The Market Information Systems Research and Development (D) Working Group Review of Artificial Intelligence Techniques in Market Analysis

Executive Summary

This report fulfills the Market Information Systems Research and Development (D) Working Group charge to evaluate the potential benefits of artificial intelligence (AI) in relation to market analysis. After careful consideration, the Working Group concluded that there may be possible benefits to improve analysis techniques. Several caveats are discussed as well. AI may not be suitable for data currently available to state insurance regulators. In addition, some of the techniques perform complex data mining operations, which can produce results that lack a clear interpretation. Lastly, AI techniques are designed for, and many require, very large datasets. As such, AI should be contemplated in the context of a long-range plan, beginning with repairing known issues with existing data, and employing more rigorous traditional statistical techniques to assess predictive accuracy of analytical tools. Subsequently, state insurance regulators can consider the acquisition of data appropriate to AI.

Introduction

In early 2021, the Market Information Systems Research and Development (D) Working Group received a charge from the Market Information Systems (D) Task Force to explore possible applications of artificial intelligence (AI) methods in market analysis. An early difficulty encountered by the Working Group is that the term "AI" itself has a variety of contested meanings. In addition, private sector entities have adopted the term as a marketing concept and inappropriately apply the label to products simply as a selling point. As such, the term has come to acquire a variety of meanings and is an "essentially contested concept."[1]

At its most general level, the term "AI" implies machine capacities that mimic or are analogous to processes of human reasoning and learning and entail some degree of machine autonomy in which learning occurs without significant human intervention. Beyond this general description, the Working Group did not feel that an attempt to define the term more strictly would be fruitful. Rather, the term is employed simply as a shorthand reference for a collection of various techniques that algorithmically seek patterns in data that are predictive of some future outcome. Common methods include machine learning, neural networks, and decision tree analysis. These processes are often contrasted to the traditional hypothetical-deductive methods of model specification associated with classical statistics. However, there does not appear to be a bright line of demarcation so that a particular technique can be firmly fixed within either category.

In addition, the Working Group focuses on what is commonly called "narrow AI," in which machine algorithms are employed for narrowly defined and limited tasks. More advanced systems, called "general AI," possess generalized autonomous problem-solving capacities that are comparable to the

[1] The term "essentially contested concept" was coined by W.B. Gallie in the seminal presentation to the Aristotelian Society in 1956.

processes of the human brain, and they are able to adapt to novel situations or information (Macnish et al., 2019).

It is important to emphasize the ways in which AI modeling techniques contrast to the standard scientific model employed in classical or traditional statistics:

Classical Statistics: Method of hypothetical-deductive reasoning in which hypotheses are clearly and narrowly specified *prior* to data testing, often with a prior understanding of the underlying causal nature of the relationships between variables. **Purpose:** To further causal understanding.

AI: Often employs a type of "data mining" in which a machine pattern-seeking algorithm is released "into the wild" to identify possible correlations between variables that may be predictive of some independent variable. Hypotheses are not specified prior to data analysis, and the algorithm may very well identify correlations that would not have occurred to an analyst and whose causal relationship is constructed post-hoc (to the degree that AI users are concerned with causality at all). **Purpose:** Predict future outcomes or events.

The difference between these two approaches is not trivial, and significant disagreements about the advantages and disadvantages of AI remain. It is of note that AI did not emerge principally from university statistics departments, but rather from the field of computer science. Many statisticians remain skeptical of the techniques and have offered up a variety of caveats for their use. For example, recently the American Statistical Society (ASA) reacted to the "reproducibility crisis" afflicting some disciplines that have discovered, with much consternation, that a large volume of published works could not be replicated. The concern was that increasingly less rigorous statistical methods departing from the hypothetical-deductive approach were becoming more prominent in a variety of fields, undermining confidence on research findings. Remarking on departures from a rigorous hypothetical-deductive approach with "data mining" and like methods in which pattern seeking is largely ceded from a researcher to a machine, the ASA warned about improper inferences that might result from such techniques. The ASA centered its discussion on the p-value, related to the probability that some observed relationship occurred by chance along. A low p-value is often employed to minimize the probability that chance relationships will be misinterpreted as a relationship that is a meaningful, non-random outcome:

"Conducting multiple analyses of the data and reporting only those [analyses] with certain p-values…renders the reported p-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significant chasing, significance questions, selective inference and a 'p-hacking' leads to a spurious excess of statistically significant results…and should be vigorously avoided" (Wasserstein & Lazar, 2016).

To translate the ASA's statement into more easily understood and less technical terms, the ASA is warning against *false positives* in which an analysis produces random or chance correlations between items that are not meaningfully related—that is, where a chance relationship is mistaken for a true causal relationship. That AI largely jettisons causal understanding as its primary goal (to the degree that causality is a concern at all) increases the probability that statistical results may be uninterpretable in any meaningful sense. This is clearly evinced by the increasing debate among state insurance regulators and insurers regarding the meaning of statistical relationships appearing in predictive

models that lack intuitive or, in many cases, even plausible explanations. See Appendix A for further discussion of the ASA statement.

The discussion above is not intended to sway state insurance regulators one way or the other with respect to AI. The purpose is simply to proffer some caveats shared by many statisticians. A final caveat is the AI techniques were developed to analyze very large data sets consisting of millions of records and possibly thousands or tens of thousands of variables. It is said to have an advantage in that algorithms can perform a large volume of analyses across different constellations of variables in a way that would be highly impractical employing traditional (and manual) model building. For small data sets, such as the limited data currently available to market analysts, it is unclear whether the expense associated with developing AI techniques can be justified, nor whether AI is at all superior to traditional model building methods. This is not an unimportant point and is discussed in more depth elsewhere in this recommendation.

Current Status of Market Analysis

Quantitative market analysis relies on just a handful of data sources:

The Complaint Database System (CDS): The NAIC compiles complaints against insurers received by state insurance regulators. Thus, each state has access to a national-level database. Complaint indices are "normalized" by expressing the volume of complaints to premium, compared with the overall industry total.

The Regulatory Information Retrieval System (RIRS): Regulatory actions in relation to insurance entities are captured in the RIRS database. Actions range from intervention in financially troubled entities to violations of producers and insurance carriers. Each record identifies the cause of the action, as well as any orders, fines, or restitution amounts. The RIRS database is currently being substantially revised to capture significantly more detail.

The Market Actions Tracking System (MATS): The MATS database captures information pertaining to market conduct exams, as well as actions short of exams. Data captured include area of scrutiny (claims, underwriting, etc.) and the outcome of the market action (order, fine, etc.). By matching MATS actions with RIRS, additional detail about the nature of the violation can be assessed.

The Market Conduct Annual Statement (MCAS): The MCAS was developed to capture data with the primary purpose of assessing an insurer's market performance and identify potential market irregularities. The data focus primarily on claims handling and underwriting, and data are scrutinized with respect to claims processing times and denials, nonrenewal and cancellation practices, and overall turnover in a book of business. Data are captured by line and coverage. To date, MCAS data are collected for life and annuities, private automobile, homeowners, health (both on and off the federally facilitated marketplace [FFM]), long-term care (LTC), lender-placed insurance, disability income, and private flood.

Miscellaneous Data Sources: Some financial data has been incorporated into market information systems. Insurers that are under financial stress, or that rapidly expand into or contract out of a line

of business, or that exhibit high defense or other adjudication costs, may be subjected to additional analysis. While financial indicators are only indirect or proxy measures of potential market issues, and by themselves may have no clear market-based interpretation, interpretation within the context of a host of other indicators may be reflective of the present of a market-relevant issue.

The NAIC, in conjunction with state insurance regulators, has developed a broad scope "market score" that incorporates much of the data referenced above, which is made available to regulators via the Market Analysis Prioritization Tool (MAPT). One such data are "normalized" by the premium volume and scope of company operations as necessary. For example, several RIRS-based ratios express the volume of RIRS actions in relation to premium volume, the number of states in which they have significant premium, and a composite ratio that incorporates both premium and scope. Each ratio is given a score, and their contribution to the overall score weighted according to their perceived predictive relevance. For example, financial ratios are accorded significantly less weight than complaints, as their relationship to market misconduct is considered more speculative and indirect.

An important caveat is that predictive analytics is not well developed in market regulation. The ratios employed in the Market Analysis Review System (MARS) have not been subjected to rigorous statistical tests that demonstrate their analytic utility. While some work has been performed in this regard, such work is significantly hampered by a dearth of appropriate data. For example, future RIRS actions are often employed as the dependent variable (the outcome of interest to be predicted). However, this presents all manner of statistical challenges. While it is certainly reasonable to use prior outcomes (past RIRS actions) to predict future outcomes (the RIRS actions to be predicted), employing RIRS actions as both dependent and independent variable introduces significant complexities in the interpretation of any observed relationship between the two. One can imagine, for example, that the use of RIRS actions in market analysis invites greater scrutiny to a given insurer, and that in turn generates future regulatory actions precisely because the company received additional scrutiny. Companies that have no "prior offenses" fail to attract regulatory scrutiny, so that any infractions may escape regulatory action *for precisely that reason*. This problem is certainly not insurmountable, but it must be explicitly recognized in any model building exercise, whether with AI or with more conventional statistical techniques.

In general, the paucity of rich data sources has significantly hampered the adoption of more rigorous analytical techniques. To return to RIRS, these data are not rich sources of detailed information. Schematics are not well designed "from the ground up." Essential data are missing, such as line of business.

Any consideration of AI or any other analytical techniques must necessarily view the utility of such techniques within the context of available data. Regardless of the validity of a technique in general, it will have limited utility if data are themselves limited. Any recommendation to employ such methods must therefore at the same time recommend a thorough review of available data.

Importantly, results of quantitative analysis are always treated as merely suggestive and tentative and are regarded as at most a precursor to more qualitative analysis. It currently is employed to prioritize entities that may merit additional scrutiny and to narrow focus on a much more limited subset of companies out of a larger pool of companies. It therefore primarily prioritizes limited regulatory resources.

State insurance regulators avail themselves of the formal analytical processes adopted by the NAIC. Quantitative or "baseline" analysis identifies entities with anomalous indicators that significantly depart for industry-wide values. A "level 1" analysis may be pursued, in which an analyst devotes additional scrutiny to such things as complaint trends, common reasons complaints are lodged against an insurer, similarities in RIRS actions, etc. If concern still remains (or additional concerns are identified) subsequent to level 1 analysis, a structured level 2 analysis may be performed. A level 2 analysis requires a much greater commitment of time and resources. For example, rather than just manually reviewing complaint data to identify patterns, an analyst may manually review actual complaint documentation to garner a more detailed understanding of the nature of complaints.

As a preliminary to the following discussion, AI/statistical analysis may have two primary functions within the context of the current market analysis structure:

1. More accurately identify companies that merit the additional expenditure of resources necessary to perform the more labor-intensive level 1 and level 2 analyses. Analysis processes that more efficiently identify problem companies for this purpose are by definition more effective and more effectively target resources by avoiding "false positives" (for lack of a better word).
2. Potentially, AI methods could assume many of the functions that are currently performed manually. For example, many of the pattern-seeking analysis performed by analysts in a level 1 review could conceivably be more efficient if automated. Potentially, AI could identify patterns that might elude a human analysis. A very advanced level of AI could perhaps assume complex analysis involved with manually reviewing complaint files and documents. However, while the possibility is raised here, it is not further pursued. That level of AI suitable for tasks may not even exist as yet, or if it does, it may be so specialized that it may not be available to state insurance regulators. Even if available, the likely enormous costs themselves would render them highly impractical.

Whether such AI exists, is available at a practical cost, and can actually out-perform more conventional analyses are questions that the Market Information Systems Research and Development (D) Working Group is simply unable to satisfactorily address. The Working Group merely suggests initially limiting the scope of ambitions to a few methods that are commonly, if not universally, recognized as AI, such as machine learning or neural networks. More expansive or ambitious efforts may result in a fruitless search for "unobtanium."[2]

Given very large data sets, well beyond what is currently available to market analysts, AI may have clear advantages to more conventional approaches. The slow, methodical, hypothetical-deductive approach that forms the core of conventional statistics may have advantages in terms of generating valid causal conclusions. However, AI may have certain advantages with respect to confronting the enormity of modern data. As AI is well-suited to performing much more expansive analysis and pattern-seeking routines over vast quantities of data, it may well identify predictive patterns that would have escaped conventional analysis or that are counterintuitive such that some hypotheses may never

[2] A tongue-in-cheek term originating among engineers in the 1950s. It is defined by Wikipedia as "… any hypothetical, fictional, or impossible material, but it can also mean a tangible but extremely rare, costly, or reasonably unobtainable material. Less commonly, it can refer to a device with desirable engineering properties for an application, but which are exceedingly difficult or impossible to achieve."

have occurred to an analyst employing a standard hypothetical-deductive approach. However, there are distinct disadvantages as well, and they are shared by other approaches often termed "data mining." The fact is that patterns may lack an intuitive meaning, and the manner in which such patterns are identified and render interpretation may be unclear. Additionally, patterns may generate numerous "false positives," apparent patterns or correlations that are purely random and possess no meaning or any real predictive power whatsoever. This is not fatal for AI techniques, but it introduces much in the way of caveats and requires significant remedial measures to be employed. This problem is so significant that it merits a much fuller discussion in a separate section below.

The Work of Market Information Systems Research and Development (D) Working Group

The Working Group solicited input from various parties. Two parties delivered presentations to the Working Group:

1. On June 16, 2021, the Working Group discussed a presentation regarding AI methods currently being explored by NAIC staff to predict which insurers are likely to experience financial stress, including insolvency. Beginning in January 2021, an outside consulting group was retained to develop both AI as well as more traditional statistical techniques to construct predictive models of insolvency risk. The efforts are ongoing at the time of writing. Presenters believed the methods were promising and could significantly advance financial risk surveillance. Among AI and statistical models explored were decision tree analysis, generalized linear models (GLMs), and logistic regression.

2. During the Working Group's June 21, 2021, meeting, Birny Birnbaum (Center for Economic Justice—CEJ) encouraged the Working Group to adopt a long-term perspective and develop a multiyear plan to explore AI techniques that might be beneficial to market analysis. He also indicated that state insurance regulators have to date failed to acquire granular transactional data that could be exploited by AI methods to afford a much more robust surveillance system to reduce consumer harm to the extent possible.

   After the meeting, the Working Group convened a subject-matter expert (SME) group with the intent of creating a draft recommendation to be submitted to the Working Group.

Recommendations

The Working Group recommends developing a long-range plan, in a sequence of five steps.

I. Existing Market Analysis Data

As noted above, market analysis suffers from a paucity of detailed data. Some movement in expanding data and remedying deficiencies was made with a complete redesign of the RIRS data, which will facilitate analysis of factors related to an entity sanctioned by state insurance regulators. If implemented, RIRS will also capture much more detailed data related to the specific misconduct that garnered a regulatory response. The RIRS proposal is currently under discussion with the Market Information Systems (D) Task Force, to which Working Group reports.

The remainder of available data also suffers from significant deficiencies. Insurers employ a variety of definitions to produce MCAS data. Even such a fundamental concept as a "claim" is reported

differently by different insurers, making market-wide analysis challenging. For example, the MCAS defines a claim in the conventional sense of "a demand for payment." Investigation by the Missouri Department of Commerce & Insurance (DCI) has determined that the definition is interpreted in wildly divergent ways across the industry that simply makes meaningful comparison impossible and renders key market indicators or ratios largely meaningless. Some insurers set up a claim on a coverage that is reasonably related to the facts of the incident as relayed by a claimant. Other insurers set up all possible coverages on a policy as a claim in their internal systems regardless of whether those coverages might be reasonable implicated in a claim. As might be imagined, those carriers have significantly higher ratios of claims closed without payment. This and other issues remain with the MCAS and significantly impair market analysis.

Recommendation 1: Survey currently available market analysis data, and identify substantive deficiencies based on the nature and substance of the data elements collected. Ensure that all data are consistently reported across insurers to the degree practical and ensure adherence to definitions of data elements.

II. Existing Methods of Market Analysis

Current quantitate methods of market analysis are large based on *ad hoc* and *intuitive* understanding of how data indicators might be related to market misconduct. For example, one of the earliest indicators developed are complaints received by state insurance regulators regarding insurers. It is probably not unreasonable to interrogate complaint data to identify trends over time, as well as just overall complaint volume, to attempt to identify potential problems in a market. Similar indices consider the volume of RIRS actions, as well as the gravity of infractions in terms of potential consumer harm. It is the opinion of many state insurance regulators that such indicators possess a rational relationship to market misconduct and are relevant to identify market actors that might benefit from a heightened level of regulatory scrutiny.

While the Working Group agrees with the rationale behind such market indicators, analytical tools have not to date been subjected to more rigorous statistical methods to clearly identify the predictive power and assess their relative importance or weight. For example, the MAPT, maintained by the NAIC and available to state insurance regulators, employs overall insurer scores based on various indicators. However, the weight of these indicators employed in the score were assigned by state insurance regulators based on experience, as well as assessment of whether a likely relationship have a clear rational meaning. For example, complaint ratios are weighted significantly more heavily than things like financial indicators. The Working Group believes subjecting the scoring system to rigorous statistical analysis could yield significant benefits in identifying problem market actors.

Recommendation 2: In conjunction with recommendation 1 (assess data quality), state insurance regulators should adopt a much more rigorously statistical approach to identify the predictive power of market scoring systems, assess how each variable should be weighted in terms of its unique contribution to productiveness, and drop those that lack analytic utility. In addition, effort should be made to integrate data into a single overall analysis. For example, the MAPT does not incorporate MCAS data, which is typically subject to a separate analysis. The Working Group believes that a "piecemeal" approach is likely less effective than a more integrated approach.

It is noted that the current state of data will likely prove limiting and that such efforts may not make much progress until additional data are made available (such as the proposed revisions to the RIRS data, currently subject to NAIC discussion).

III. Available Approaches: Exploring AI

In additional to more traditional statistical tools, such as various types of regression models and correlation analyses, AI may offer additional benefits. Some commercial statistical packages have incorporated AI methods. The statistics package SAS, which is widely used in both the private and public sectors, makes some AI techniques available in its standard statistical module.[3] In addition, SAS has developed a module called Enterprise Miner, which incorporates both data mining and some lower-level AI routines. (For those familiar with the terms, it performs such things as decision-tree analysis, neural networks, and like forms of analyses). Other modules make machine learning available—a potentially powerful type of analysis that modifies prior predictive algorithms as new data become available.

Recommendation 3: In undertaking recommendation 2, incorporate various promising AI modes of analyses, as well as traditional statistical modeling. Constantly assess the precision of model outcomes relative to objectives such as identifying potential market issues.

IV. Qualitative Analysis

The current model of market analysis incorporates a multistage hierarchical structure. First, quantitative analysis such as that produced by the MAPT identifies potential market problems and narrows focus to entities that appear to exhibit potential areas of regulatory concern. Having narrowed down the focus of analysis to a much more limited pool of candidates, market analysts in the states engage in more manual or qualitative analysis of additional information sources. For example, an analyst may review a selection of complaint files to identify additional patterns of market behavior to better understand their nature and substance.

As noted above, AI techniques such as text analysis could potentially expand such exercises and improve the identification of concerning patterns at a deeper level, as well as assess ways to improve the efficiency of other qualitative tasks.

Recommendation 4: Assess ways AI can improve both the efficiency of *qualitative* analysis and facilitate pattern recognition across larger volumes of textual evidence, including most especially complaints, but perhaps other textual sources. For example, the "level 1" analysis formalized in NAIC market system may include a review of the "management discussion and analysis" of the financial annual statement.

[3] SAS is markets in "modules," each consisting of a different suite of capabilities that can be tailored to a user's need. For example, "base SAS" provides standard data handing programs. A "statistics module" provides a wide-ranging set of analytical routines.

V. Longer-Range Planning

As noted above, data mining and AI techniques were developed primarily as tools to analyze large volumes of data. For data past a certain magnitude, including especially those containing many hundreds or even thousands of variables, the traditional hypothetical-deductive cornerstone that is the cornerstone of traditional statistical inference may be ill-suited as well as cost-prohibitive in terms of time and resources. If the purpose is solely prediction as opposed to causal understanding, AI can fine-tune predictive algorithms by testing relationships that may be unlikely to occur to a statistician employing causal modeling.

Currently, such large volumes of data are unavailable to market analysts, though they could potentially be obtained. More granular data pertaining to claims, underwriting, and other areas of company operations are routinely collected via the "standard data requests" adopted as a supplement to the *Market Regulation Handbook* and commonly employed in market conduct exams.

However, AI and data mining can churn up counterintuitive statistical relationships that defy ready interpretation. In addition, it is likely to detect proxy relationships that are not understood. Proxy relationships, in which a third variable is substituted for an underlying variable of interest, are often employed in statistical models. This is often due to the accessibility or cost of obtaining data of the actual causal variable of interest. However, when employed in traditional statistical analysis, the nature of the relationship between the proxy variable and the actual variable of interest is generally well understood. This is not true of AI techniques that employ or resemble data mining.

The techniques are also likely to generate some number of purely chance relationship, where a correlation is generated by random chance. Inferential statistics seek to minimize mistaking a chance relationship for a meaningful association. Typically, the use of a p-value requirement of 0.05 or less limits the probability of accepting a random relationship to no more than 5% of occurrences. However, a 5% threshold means that over time, false, or chance relationships will be misinterpreted of a true correlation.

This fact is not fatal for the use of AI in market analysis, but it does represent a strong caveat for those employing the techniques, at least those that share elements with data mining. Careful interpretations of p-values should recognize an increased possibility of false positives. Observed relationships should be assessed and validated over time to ensure correlations are stable. In addition, once relationships are identified via AI and found useful, standard statistical models should also be employed to test whether different techniques yield superior predictive power. Additional discussion of caveats is presented in the appendix.

That said, there is much potential of AI in market analysis, *assuming that additional, more granular, data are available.* As noted, such techniques are most suited for large datasets whose very size would make a standard statistical approach impractical just given the sheer number of possible correlations available for testing.

Recommendation 5: Systematically explore potential data sources suitable for AI techniques, with an eye for discovering patterns and relationships in relation to some well-defined outcome one is attempting to predict. This may be identifying entities that may merit additional regulatory scrutiny in

a way that is currently done by the less sophisticated methods employed in the MAPT or with the MCAS. Larger volumes of data, such as the standard data requests, can be subjected to AI to identify problematic claims handling, underwriting, and other insurer practices.

Summary of Recommendations

Recommendation 1: Survey currently available market analysis data, and identify substantive deficiencies based on the nature and substance of the data elements collected. Ensure that all data are consistently reported across insurers to the degree practical, and ensure adherence to definitions of data elements.

Recommendation 2: In conjunction with recommendation 1 (assess data quality), state insurance regulators should adopt a much more rigorously statistical approach to identify the predictive power of market scoring systems, assess how each variable should be weighted in terms of its unique contribution to productiveness, and drop those that lack analytic utility. In addition, effort should be made to integrate data into a single overall analysis. For example, the MAPT does not incorporate MCAS data, which is typically subject to a separate analysis. The Working Group believes that a "piecemeal" approach is likely less effective than a more integrated approach.

Recommendation 3: In undertaking recommendation 2, incorporate various promising AI modes of analyses, as well as traditional statistical modeling. Constantly assess the precision of model outcomes relative to objectives, such as identifying potential market issues.

Recommendation 4: Assess ways AI can improve both the efficiency of *qualitative* analysis and facilitate pattern recognition across larger volumes of textual evidence, including most especially complaints, but perhaps other textual sources. For example, the "level 1" analysis formalized in NAIC market system may include a review of the "management discussion and analysis" of the financial annual statement.

Recommendation 5: Systematically explore potential data sources suitable for AI techniques, with an eye for discovering patterns and relationships in relation to some well-defined outcome one is attempting to predict. This may be identifying entities that may merit additional regulatory scrutiny in a way that is currently done by the less sophisticated methods employed in the MAPT or with the MCAS. Larger volumes of data, such as the standard data requests, can be subjected to AI to identify problematic claims handling, underwriting, and other insurer practices.

Appendix: Caveats

Recently, some fields of scientific inquiry have experienced much consternation and hand-wringing due to the so-called "replicability crisis" resulting from the realization that many studies published in top-tier journals could not be replicated. In 2015, Open Science Collaboration published research into the replicability of psychological studies. Of the 100 studies that were subjected to testing, replications yielded statistically significant results in only 36% compared to 97%of the original publications (Open Science Collaboration, 2015). Similar reproducibility issues were found in other fields.

Attention was directed at quantitative methods, particularly those made possible by modern computing power. Researchers can run countless variations of models, including multiple different variables, cross-effects, and other tweaks, until they eventually produce positive or statistically significant results. The inevitable outcome of the lack of rigor of such methods is that many chance correlations will be mistaken for meaningful relationships.

Think of it this way. The probability of obtaining all heads from 10 flips of a fair coin is 1/1024. So, if a researcher actually performed the experiment 1,024 times and obtained 10 heads at least once, it would obviously be improper to infer that the coin was a two-headed coin. Without knowledge of the total number of trials, one might reject the "null hypothesis" that the coin is fair, and results would be "statistically significant" with a p-value of (1/1,024) = 0.00098, well below the 0.05 maximum threshold to establish statistical significance. But the true p-value can only be calculated with knowledge of the total number of trials prior to obtaining the recorded result, such that the true p-value is well above the maximum threshold.

There are no allegations of willful misconduct so much as careless and sloppy methods, producing much introspection about how statistics methods are taught to scientists at colleges and universities. The problem is so significant that the following year, the American Statistical Association (ASA) released a statement regarding misuse of p-values and practices known as "p hacking" or "data dredging." A letter from the ASA is reprinted below, with a link to the full statement (used with permission).

Really, this is a warning for state insurance regulators not to adopt a casual attitude about apparent relationships turned up by the methods. When such methods are employed, modelers should be on constant guard against mechanical interpretations of model outputs. It is important to fully understand what is going on in the "black box" of an AI algorithm, the results of all statistical tests performed, and the totality of processes generating final results.

A high number of false positives that prompt regulatory follow-up can risk draining away regulatory resources going down blind allies.

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of QuantitativeScience

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice "emphasizes principles of good study design and conduct, a varietyof numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean."

"The *p*-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, the ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post p<0.05 era.'"

"Over time it appears the *p*-value has become a gatekeeper for whether work is publishable, at least in some fields," said Jessica Utts, ASA president. "This apparent editorial bias leads to the 'file-drawer effect,' in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important scientifically is never seen in print. It also leads to practices called by such names as '*p*-hacking' and 'data dredging' that emphasize the search for small *p*-values over other statistical and scientific reasoning."

The statement's six principles, many of which address misconceptions and misuse of the *p*-value, are the following:

1. *P-values can 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 theprobability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only onwhether 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 theimportance of a result.*

6. *By itself, a p-value does not provide a good measure of evidence regarding a model orhypothesis.*

The statement has short paragraphs elaborating on each principle.

In light of misuses of and misconceptions concerning *p*-values, the statement notes that statisticians often supplement or even replace *p*-values with other approaches. These include methods "that emphasize estimation over testing such as 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."

"The contents of the ASA statement and the reasoning behind it are not new—statisticians and other scientists have been writing on the topic for decades," Utts said. "But this is the first time that the community of statisticians, as represented by the ASA Board of Directors, has issued a statement to address these issues."

"The issues involved in statistical inference are difficult because inference itself is challenging," Wasserstein said. He noted that more than a dozen discussion papers are being published in the ASA journal *The American Statistician* with the statement to provide more perspective on this broad and complex topic. "What we hope will follow is a broad discussion across the scientific community that leads to a more nuanced approach to interpreting, communicating, and using the results of statistical methods in research."

About the American Statistical Association

The ASA is the world's largest community of statisticians and the oldest continuously operating professional science society in the United States. Its members serve in industry, government and academia in more than 90 countries, advancing research and promoting sound statistical practice to inform public policy and improve human welfare. For additional information, please visit the ASA website at www.amstat.org.

For more information:

Ron
Wasserstein

Citations

Macnish, K., Ryan, M. & Stahl, B. (2019). Understanding ethics and human rights in smart information systems: A multi-case study approach. *The Orbit Journal*, *2*(2), 1–34.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://science.sciencemag.org/content/349/6251/aac4716

Wasserstein, R.L., & Lazar, N.A. (2016). The ASA statement on p-values: Context, process and purpose. The American Statistician, *70*(2), 129–133.