CHAPTER

20

Significance versus Luck in the Age of Mining: The Issues of *P*-Value "Significance" and "Ways to Test Significance of Our Predictive Analytic Models"

PREAMBLE

One of the greatest challenges we face in predictive analytics studies is the common perception that the phrase "it has been proved" means something. Many times, it does not. The reason is that the traditional statistical analysis approaches have based findings in their studies upon the meaning and application of the *P*-value. The *P*-values reflect the estimated probability that you are wrong in your prediction when you think you are right. The smaller the *P*-value, the greater is the probability that you are right in your conclusion. The problem is that the theory underlying the *P*-value does not fit reality in the real world in most cases, but it is used any way. This chapter will explore the nature of "statistical proof" and "accuracy" of predictive analytic models.

INTRODUCTION

Gartner (2017) thinks of predictive analytics as "an approach to data mining" that has four attributes:

1. An emphasis on prediction (rather than description, classification, or clustering)
2. Rapid analysis measured in hours or days (rather than the stereotypical months of traditional data mining)
3. An emphasis on the business relevance of the resulting insights (no ivory tower analyses)

Handbook of Statistical Analysis and Data Mining Applications
https://doi.org/10.1016/B978-0-12-416632-5.00020-7

4. (increasingly) An emphasis on ease of use, thus making the tools accessible to business users
   These attributes emphasize by Gartners appear to be somewhat biased toward business users, which is understandable, when predictive analytics is being used today in all kinds of endeavors. Whatever the domain, another attribute is important.
   Thus, another, fifth element, will be added in our discussion:
5. Significance (the "Real Accuracy") of the model
   This last "attribute," "significance," is what we will concentrate on in this chapter.

THE PROBLEM OF SIGNIFICANCE IN TRADITIONAL *P*-VALUE STATISTICAL ANALYSIS

We cited in our first edition (2009, Handbook of Statistical Analysis and Data Mining Applications) that medical research/scientific articles have as much as "70% errors" in experimental design or algorithm choice, based on a study of 72 cancer trials analysis articles indexed in Medline and PubMed services (Murray et al., 2008).

The controversy over the proper use of *P*-values in studies utilizing traditional statistics has increased dramatically since 2009. Nowhere is this more apparent, important, and even critical than in medicine. On the one hand, the traditional statistical tool of *P*-value was adopted to assist with reproducibility and allow comparison of studies by different clinical investigators. On the other hand, the pursuit of "*P*-value perfection" can have tragic consequences. The title of a June 2015 paper was "Science is Heroic, with a tragic (statistical) flaw" (Siegfried, 2015a). The gist of this article was "…the standard statistical methods for evaluating evidence are usually misused, almost always misinterpreted and are not very informative even when they are used and interpreted correctly …" The problems persist because the quest for "statistical significance" is mindless. Determining significance has become a surrogate for good research. The conclusions, among others, were that "scientific studies are not as reliable as they pretend to be…no more reliable than public opinion polls," and leading one to extrapolate that such flaws could be fatal when relied upon in medical research and treatment.

Gigerenzer and Marewski (2015) write in the Journal of Management: "…Among multiple scientific communities, "statistical significance" has become an idol, worshiped as the path to truth. Advocated as the only game in town, it is practiced in a compulsive, mechanical way — without judging whether it makes sense or not."

This publication was followed in July of 2015 by a second part to the Siegfried paper, titled "Top 10 ways to save science from its statistical self" (Siegfried, 2015b), in which Siegfried stated unequivocally that

> Statistics is to science as steroids are to baseball. Addictive poison. But at least baseball has attempted to remedy the problem. Science remains mostly in denial. True, not all uses of statistics in science are evil, just as steroids are sometimes appropriate medicines. But one particular use of statistics — testing null hypotheses — deserves the same fate with science as Pete Rose got with baseball. Banishment.

Testing of null hypothesis is the hallmark of scientific methodology of the past century. But it has also been the prime reason making many research results irreproducible, if not erroneous. This has been particularly rampant in the medical sciences apparently because most

of the users in the medical sciences do not know how to apply traditional *P*-value statistics properly to their research designs.

A third article rapidly followed on August 27, 2015 titled: "Psychology results evaporate upon further review" (Bower, 2015), which reported that only 35 of 97 "statistically significant results" could be replicated by a group of 270 researchers led by psychologist Brian Nosek of the University of Virginia in Charlottesville (Nosek, 2015). This group used a complicated "replication/reproducibility" analysis which among other things attempted to get at the question of whether the "expertise/lack of expertise" in the replicating scientific team or the strength of the initial evidence (e.g., significance level of the *P*-value) was more important in determining reproducibility; the initial study's *P*-value strength appeared to win out. Some of the results are illustrated in Fig. 20.1, which shows clearly that a majority of the studies could not be replicated.

A fourth article at the end of 2015 was titled: "Year in Review: Scientists tackle the irreproducibility problem" (Saey, 2015a). The summary for the year stated that lack of the ability to replicate published scientific results has been an issue for years, particularly in the medical

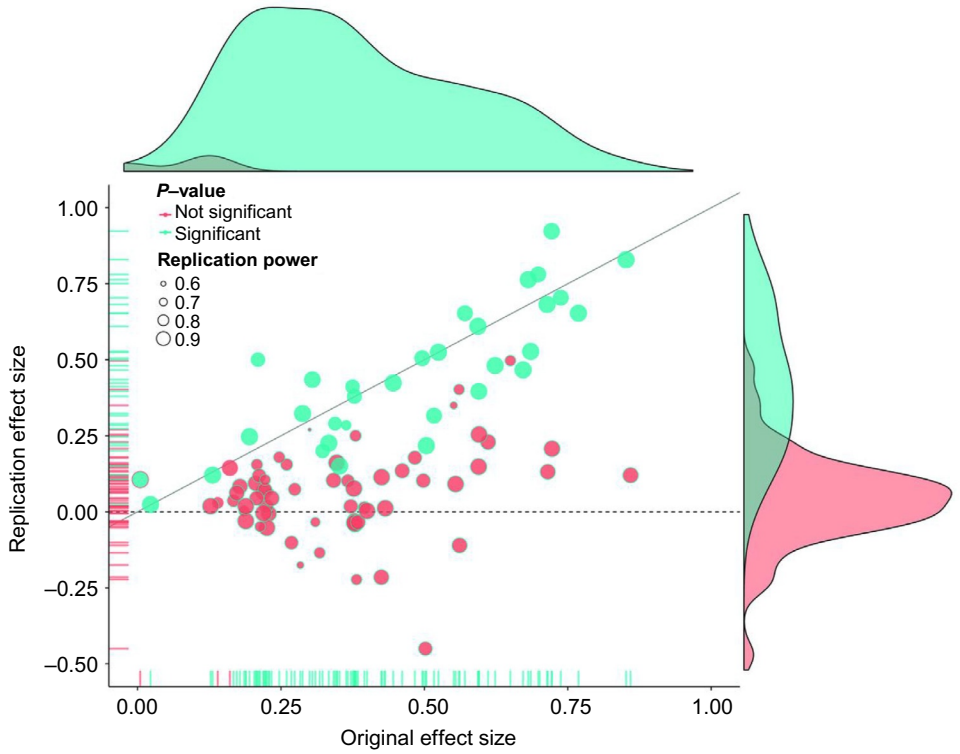

FIG. 20.1 Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant *(green)* and nonsignificant *(red)* effects (Nosek, 2015).

FIG. 20.2 The extension of reproducibility. *From https://www.sciencenews.org/article/year-review-scientists-tackle-irreproducibility-problem—Image source: THEEVENING/ISTOCKPHOTO.*

and social sciences (Fig. 20.2). Several research groups that were studying this problem found that the replication results were indeed not good (Saey, 2015b).

Finally, a fifth article published at the beginning of 2016 titled: "Experts issue warning on problems with p-values" (Siegfried, 2016) brought this issue to a head. This article had a subtitle of "Misunderstandings about common statistical test damage science and society." By this time, the problems with the "blind faith in *P*-values" had been reported sufficiently that the scientific community was beginning to listen. And thus, a "Watershed" announcement that came out of the American Statistical Association (ASA) during March 2016 found the world ready to listen. The ASA stated clearly "While the p-value can be a useful statistical measure, it is commonly misused and misinterpreted," the statistical association report stated, continuing "this has led to some scientific journals discouraging the use of *P*-values, and some scientists and statisticians recommending their abandonment."

But the American Statistical Association apparently was aware of this *P*-value use problem, as 2 years previously in 2014 they formed an internal group to study this problem, with a goal to produce a document on the proper use of *P*-values for the guidance of researchers, practitioners, and science writers who are not statisticians. The results of this study group were finally published in June of 2016 (Wasserstein and Lazar, 2016).

The ASA study group came up with six "principles of use of *P*-values," as follows:

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 the probability that the data were produced by random chance alone.
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 ASA study group's conclusion was this:

> Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of 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. No single index should substitute for scientific reasoning.

One of the coauthors of this book has been involved with scientific medical research over a period of more than 40 years. During this time, he has been exposed over and over again to the misuse of traditional statistics in medical research publications. In Chapter 12 of this book, the authors cited, as an example, a misuse of parametric statistics in the medical subdiscipline of radiology. Eklund et al. (2017) found that in using functional magnetic resonance imaging (fMRI), there was greater than a 70% error rate (false-positives) in identifying Alzheimer's disease in over 40,000 scientific papers indexed by PubMed over the past 20 years. This does not mean that doctors cannot trust MRI results, but it does mean that one must question the validity of *conclusions* of many studies using traditional parametric statistical techniques to diagnose Alzheimer's disease. *Bottom line* is if the paper states "It has been scientifically proven," it is very likely to be false.

It is becoming increasingly clear that most research studies (at least in the medical field) analyzed with traditional statistical methods cannot be replicated. This situation in Medicine today is rather ironic, considering the fact that the need to replicate studies was the primary motivation for R.A. Fisher to develop his statistical analysis methods. Another interesting, specific example of this comes from the lab of immunologist Dr. Tim Errington at the University of Virginia's Centre for Open Science (Feilden, 2017). Dr. Errington runs "The Reproducibility Project," which since 2011 has attempted to repeat the findings reported in cancer studies. Only two of the five "landmark cancer studies" in this project have been reproducible—meaning 60% of these so called "*landmark*" studies appear bogus. Fig. 20.3 shows some of the vials of test materials involved in one of these studies.

Replication is supposed to be a hallmark of scientific integrity. The concern over this lapse of "scientific accuracy/integrity" has been growing for some time, especially over the past 3 years, such that the University of Washington (Seattle) is offering a new course for spring 2017 titled "Calling Bullshit in the Age of Big Data." Among the learning/behavioral objectives for this course is the following: After taking this course, you will be able to provide a statistician or fellow scientist with a technical explanation of why a claim or conclusion is "in error."

Below are three links that go to this new course, including a syllabus:

- http://callingbullshit.org/syllabus.html#Big;
- http://www.recode.net/2017/2/19/14660236/big-data-bullshit-college-course-university-washington;
- https://www.statnews.com/2017/02/17/science-fights-alternative-facts/

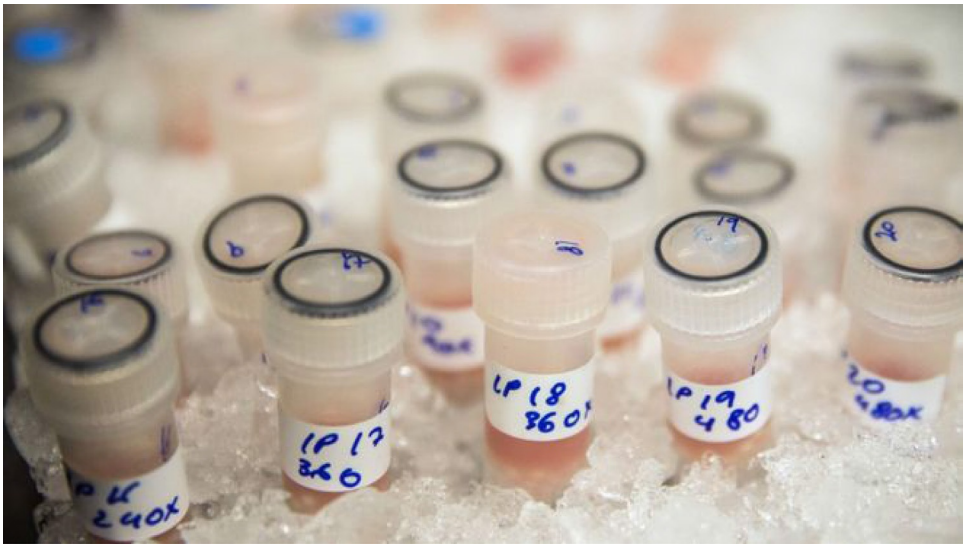

FIG. 20.3 Some of the vials of testing biomaterials analyzed among the studies reported by Tom Feilden. Only two of five "landmark cancer studies" results were confirmed in replication studies carried out at the University of Virginia. *(Getty Images) From Feilden, T., February, 2017. http://www.bbc.com/news/science-environment-39054778, http://www.bbc.com/news/science-environment-39054778.*

This University of Washington course is being taught by Carl Bergstrom, a biologist, and Jevin West, a professor in UW's Information School. After the course was announced, they woke up the next morning to chaos. They had 20,000 visitors to the course website, their mailboxes were full, and they were getting book offers. Bergstrom and West are longtime scientific collaborators who spent years grumbling about seeing inflated claims, manipulated algorithms, and twisted interpretations of scientific research, not only in the popular press but also in grant applications and scientific papers. So they decided to put together this "fun class"; they apparently "hit a nerve," as the class (capped at 160 students), filled up within 1 min of online registration. As a result, the materials are now available free online, and it is expected that lectures will be posted as well. The presence and "instant filling of students" in this course upon its announcement is just one indicator of the concern people have about being able to recognize what is true and false.

In addition to Bergstrom and West at the University of Washington, the author writing this chapter has had these concerns for decades, being a scientific researcher himself, so he finds himself in agreement with the following quote:

> What I'm finding among scientists is an uneasiness that goes back years, even decades, about an eroding appreciation of science, how it works, and how it's incorporated into our society. And it seems to be in a crescendo right now," physicist Rush Holt, chief executive of the AAAS (American Association for the Advancement of Science), told STAT NEWS in Feb, 2017. **(*Joseph, 2017*).**

Most of this "mistrust" of science has to do with scientists' misuse and misinterpretation of traditional *P*-value statistics, which were developed in the last (20th) century. Since about the year 2000, we have had at our disposal the availability of modern machine learning and

predictive analytic technologies that have the ability to get at "truth" much more cleanly. Thus, we need to embrace these methods and use them.

USUAL DATA MINING/PREDICTIVE ANALYTIC PERFORMANCE MEASURES—TERMINOLOGY

Many, if not most, of the "performance measures" that can be used in predictive analytics are listed below.

Continuous Data (Continuous Numerical) Regressions

- Mean square error (MSE)—average of the squares of the differences between the predicted and actual values
- Mean absolute error (MAE)—similar to MSE but uses absolute values instead of squaring
- Bias—the average of the differences between the predicted and actual values
- Mean absolute percentage error (MAPE)—average of the absolute errors, as a percentage of the actual values
- Correlation coefficient between actual AND predicted output (works only as a good measure if the relationship between actual and predicted is linear, which is most often not the case)
- *F* measures

Categorical Data (Yes/No; 1, 2, or 3rd Class; High/Low; etc) Classifiers

- Accuracy (as measured by the data mining/machine learning algorithm) (the percentage of the time that the "predicted class" equals the "actual class")
- Weighted (cost-sensitive) accuracy (e.g., medical diagnosis where one can be in error in one of two ways: (1) keeping a healthy person in the hospital (low cost) or (2) sending sick person home (high cost))

General Methods

When the concern is the entire population,

- Percent correct classification (PCC)—overall accuracy without regard to what type of errors are present
- Confusion matrix—provides summary of different kinds of errors
- Type I and type II errors
- Precision and recall
- False alarms and false dismissals
- Specificity and sensitivity

When the concern is a "subset" of the population (that will be "treated" or "affected"),

- Lift
- Gain
- ROC
- Area under the ROC curve (AUROC)—like measuring separation across an entire spectrum

Further definitions and descriptions of the above measures can be found at the following three citations:

- Abbott, Dean, June, 2015; http://www.predictiveanalyticsworld.com/patimes/defining-measures-of-success-for-predictive-models-0608152/5519/.
- Abbott, Dean, 2014; Applied Predictive Analytics: Principles and Techniques for the Professional Analyst, 1st Edition; Wiley.
- Abbott, Dean, 2006; http://abbottanalytics.blogspot.com/2006/11/error-measures.html.

Most Current Data Mining Software Packages Allow Ranking of Models by a Criterion Like ROC, Lift Chart, Gain Chart, or Similar

For most practitioners of predictive analytics, these are preferred way to assess model validity. Software described in this book (e.g., Statistica Data and Text Miner) includes accuracy/significance criteria like ROC, lift and gain charts, train, test, and V-fold cross validation measures so they can be easily incorporated into the modeling process. By reading the previous chapters in this book and working through some of the tutorials, the reader should get an idea of how these "performance measures" are used in predictive analytics.

UNIQUE WAYS TO TEST ACCURACY ("SIGNIFICANCE") OF MACHINE LEARNING PREDICTIVE MODELS

Predictive analytics are used to produce "models" to predict or forecast future behavior. But how reliable or "significant" (to use older *P*-value statistical terminology) are these models? In other words, how can one validate the "accuracy scores" of the models? (Garment, 2014a,b).

Three methods some of the leading data mining/predictive analytic consultants use to validate the models are discussed in detail below.

COMPARE PREDICTIVE MODEL PERFORMANCE AGAINST RANDOM RESULTS WITH LIFT CHARTS AND DECILE TABLES

The basis of this method is that lift charts and decile tables compare the results of a model with what the results would be if no model was used.

Here's how it works (SlidesShare, 2014):

1. Choose a binary (yes/no, 1 or 2, high or low, etc.) variable as the TARGET variable.
2. Randomly split lead data into two samples, using percentage of your own choosing for the data in each sample, but the following percentages are generally acceptable (if the data set is extremely large in relationship to the number of predictor variables, then a smaller percentage for the test set could be appropriate): 60% = modeling sample, 40% = hold-out sample.
3. Use data mining algorithms to find the best set of predictor variables that work in the modeling sample and identify highly responsive leads.

4. Score leads on a scale of 1–100, 100 being the most likely to convert.
5. Rank order leads by score.
6. Split leads into 10 sections (deciles).
7. Evaluate the results in a decile table.
8. These data are then plotted on a lift chart to illustrate the performance of the model.

Let's look at a specific example of a color catalog sales campaign. From historical data of the previous year, the responses to all the catalogs sent out were divided into 10 deciles, from the top decile, which got the highest percentage of buyers, to the lowest 10%, which had the fewest buyers. These results are presented in Table 20.1 and Fig. 20.4.

In Fig. 20.4, the cumulative predictive performance model is represented by the curved blue line. The diagonal red line is the performance expected purely by chance. The red X indicates the gain of the first decile above not using any modeling; thus, the gain of the first decile is 4.0 times greater than doing nothing to decide to whom the color catalogs should be sent. The green line represents the ideal, "perfect" model, where contacting only 0.8% of leads would yield 100% of sales.

One would use this approach to model performance evaluation by training models with different algorithms and comparing their performances by different model lines in the cumulative gain chart. The algorithm with the curved line located at the highest position above the diagonal red line is judged to the best model (Karl Rexer (personal communication and Garment, 2014b)). This process is more helpful than traditional statistical evaluation metrics (Rexer, pers. Com., Garment, 2014b). The machine learning model evaluation methods learn to recognize patterns in the data case by case (the way humans do it), rather than using an evaluation metric (e.g., the *P*-value) based on an average over the entire data set of cases.

TABLE 20.1 Historical Data Grouped into 10 Deciles From Highest Number of Sales to Lowest Number of Sales

Decile (based on model scroe)	Number of Leads (Hold-out Sample)	Sales	Conversoion Rate (%)	Lift (Above random sample)
1	8,000	252	3.1%	4.0
2	8,000	115	1.4%	1.8
3	8,000	70	0.9%	1.1
4	8,000	59	0.7%	0.9
5	8,000	40	0.5%	0.6
6	8,000	29	0.4%	0.5
7	8,000	31	0.4%	0.5
8	8,000	15	0.2%	0.2
9	8,000	14	0.2%	0.2
10	8,000	8	0.1%	0.1
TOTAL	80,000	632	0.8%	

The "lift" column signifies how much more successful the model is likely to be than if no predictive model was used to target leads.

From Karl Rexer, personal communication—http://www.rexeranalytics.com/, http://www.plottingsuccess.com/3-predictive-model-accuracy-tests-0114/, and http://www.kdnuggets.com/2014/02/3-ways-to-test-accuracy-your-predictive-models.html.

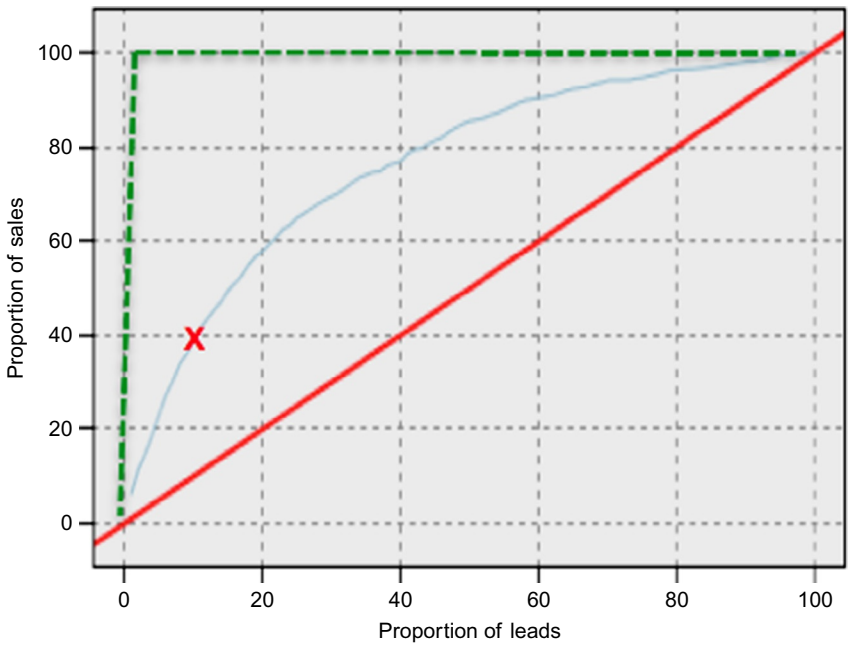

FIG. 20.4 The data in the above table can be plotted in cumulative gain chart.

EVALUATE THE VALIDITY OF YOUR DISCOVERY WITH TARGET SHUFFLING

Target shuffling is a process that reveals how likely it is for results (a "predictive analytic"/" data mining" model) to have occurred by chance. This is done by actually changing the "target variable data" from its case and putting it with a new case (or "shuffling" the target result of each case to other cases) randomly. This is done many times, maybe 500 or 1000, with the predictive analytic model run on each "new shuffle" and then all of these model's results compared. This comparison can be done most easily by graphing. If the original/real data model falls at the extreme of the "shuffled data models" outside of the 0.05 level of a normal statistical curve, then this can be called this "significance of the model" to satisfy traditional statistical thinking methods.

Fig. 20.5 shows an example of how target shuffling works, as discussed thoroughly in a LinkedIn post (SlidesShare, 2014):

1. Randomly shuffle the output (target variable) on the training data to "break the relationship" between it and the input variables.
2. Search for combinations of variables having a high concentration of interesting outputs.
3. Save the "most interesting" result and repeat the process many times.
4. Look at a distribution of the collection of bogus "most interesting results" to see how much of apparent results can be extracted from random data.
5. Evaluate where on (or beyond) this distribution your actual results stand.
6. Use this as your "significance" measure.

Further information on "target shuffling" can be found in a video by Elder (2017).

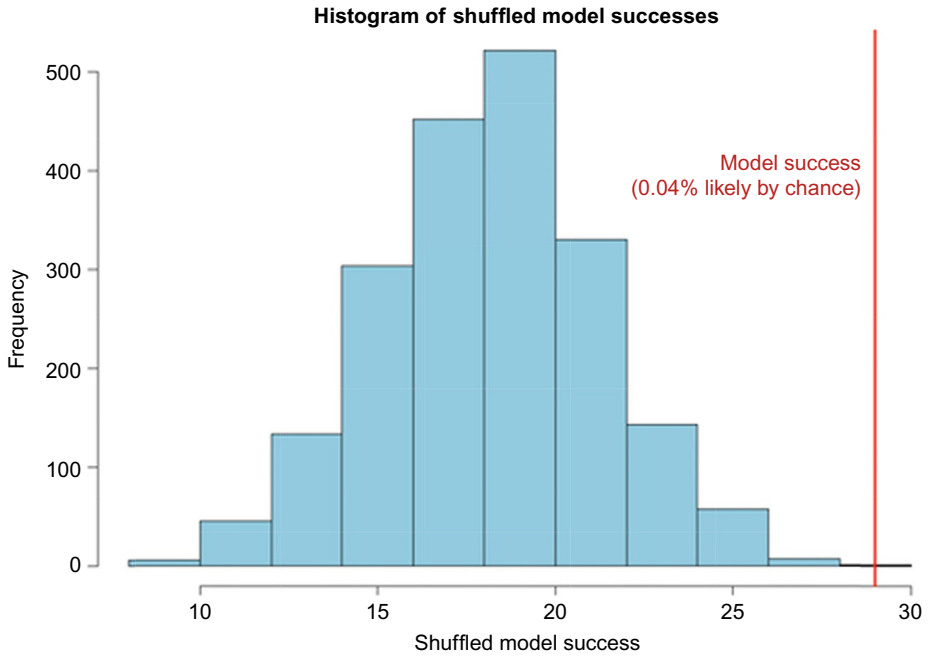

FIG. 20.5 An example of TARGET SHUFFLING results graphed; in the histogram pictured above, the original predictive analytic model scored in the high 20's. Only 0.04% of the random, shuffled models performed better, meaning the model is significant to that level and would meet the criteria of a publishable result in any journal that requires "significance" at the traditional *P*-value of <0.05% (John Elder, personal communication; and Garment, 2014a,b).

TEST PREDICTIVE MODEL CONSISTENCY WITH BOOTSTRAP SAMPLING

Bootstrap sampling tests a model's performance on certain of the data over and over again to provide an extra estimate of accuracy. As Dean Abbott states (*personal communication*—http://www.abbottanalytics.com/; Garment, 2014a,b; and SlidesShare, 2014) that *he* uses this method to test the consistency of his predictive models and to determine if they're not just statistically significant, but *operationally significant*. "You can have a model that is statistically significant, but it doesn't mean that it's generating enough revenue to be interesting or valuable….." he explains. Data miners and predictive analytic modeling consultants usually do not use the type of traditional statistics taught in college classes to analyze their predictive analytic models but instead use data to validate the models. Bootstrapping is a way to iteratively do this and can even be used effectively many times with small data sets if "small" is the only data available.

Here's one way that bootstrap sampling works:

1. Take a random sample of data and split it into three subsets: training, testing, and validation.
2. Build model on the training subset.

3. Evaluate model on the testing subset.
4. Repeat this training and testing process several times using randomly chosen sets of the data.
5. Once you're convinced your model is consistent and accurate, deploy it against the final validation subset.

Another way to do bootstrapping is with V-fold cross validation methods. This is automatically available in the Statistica data mining software illustrated in previous chapters of this book and as follows:

1. Take a random sample of data and split it into three subsets: training, testing, and a V-fold cross validation sampling. V in V-fold represents how many times one will take cross validation subsamples; if V is set to 10, then 10 separate subsamples of the data are taken and run through the model. But V can be set to 100, 1000, or whatever. Generally, 10 is satisfactory.
2. Build model on the training subset, getting a "test set accuracy score."
3. Evaluate model on the testing subset, getting a "test set accuracy score."
4. Evaluate the model on the V-fold cross validation, getting a "V-fold cross validation accuracy score."
5. If all three of these accuracy scores are about the same, then we have a good model that should perform on new data with this accuracy.

For readers who want to further study these issues of "false significance" and "random noise in data" that can ensure that scientific discoveries are untrustworthy, the following references are suggested:

- Siegel, E., 2016. http://www.predictiveanalyticsworld.com/book/ (Chapter 3).
- Siegel, E., 2014. http://www.predictiveanalyticsworld.com/patimes/breakthrough-avert-analytics-treacherous-pitfall/3366/, where one can read about "Are Orange Cars Lemons".
- Gelman, A., Fung, K., 2016. http://www.slate.com/articles/health_and_science/science/2016/01/amy_cuddy_s_power_pose_research_is_the_latest_example_of_scientific_overreach.html.

POSTSCRIPT

This chapter presents the final piece of what might have appeared to you as a puzzle in the name of this book. This "handbook of statistical analysis and data mining applications" is a comprehensive presentation of the elements of data mining analysis, but not for statistical analysis. Rather, this book (and particularly this chapter) has presented some of the comparisons between the methods and credibility of traditional statistical analysis and data mining (predictive analytics) methods for building models of patterns in data sets. There are right ways to use traditional statistical analysis methods, but few researchers in science or medicine perform them or evaluate them properly. This problem is one of the reasons we wrote this book.

References

Bower, B., 2015. Psychology results evaporate upon further review. Science News. https://www.sciencenews.org/article/psychology-results-evaporate-upon-further-review.

Eklund, A., Nichols, T., Knutssen, H., 2017. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. PNAS 113 (33), 7900–7905.

Elder, J., 2017. http://www.elderresearch.com/target-shuffling-video.

Feilden, T., 2017. Most scientists 'can't replicate studies by their peers'. BBC News: Science & Environment, 22 February 2017. http://www.bbc.com/news/science-environment-39054778.

Garment, V., 2014a. 3 ways to predict the accuracy of your predictive models. KDNuggets. http://www.kdnuggets.com/2014/02/3-ways-to-test-accuracy-your-predictive-models.html.

Garment, V., 2014b. 3 ways to test the accuracy of your predictive models. Plotting Success. http://www.plotting-success.com/3-predictive-model-accuracy-tests-0114/.

Gartner, 2017. http://www.gartner.com/it-glossary/predictive-analytics/.

Gelman, A., Fung, K., 2016. http://www.slate.com/articles/health_and_science/science/2016/01/amy_cuddy_s_power_pose_research_is_the_latest_example_of_scientific_overreach.html.

Gigerenzer, G., Marewski, J., 2015. Surrogate science: the idol of a universal method for scientific inference. J. Manage. 41 (2), 421–440. http://journals.sagepub.com/doi/full/10.1177/0149206314547522 https://doi.org/10.1177/0149206314547522.

Joseph, A., 2017. In Trump era, a leading science group exhorts its members: do not 'retreat to the microscope'. Science News. https://www.statnews.com/2017/02/16/aaas-qa-trump-science/.

Murray, D., Pais, S., Biltstein, J., Alfano, C., Lehman, J., 2008. Design and analysis of group-randomized trials in cancer. J. Natl. Cancer Inst. 2008, 483–491.

Nosek, B., 2015. Estimating the reproducibility of psychological science. Science 349 (6), 251. https://doi.org/10.1126/science.aac4716.

Saey, T.H., 2015a. Year in review: scientists tackle the reproducibility problem. Science News. https://www.sciencenews.org/article/year-review-scientists-tackle-irreproducibility-problem.

Saey, T.H., 2015b. Is redoing scientific research the best way to find truth? Science News. https://www.sciencenews.org/article/redoing-scientific-research-best-way-find-truth.

Siegel, E., 2016. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. Wiley, Hoboken, NJ (Chapter 3).

Siegfried, T., 2015a. Science is heroic, with a tragic (statistical) flaw. https://www.sciencenews.org/blog/context/science-heroic-tragic-statistical-flaw.

Siegfried, T., 2015b. Top 10 ways to save science from its statistical self: null hypothesis testing should be banished, estimating effect sizes should be emphasized. https://www.sciencenews.org/blog/context/top-10-ways-save-science-its-statistical-self.

Siegfried, T., 2016. Experts issue warnings on problems with P values. Science News. https://www.sciencenews.org/blog/context/experts-issue-warning-problems-p-values.

SlidesShare, 2014. 3 tests experts use to validate predictive model accuracy. https://www.slideshare.net/SoftwareAdvice/3-tests-experts-use-to-validate-predictive-model-accuracy.

Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on p-values: context, process, and purpose. Am. Stat. 70 (2), 129–133. https://doi.org/10.1080/00031305.2016.1154108. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108.