meta-Analysis is a collection of quantitative procedures that synthesizes the findings from a review of the research on the topic under investigation. We will briefly discuss meta-analysis in Section IV.

B) A Deeper Investigation Into Inferential Statistics

1) Effect Size, Statistical Power, and the Probability of Type II Errors

Effect size, practical versus statistical significance, statistical power, and the probability of Type II errors receive limited attention in introductory statistics courses if they are even mentioned at all. Advanced statistics courses would delve more deeply into these topics. NHST focuses on whether an effect exists (the alternate hypothesis) or whether it does not exist (the null hypothesis).

In contrast to most introductory textbooks, *Clear-Sighted Statistics* presented an elementary discussion of effect sizes. We used the tables for the interpretation of effect sizes which Jacob Cohen cautiously introduced in his ground-breaking book, *Statistical Power Analysis in the Social Sciences*. These thresholds have become the standard way to interpret the magnitude of effect size. Gene Glass, one of the leading developers of meta-analysis and an effect size theorist, argued against reducing effect size to “tee shirt” sizes. Glass and his co-authors wrote:

> There is no wisdom whatsoever in attempting to associate regions of the effect size metric with descriptive adjectives such as “small,” “moderate,” “large,” and the like. Dissociated from a context of decision and comparative value, there is little inherent value to an effect size of 3.5 or .2. Depending on

> what benefits can be achieved at what cost, an effect size of 2.0 might be "poor" and one of .1 might be "good."[8]

There is much more to learn about effect sizes. There are many types of effect size each with their own advantages, disadvantages, and applications. A more nuanced interpretation of effect size than the one presented in *Clear-Sighted Statistics* should be studied in advanced courses because the magnitude of the effect size helps us determine whether our findings have practical significance. Practical significance focuses on a very important question: Do the findings have real world importance? In addition, we did not develop confidence intervals for effect sizes, which some contemporary statisticians argue is very important.

2) ANOVA Tests:

Advanced statistics courses would explore more sophisticated ANOVA analyses:

1. **One-Way Repeated Measures ANOVA:** These ANOVA tests make repeated measures of over time.

2. **Two-Way Anova Without Replication:** An ANOVA Test with two sets of independent variables or treatments. Excel's Data Analysis ToolPak can conduct this test.

3. **Two-Way Anova With Replication:** An ANOVA Test with two sets of independent variables or treatments. Replication refers to whether the researcher is replicating the test with multiple groups. Excel's Data Analysis ToolPak can conduct this test.

4. **Factorial ANOVA:** These tests are similar to two-way ANOVA test with additional independent variables, treatments, or factors.

5. **MANOVA:** Multivariate Analysis of Variance: An ANOVA test with more than one dependent variable.

6. **ANCOVA (Analysis of Covariance):** An extension of ANOVA used to determine whether the treatments are equal across independent variables.

7. **MANCOVA (Multivariate Analysis of Covariance):** An extension of ANCOVA for multiple dependent variables.

8. **Kruskal-Wallis H test:** A nonparametric version of a One-Way ANOVA test.

3) Regression:

While *Clear-Sighted Statistics* covered simple linear regression, advanced statistics courses focus on more sophisticated types of regression. These include:

1. **Multiple Regression:** Linear regression for modeling the relationship between one dependent variable and more than one independent or predictor variables.

2. **Spearman's Rho:** A nonparametric test of the strength of the association between two variables.

3. **Kendall's Tau:** A nonparametric test of the strength of the association between two ordinal-level variables.

4. **Logistic Regression:** A regression model used when the dependent variable is binary: Either/Or, Yes/No, etc.

5. **Multinomial Logistic Regression:** A logistic regression model with more than two outcomes.

6. **Structural Equation Modeling:** Very sophisticated models that use mathematical and computer algorithms to construct causal models.

4) Nonparametric Techniques

There are a variety of sophisticated tests used by analysts to determine whether parametric tests like z-tests, t-tests, or ANOVA tests are appropriate. These tests would be covered in more advanced statistics courses. Dedicated statistical software, like SPSS, makes running these tests very easy. More advanced statistics classes would introduce dedicated statistical software like SPSS, Stata, or R to test the assumptions of parametric tests. Should our data fail to meet the requirements for parametric tests, we would use the appropriate nonparametric test. We should consider using nonparametric techniques when:

- The data are not normally distributed.
- The sample size is too small to run parametric tests (of course, small samples reduce statistical power and increase the risk of Type II errors).
- The data contain outliers that cannot be removed.
- The data are heavily skewed, and as a result, a decision is made to use the median instead of the mean.

While nonparametric tests have less statistical power than parametric tests, they are more robust. Recall that robustness means the test provides useful results even when one or more key assumptions are violated.

While *Clear-Sighted Statistics* covered only one nonparametric technique, chi-square tests, advanced statistics courses would cover other nonparametric tests. Table 1 shows some of the basic parametric tests and their nonparametric equivalents.

Table 1: Parametric Tests and Their Nonparametric Equivalent

Parametric Test	**Nonparametric Tests**
One-sample z-test, One-sample t-test	Sign test
One-sample z-test, One-sample t-test	Wilcoxon Signed Rank test
Two-sample t-test for independent means	Wilcoxon-Mann-Whitney test
One-way ANOVA test	Kruskal-Wallis test and Mood's Median test
Two-way ANOVA test	Friedman test
Coefficient of Correlation	Spearman Rank Correlation

5) Bayesian Inference:

We briefly touched on Bayes' Theorem when we reviewed probability. Bayesian inference, however, is a sophisticated topic that might be covered in more advanced statistics courses. Historically, statisticians tend to be a quarrelsome lot. Tukey joked that the "collective noun for a group of statisticians is a quarrel."[9] One of the longest and most acrimonious debates in statistics has been between the frequentists and the Bayesians. Nearly all introductory statistics textbooks approach statistics from a frequentist orientation, which is based on objective (classical and empirical) probability. Confidence intervals and the Fisher/Neyman-Pearson NHST are central to frequentist techniques. Bayesian inference is based on Bayes' Theorem and subjective probability.

While frequentists still dominate the science of statistics, Bayesian inference has been slowly gaining acceptance since the 1970s. Its advocates see Bayesian inference as either a useful supplement to frequentist statistical inference, or its replacement.

We have spent a good deal of time on frequentist NHST, which has become the *sine qua non* of inferential statistics and the cornerstone of most social sciences. Yet, NHST is not widely used in physics, chemistry, or biology.

Let's devote a couple of paragraphs to Bayesian inference. This discussion is based on two sources: 1) Ben Lambert's introductory textbook, and a wonderful open access peer-review monograph written by Alonso Ortega and Gorka Navarrete.[10]

Bayesian inference uses Bayes' Theorem to update evidence in support of *both* the null and alternate hypotheses as more information becomes available. Advocates of Bayesian inference contend that this approach allows us to move away from the dichotomous frequentist approach that requires either rejecting or failing to reject the null hypothesis. Its advocates contend that using Bayesian inferences gives researchers a better perspective on the data and the extent to which it supports the null hypothesis *or* the alternate hypothesis. This is something traditional NHST does not do. As we discussed in Module 13, NHST focuses on whether the data falsifies the null hypothesis. A central feature of NHST is the p-value, which is evidence against the null hypothesis or what one philosopher of science said is the degree to which the data are embarrassed by the null hypothesis.[11] P-values provide evidence against the null hypothesis. It never produces evidence in favor of the null or alternate hypotheses.[12]

According to the Bayesians, frequentist NHST has three serious flaws:[13]

1) It only provides evidence against the plausibility of the null hypothesis, but fails to provide any evidence in favor of the alternate hypothesis.

2) Its inferences are made on hypothetical data distributions (z-, t-, F-, or χ^2 distributions, among others.) instead of being based on actual data.

3) It does not provide clear rules for stopping data collection and as a result any null hypothesis can be rejected when the sample is large enough.

The advantage of Bayesian inference , its advocates argue, is that our degree of belief in the null *and* alternate hypotheses—our "prior knowledge"—is updated in light of new data. Researchers, therefore, are encouraged to think about the magnitude of evidence that supports the existence of an effect, instead of a dichotomous way of thinking where an effect either exists or does not exist.

Bayesian inference obtains information from three sources:

1) A model that specifies how latent parameters (φ) generate data (D).

2) Prior information about those parameters.

3) The observed data (likelihood).

These sources lead to the construction of *Bayes Factors*, which are the ratio of the likelihood, or probability of the alternate hypothesis to the probability of the null hypothesis. The Bayes Factor can be interpreted as the strength of evidence for the competing hypotheses.

Equation 1 shows the formula for calculating Bayes Factors:

$$Bayes\ Factor = \frac{P(Data|H_1)}{P(Data|H_0)}$$

Equation 1: Bayes Factor Equation

Table 2 shows how Bayes Factors are interpreted:[14]

Table 2: Bayes Factor Interpretation.

Bayes Factor	Interpretation
> 100	Extreme evidence for the alternate hypothesis
30 – 100	Very strong evidence for the alternate hypothesis
10 - 30	Strong evidence for the alternate hypothesis
3 - 10	Moderate evidence for the alternate hypothesis
1 - 3	Anecdotal evidence for the alternate hypothesis
1	No evidence
1/3 - 1	Anecdotal evidence for the null hypothesis
1/3 – 1/10	Moderate evidence for the null hypothesis
1/10 – 1/30	Strong evidence for the null hypothesis
1/30 – 1/100	Very strong evidence for null hypothesis
<1/100	Extreme evidence for the null hypothesis

A major obstacle to Bayesian inference, besides the hostility of frequentists like Ronald A. Fisher, was that it requires complex calculations. Starting in the 1990s, however, Bayesian software was introduced. Here are just a few statistical applications for Bayesian analysis: BayesiaLab, JAGS, JASP, Stan, and WinBugs.

IV. Where is Statistics Going? The Many Second Thoughts About NHST

We may be witnessing a *paradigm shift* in inferential statistics. In 1962, philosopher of science, Thomas S. Kuhn coined the term *paradigm shift* in his widely read book, *The Structure of Scientific Revolutions*.[15] A paradigm shift is the messy way important changes in science happen when the scientific community fundamentally alters accepted thinking and methods. According to Kuhn, there are four stages to a paradigm shift:

1) **Normal Science:** In this stage the dominant paradigm is active and widely supported. Kuhn's examples of normal science include Newtonian physics, caloric theory, and the theory of electromagnetism. The dominant paradigm defines how science is conducted.

2) **Extraordinary Research:** The dominant paradigm becomes suspect when researchers find anomalies. This throws the scientific discipline into a state of crisis. "Confronted with anomaly or with crisis," Kuhn declares, "scientists take a different attitude toward existing paradigms, and the nature of research changes

accordingly."[16] In essence, scientists begin to experiment with new ideas and new methods.

3) **Adoption of a New Paradigm:** Scientists conducting extraordinary research eventually develop a new paradigm. This is the messy stage. Many scientists refuse to adopt the new paradigm. To illustrate this point, Kuhn quotes Max Planck, the German theoretical physicist, "…a new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents die, and a new generation grows up that is familiar with it."[17] The adoption of a new paradigm takes time because the older scientists have to die off.

4) **Aftermath:** The new paradigm becomes dominant. When dominant, it becomes institutionalized; which is to say, it guides the conduct of science and how the discipline is taught to students.

A) The NHST Paradigm:

Is the science of statistics now in the midst of a paradigm shift? Is NHST on its way out, ready for the dustbin of history like the geocentric model of the universe that placed Earth at its center?

The NHST paradigm was established in the mid-1920s and 1930s. As practiced in the second half of the twentieth century and the first part of the twenty-first century, NHST is an amalgam of two different and incompatible approaches. The first approach is Ronald A. Fisher's *Significance Testing*, which gave us the null hypothesis, significance levels, and p-values. The second approach, called *Hypothesis Testing,* was developed by Jerzy Neyman and Egon Pearson, the son of Karl Pearson. The Neyman-Pearson approach gave us two hypotheses (the null and alternate) and Type I and Type II errors. We need not get into the details, but Fisher and Neyman strongly disagreed about their approaches to inferential statistics. Today some of the critics of NHST claim that many of the problems with how NHST is conducted stem from the inherent incompatibility of Fisher's significance testing and Neyman-Pearson's hypothesis testing.

Critiques of NHST have a long history. Stephen T. Ziliak and Deirdre N. McCloskey, who are articulate and thought-provoking critics of NHST, point to a long-forgotten letter written by William Gosset to Egon Pearson in 1926 concerning weaknesses in Fisher's approach to Significance Testing. Neyman and Pearson later operationalized Gosset's comments in their version of Hypothesis Testing.[18]

In 1951, Frank Yates, a close colleague of Ronald Fisher, placed an article in the *Journal of the American Statistical Association* that praised and criticized significance testing. Yates wrote that Fisher's *Statistical Methods for Research Workers* has caused researchers to "…pay undue attention to the results of the tests of significance they perform on their data…and too little to the estimates of the magnitude of the effects they are estimating."[19]

Comments' similar to Yates' were made during the next twenty years. In 1966, David Bakan declared that NHST flaws are apparent to everyone just like the state of undress of Hans Christian Anderson's foolish emperor parading around in his underwear to show his subjects his wonderful invisible robe. Bakan, an advocate of Bayesian inference, wrote,

> "*…the test of significance* does not provide the information concerning psychological phenomena characteristically attributed to it; and that, furthermore, a *great deal of mischief* has been associated with its use. *What is said in this paper is hardly original*. It is, in a certain sense, what '*everybody knows*.' To say it 'out loud' is, as it were, to assume the role of the child who pointed out that the *emperor was really outfitted in his underwear*." (Italics added)."[20]

In 1970, sociologists Ramon E. Morrison and Denton E. Henkel edited an anthology of 31 articles titled *Significance Test Controversy*. One contributor compared NHST to "a potent but sterile intellectual rake [a shamelessly immoral person or knave] who leaves in his merry path a long train of ravished maidens but no viable scientific offspring."[21]

A long-time critic of NHST, Jacob Cohen published an article in 1994 in *American Psychologist* entitled “The Earth is Round (p<.05).” Cohen wrote, “After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists.”[22] Cohen cites three problems with NHST:

1) The near-universal misinterpretation of p-values as the probability that the null hypothesis is false.
2) The misinterpretation that the complement of the p-value is the probability of successful replication of the study.
3) The mistaken assumption that if one rejects the null hypothesis, the theory that led to the test is affirmed.[23]

In the quarter century since Cohen’s often-cited article, legions of statisticians have published critiques of NHST. What’s wrong with NHST? It does not tell us what we want to know: The probability that the null hypothesis is true. It ignores the size of the effect. And, as John P. A. Ioannidis, along with other scholars, pointed out most published research using NHST is false because of low statistical power and the inability to replicate the studies’ findings.[24] In “Why Most Published Research Findings are False,” Ioannidis developed six corollaries, which take us beyond the narrow issue of NHST:

1) The smaller the sample sizes, the less likely the research findings are “true.”
2) The smaller the effect sizes, the less likely the research findings are “true.”
3) The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are “true.” Which is to say, the more variables the less likely the model is “true.”
4) The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are “true.”
5) The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are “true.”

6) The hotter a scientific field (with more scientific teams involved), the less likely the research findings are "true."

The calls for downplaying NHST testing or replacing it entirely have been increasing. In 2016, *The American Statistician* published an editorial on the use of p-values. In their editorial, Ronald L. Wasserstein and Nicole A, Lazar noted:

> Statisticians and others have been sounding the alarm about these matters for decades, to little avail. We hoped that a statement from the world's largest professional association of statisticians [the American Statistical Association] would open a fresh discussion and draw renewed and vigorous attention to changing the practice of science with regards to the use of statistical inference.[25]

This editorial was published under the title, "ASA Statement on P-Values and Statistical Significance."

Here is a summary what the author of the ASA editorial said about p-values, which they define as "…the probability under a specific statistical model that the statistical summary of the data…would be equal to or more extreme than its observed value:"[26]

1) P-values indicate how incompatible the data are with a specified statistical model.

2) P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. [Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.]

3) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

4) Proper inference requires full reporting and transparency.

5) A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis. The statement states that "…data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible." These approaches include: "…confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates."[27]

This statement, however, stopped short of calling for an end of NHST and its p-values.

In March 2019, the scientific journal *Nature* published an article by Valentin Amrhein, Sander Greenland, and Blake McShane calling for the end of NHST. This article had 800 signatories.[28] In addition, during the same week, *The American Statistician*, published another editorial on NHST, "Moving to a World Beyond '$p < 0.05$.'" It precedes 43 articles from prominent statisticians that deal with the contentious issue of how to move beyond NHST.

The authors of this editorial wrote:

> The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the [43] articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way. [29]

The editorialists acknowledge that voices in the 43 papers in this issue "do not sing as one. At times in this editorial and the papers you'll hear deep dissonance, the echoes of 'statistics wars' still simmering today."[30]

Clearly inferential statistics may be entering Stage Three of a paradigm shift as outlined by Kuhn: Adoption of a New Paradigm. The exact details of this new paradigm are still fuzzy. To quote Yogi Berra, the New York Yankees baseball player, "It's tough to make predictions, especially about the future."[31] Throwing caution to the wind, here is a broad outline of five things to expect. It would be a fool’s errand to predict the likely new statistical inference techniques with any more detail.

A) The terms statistically significant and statistically insignificant will be deemphasized, if not banished.

B) The use of the standard 0.05 significance level will decline along with the preeminence of p-values. The significance levels used, or more likely the confidence levels, will be selected based on the practical importance of the effect size for the issue under investigation.

C) Confidence intervals will replace NHST. Confidence intervals, which we described as the inverse of NHST, provide a range of plausible estimates of the population parameter rather than a single dichotomous conclusion of “significant or not significant.” As a consequence, confidence intervals provide more useful information than NHST.[32]

We looked at the issue of a person’s political affiliation—Republican, Independent, or Democrat—and their attitude toward the legalization of marijuana. We conducted this analysis three ways. We performed two different null hypothesis tests and we constructed confidence intervals.

In Module 15, we used a two-sample z-test for proportions. This test is unsatisfactory because we can only compare two samples at a time. We would have to conduct this test three times: 1) Republicans to Independents, 2) Republicans to Democrats, and 3) Independents to Democrats. The problem with this method is that the probability of a Type I error would greatly increase.

In Module 17, we again examined this question using a chi-square contingency table. We concluded that a person’s attitudes toward the legalization of marijuana are dependent on political affiliations. The problem is that this test is an omnibus test and we would have to conduct a *post hoc* analysis to determine which of the three pairs are unequal: 1) Republicans vs. Independents, 2) Republicans vs. Democrats, and 3) Independents vs. Democrats. One *post hoc* analysis is the LSD confidence interval that we used on Module 16 for ANOVA tests. So, why not skip this test and resort to using confidence intervals?

The clearest answer to the question of the association of political affiliation and attitudes toward the legalization of marijuana is found in Module 11 on confidence intervals, when we addressed this issue with confidence intervals. Because the confidence intervals for Republicans, Independents, and Democrats do not overlap, we concluded that Democrats are more likely to favor the legalization of marijuana than Independents, who are more likely to favor the legalization of marijuana than Republicans. This conclusion can be presented is a chart the provides a clear illustration of this conclusion. See Figure 1:

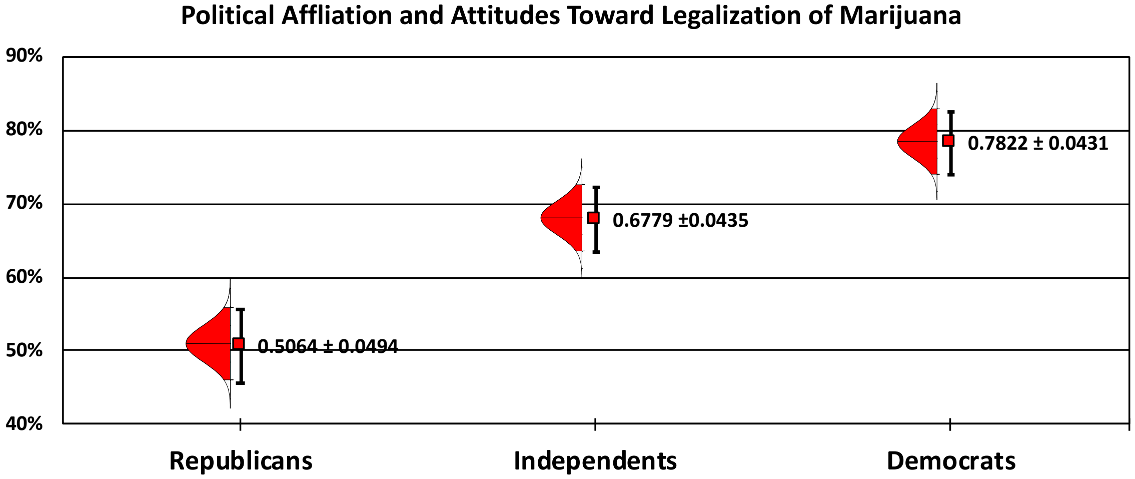

Figure 1: Confidence Interval Chart

D) Effect sizes will be reported using confidence intervals. Researchers will cease reporting the magnitude of effects in “tee shirt” sizes—small, medium, and large—based on effect size thresholds developed decades ago. They will use their judgment and the collective wisdom of experts in the field to determine the practical importance of the effect.

E) Meta-analysis will continue to grow in importance. Meta-analysis is a collection of statistical techniques for combining the results of multiple studies. Meta-analyses are useful for reconciling discrepancies regarding the effect size found in the research literature.

One final point: We are in an uncertain transition to a new paradigm. While statisticians argue vociferously for one approach or another, it would be easy to fall into a naïve cynicism. After all, if the leading statisticians are arguing that the methods developed a hundred years ago are seriously flawed and yet they cannot agree on the best way forward, you may jump to the conclusion that this statistics stuff is just a bunch of useless malarkey, hogwash, and hokum. Resist this temptation. All sciences undergo paradigm shifts. So, stay skeptical. Do not dismiss the discipline of statistics. Remember John Tukey's sage advice: "The most important maxim for data analysis to heed, and one which many statisticians seemed to have shunned is this: 'Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise.'"[33]

V. Exercises

1. **What is a major lesson you should have learned from *Clear-Sighted Statistics*?**

2. **Inferential Statistics is based on probability. How does this affect our understanding of the "truth"?**

3. **True or False: Statisticians are convinced that the way Statistics is taught and practiced is beyond reproach.**

4. **When Gov. Cuomo speaks about COVID, he uses a 3-day average, not a daily average. Why?**

5. **When should nonparametric tests be considered?**

6. **Multiple Choice: Compared to parametric tests, nonparametric tests have…**
   - ○ More Statistical Power
   - ○ The same Statistical Power
   - ○ Less Statistical Power

7. **Multiple Choice: p-values provide…**
   - ○ No useful information whatsoever
   - ○ Evidence in favor of the null hypothesis
   - ○ Evidence in favor of the alternate hypothesis
   - ○ Evidence against the null hypothesis

8. **Multiple Choice: Bayesian Statistics…**
   - ○ Is based on Bayes Theorem and subjective probability
   - ○ Updates evidence supporting both the H_0 and H_1
   - ○ Rejects the dichotomous approach of traditional NHST
   - ○ Was not eagerly adopted by most statisticians
   - ○ All of the above

9. **True or False: Since at least 1950, Statisticians have criticized NHST**

10. **In Jacob Cohen's article, "The Earth is Round ($p < 0.05$)," what are his criticisms of NHST?**

11. **What are John P. A. Ioannidis' six corollaries on NHST?**

CLEAR-SIGHTED STATISTICS

EDWARD VOLCHOK

*   *   *

[1] Richard P. Feynman, "Cargo Cult Science: Some Remarks on Science Pseudoscience, and Learning How Not to Fool Yourself," *California Institute of Technology 1974 Commencement Address*, http://calteches.library.caltech.edu/51/2/CargoCult.htm. Here is how Feynman described cargo cults in his commencement address: "In the South Seas there is a Cargo Cult of people. During the [Second World] war they saw airplanes land with lots of good materials, and they want the same thing to happen now. So they've arranged to make things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit in, with two wooden pieces on his head like headphones and bars of bamboo sticking out like antennas—he's the controller—and they wait for the airplanes to land. They're doing everything right. The form is perfect. It looks exactly the way it looked before. But it doesn't work. No airplanes land. So I call these things Cargo Cult Science, because they follow all the apparent precepts and forms of scientific investigation, but they're missing something essential, because the planes don't land."

[2] John P. Campbell, "Editorial: Some Remarks From the Outgoing Editor," *Journal of Applied Psychology*, Vol. 67., No. 6, 1982, p. 698).

[3] Dr. Fraley remarks are cited in, Stephen T. Ziliak and Deirdre N. McCloskey, *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, (Ann Arbor, MI: The University of Michigan Press, 2008). pp. 128-129.

[4] John Wilder Tukey, "John Tukey Quotes About Visualization," *AZ Quotes*. https://www.azquotes.com/author/14847-John_Tukey/tag/visualization.

[5] Cathy O'Neil, *On Being a Data Skeptic*. (Cambridge: O'Reilly Media, 2013). Kindle Edition. Location 8.

[6]"John Wilder Tukey," *National Science and technology Medals Foundation*, https://www.nationalmedals.org/laureates/john-wilder-tukey.

[7] Adam Hayes, "Six Sigma," *Investopedia*, June 25, 2019, https://www.investopedia.com/terms/s/six-sigma.asp.

[8] Gene V. Glass, Barry McGaw, and Mary Lee Smith, *Meta-Analysis is Social Research*, (Beverly Hills, CA: Sage, 1981), p. 104.

[9] David R. Brillinger, "…How Wonderful the Field of Statistics is…" Department of Statistics, University of California, Berkeley, https://www.stat.berkeley.edu/~brill/Papers/copssbrill.pdf.

[10] Ben Lambert, A Student's Guide to Bayesian Statistics, (Los Angeles: Sage, 2018). Alonso Ortega and Gorka Navarrete, "Bayesian Hypothesis Testing: An Alternative to Null Hypothesis Significance Testing in Psychology and Social Science." (InTechOpen, 2017), https://www.intechopen.com/books/bayesian-inference/bayesian-hypothesis-testing-an-alternative-to-null-hypothesis-significance-testing-nhst-in-psycholog. Sharon Bertsch McGrayne, The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy. (New Haven, CT: Yale University Press, 2011).

[11] Nicholas Maxwell, *Data Matters: Conceptual Statistics for a Random World*, (Emeryville, CA: Key College Publishing, 2008), p. 553.

[12] "Null Hypothesis Significance Testing Never Worked." *Statistical Thinking*, August 4, 2019. https://www.fharrell.com/post/nhst-never/.

[13] Alonso Ortega and Gorka Navarrete, "Bayesian Hypothesis Testing: An Alternative to Null Hypothesis Significance Testing in Psychology and Social Science." (InTechOpen, 2017), https://www.intechopen.com/books/bayesian-inference/bayesian-hypothesis-testing-an-alternative-to-null-hypothesis-significance-testing-nhst-in-psycholog.

[14] Michael D. Lee & Eric-Jan Wagenmakers, *Bayesian Cognitive Modeling: A Practical Course*. (Cambridge, UK: Cambridge University Press, 2014), p. 105.

[15] Thomas S. Kuhn, *The Structure of Scientific Revolutions*, (Chicago, IL: University of Chicago Press, 2012).

[16] Kuhn, p. 91.

[17] Kuhn, p. 150. Max Planck, *Scientific Autobiography and Other Papers*. F. Gaynor trans. (New York, 1949), pp. 33-34.

[18] Stephen T. Ziliak and Deirdre N. McCloskey, *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, (Ann Arbor, MI: The University of Michigan Press, 2008), pp. 133-134.

[19] Frank Yates, "The Influence of 'Statistical Methods for Research Workers' on the Development of the Science of Statistics," *Journal of the American Statistical Association*, Vol. 46, No. 253, March 1951, p. 32.

[20] David Bakan, "Tests of Significance in Psychological Research," *Psychological Bulletin*, Vol. 66, December 1966, p. 423.

[21] Ramon E. Morrison and Denton E. Henkel, *Significance Test Controversy*, (Chicago, IL, Butterworths, 1970). See Paul Meehl's "Theory Testing in Psychology ad Physics: A Methodological Paradox," p. 265.

[22] Jacob Cohen, "The Earth is Round (*p* <.05)," Vol. 49, No. 12. *American Psychologist*, December, 1994, p. 997. Cohen's first presented is concern about the lack of focus on statistical power and effect size in 1962. Jacob Cohen, "The Statistical Power of Abnormal-Social Psychological Research: A Review," *Journal of Abnormal and Social Psychology*, Vol. 12, No. 3, September 1957, pp. 145-153. In this article, Cohen raised the concern that psychologist focus on finding statistical significance, but ignore the role of statistical power. With the lack of statistical power, Type II errors—failing to detect a real effect—is a major shortcoming in the research literature. Lack of statistical power, in Cohen's words, "…must often result in investigations being undertaken which have little chance of success despite the actual falsity of the null hypothesis" (p. 145.).

[23] Jacob Cohen, "The Earth is Round (*p* <.05)," Vol. 49, No. 12. *American Psychologist*, December, 1994, p. 997.

[24] John P. A. Ioannidis, "Why Most Published Research Findings are False," *PLoS Med*, August 2005 Vol. 2(8) e.124. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/.

[25] Ronald L. Wasserstein and Nicole A. Lazar, "The ASA Statement on *p*-Values: Context Process, and Purpose", *The American Statistician*, Vol. 70, No. 20, March 7, 2016, p. 130. https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108.

[26] The ASA Statement on *p*-Values: Context Process, and Purpose", *The American Statistician*, Vol. 70, No. 20, March 7, 2016, p. 131. https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108.

[27] The ASA Statement on *p*-Values: Context Process, and Purpose", *The American Statistician*, Vol. 70, No. 20, March 7, 2016, pp. 131-2. https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108.

[28] Valentin Amrhein, Sander Greenland, and Blake McShane, "Scientists Rise Up Against Statistical Significance," *Nature*, Vol. 567, March 2019, pp. 305-307. https://www.nature.com/articles/d41586-019-00857-9.

[29] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar (2019) Moving to a World Beyond "p < 0.05," *The American Statistician*, 73:sup1, p. 2. https://www.tandfonline.com/doi/pdf/10.1080/00031305.2019.1583913?needAccess=true.

[30] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar (2019) Moving to a World Beyond "p < 0.05," *The American Statistician*, 73:sup1, p. 1. https://www.tandfonline.com/doi/pdf/10.1080/00031305.2019.1583913?needAccess=true.

[31] Yogi Berra, *Good Reads*, https://www.goodreads.com/quotes/261863-it-s-tough-to-make-predictions-especially-about-the-future.

[32] Geoff Cumming, *Understanding the New Statistics: Effect Size, Confidence Intervals, and Meta-Analysis*. (New York: Routledge, 2012), p. ix.

[33] John Wilder Tukey, “The Future of Data Analysis,” *Annals of Mathematical Statistics*, Vol. 31, No. 1, March 1962, pp. 13-14.

*   *   *