A comparative study of frequentist vs Bayesian A/B testing in the detection of E-commerce fraud

James Christopher Westland
Department of Information and Decision Sciences, University of Illinois Chicago, Chicago, Illinois, USA

Received 21 July 2022
Revised 16 September 2022
Accepted 17 September 2022

Abstract

Purpose – This paper tests whether Bayesian A/B testing yields better decisions that traditional Neyman-Pearson hypothesis testing. It proposes a model and tests it using a large, multiyear Google Analytics (GA) dataset.

Design/methodology/approach – This paper is an empirical study. Competing A/B testing models were used to analyze a large, multiyear dataset of GA dataset for a firm that relies entirely on their website and online transactions for customer engagement and sales.

Findings – Bayesian A/B tests of the data not only yielded a clear delineation of the timing and impact of the intellectual property fraud, but calculated the loss of sales dollars, traffic and time on the firm's website, with precise confidence limits. Frequentist A/B testing identified fraud in bounce rate at 5% significance, and bounces at 10% significance, but was unable to ascertain fraud at the standard significance cutoffs for scientific studies.

Research limitations/implications – None within the scope of the research plan.

Practical implications – Bayesian A/B tests of the data not only yielded a clear delineation of the timing and impact of the IP fraud, but calculated the loss of sales dollars, traffic and time on the firm's website, with precise confidence limits.

Social implications – Bayesian A/B testing can derive economically meaningful statistics, whereas frequentist A/B testing only provide *p*-value's whose meaning may be hard to grasp, and where misuse is widespread and has been a major topic in metascience. While misuse of *p*-values in scholarly articles may simply be grist for academic debate, the uncertainty surrounding the meaning of *p*-values in business analytics actually can cost firms money.

Originality/value – There is very little empirical research in e-commerce that uses Bayesian A/B testing. Almost all corporate testing is done via frequentist Neyman-Pearson methods.

Keywords Statistics, Fraud, A/B testing, Cyber-crime

Paper type Research paper

1. Introduction

Fraud, in all of its variety, is one of the major concerns in electronic commerce. Electronic commerce obviates various "middle-men" in retailing – e.g. cashiers, salespersons, security guards and so forth. Advantages of highly efficient transaction processing un-monitored by humans, opens numerous pathways to commit fraud. As a consequence, electronic commerce continually seeks data analytical approaches to replacing the security provided by humans, while retaining the efficiency and cost advantages of digital platforms. A/B testing can provide one such analytical approach to fraud detection. Numerous tools are available in a highly competitive and expanding market, including tools by VWO Corporation, Optimizely, Convert Experiences, SiteSpect, AB Tasty, Evolv, Google Optimize, Qubit, Adobe Target and others. Current software generally takes a frequentist approach.

Journal of Electronic Business & Digital Economics
Vol. 1 No. 1/2, 2022
pp. 3-23
Emerald Publishing Limited
e-ISSN: 2754-4222
p-ISSN: 2754-4214
DOI 10.1108/JEBDE-07-2022-0020

Auditors define control systems over fraud in three categories: preventive, detective and corrective (Westland, 2020a). Preventive controls prevent a fraud from happening, and are typically passive in nature. Detective controls detect that a transaction or group of transactions has a higher probability of fraud, and should be investigated. Corrective controls recognize that error correction is highly error prone, and control over corrections is necessary for most systems (Pumsirirat & Liu, 2018; Al-Shabi, 2019; Westland, 2000; Westland, 2002, 2004, 2017, 2020a). A/B tests provide a corrective control over groups of transactions, and their control implementations are similar to that of autoencoders, Benford tests, Sarbanes-Oxley tests and other detective control implementations (Westland, 2019, 2020b). In turn, the operations for fraud detection are generally the responsibility of internal audit and technical security staff in e-commerce operations. Auditors' responsibility is not the complete elimination of fraud; rather their main goal is reducing the cost of fraud to acceptable levels. The trade-offs in this activity are assessed in terms of the cost of fraud detection and control vs the savings from frauds that have been prevented or recovered from; in other words, the expected net savings from fraud prevention.

The search for improved tools for fraud detection is of immense importance to e-commerce firms today. Fraud has been steadily on the rise as more and more commerce transactions have migrated to digital platforms. Rice, Weber, and Wu (2014), Ashbaugh-Skaife, Collins, Kinney, and LaFond (2008), Ge, Koester, and McVay (2016), Bedard and Graham (2011) and Bedard, Hoitash, and Hoitash (2009) have provided compelling evidence that firms with weak internal controls suffer increased numbers of frauds. Recent information e-commerce fraud have grown costlier and more frequent; for example Home Depot's 2015 breach has cost it $232m so far an amount that they expect to reach billions. A 2015 breach of Ashley Madison stole 40m accounts including photos details of sexual proclivities and personal addresses Target's 2013 breach affected the accounts of 70m customers and so far has cost the firm $162m in added expense. A 2018 breach of Marriott hotels exposed private information of ~500m customers; a 2019 breach of Capital One exposed financial information of ~100m credit applicants; and a 2017 breach at Equifax exposed financial information of 143m customers. In 2014 A Guardians of Peace breach of Sony Pictures stole over 100 terabytes of confidential data. 2014 also saw the theft of 360m MySpace accounts a LinkedIn hack that took more than 100m accounts a 500-million-accounts, a hack of Yahoo, theft of 340m AdultFriendFinder accounts, their second hack in a year and numerous other breaches. Both frequency and scale of breaches have grown dramatically in the recent past.

This research conducts an empirical study using a large, multiyear dataset of Google Analytics (GA) data for a particular firm's website during a particular period of time, and which was compromised by other fraudulent websites that appropriated the firms' brand and e-commerce transactions for a portion of that time. Assessments of "fraud" and "no-fraud" are obtained through applications of the firm's actual loss function and the calculation of loss under competing decisions from a frequentist and Bayesian A/B test decision based on the empirical data. The approach is tested on six years of GA e-commerce data obtained from a major service organization that relies entirely on their website and online transactions for customer engagement and sales.

To this date, A/B testing has not been a standard tool for fraud detection, though its implementation is similar enough to other methods that A/B testing can be used to detect, prevent and recover from fraud in the same way as we have already used autoencoders (Pumsirirat & Liu, 2018; Al-Shabi, 2019; Westland, 2019, 2020b) and generalist algorithms such as the Fraud Aware Impression Regulation system (Li *et al.*, 2019). The current research analyzes and statistically tests a large database using Bayesian vs frequentist approaches to A/B tests for fraud detection. Thus the research serves as an empirical demonstration that A/B testing is effective as a fraud detection methodology, and in addition tests the effectiveness of Bayesian vs frequentist approaches to fraud detection. We used six years of GA data for a

website during two time periods: one in which a fraudulent website was fraudulently drawing off and misleading customers and another period where this fraud was not being perpetrated. This allowed us to label particular sets of transactions as fraudulent.

The presentation in this research paper proceeds as follows. Section 2 reviews the prior literature in fraud research. Section 3 introduces the research dataset, its curation and analysis. Section 4 reviews the mathematics of A/B testing methodologies; Section 5 conducts the analysis and reviews the results. Section 6 draws conclusions and offers a brief discussion. Section 7 looks at potential application so of the results and their managerial implications. Finally, Section 8 suggests limitations and future research.

2. Prior research in electronic commerce fraud detection

Fraud detection may either be supervised or unsupervised – i.e. requiring datasets that are at least partly labeled, versus being completely unlabeled. Supervised methods generate a predictive probability that a new case, or a set of cases, is fraudulent. Classification methods (Hand & Henley, 1997; Jha, Guillen, & Westland, 2012; Ahfock, McLachlan, Yang, & Zhu, 2022; Jha & Westland, 2013) such as linear discriminant analysis and logistic discrimination, have proved to be effective tools for many applications, but more powerful tools (Ripley & Ripley, 2001; Bolton & Hand, 2001; Bolton & Hand, 2002; Webb, Campbell, Schwartz, & Sechrest, 1999) such as neural networks are being applied. Rule-based methods, though dated, are still being applied. These are supervised "IF-THEN-ELSE" learning algorithms that produce classifiers. Examples of such algorithms include the BAYES implementation of the CN2 induction algorithm (Clark & Niblett, 1989), the FOIL implementation of decision tree algorithms (Quinlan, 1990) and the RIPPER evolution of IREP and C4. 5 machine learning rules (Cohen, 1995). Tree-based algorithms such as CART: Classification And Regression Trees (Breiman, Friedman, Olshen, & Stone, 1984) produce classifiers of a similar form. Combinations of some or all of these algorithms can be created using meta-learning algorithms to improve prediction in fraud detection (Chan, Fan, Prodromidis, & Stolfo, 1999).

Some work has addressed misclassification of training samples (e.g. Chhikara & McKeon, 1984) but not in the context of fraud detection. Social acquaintance analysis relating known fraudsters to other individuals using record linkage and social network methods has been proposed for some time (Wasserman & Faust, 1994) but only recently have graph analytic tools become available to really make use of this method (e.g. see Pourhabibi, Ong, Kam, & Boo, 2020; Hooi *et al.*, 2016; Zhang *et al.*, 2022; Cheng, Wang, Zhang, & Zhang, 2020).

Unsupervised methods are used when there are no prior sets of legitimate and fraudulent observations. Techniques employed here are usually a combination of profiling and outlier detection methods. Benford's law (Berger & Hill, 2011; Hill, 1995) is popularly used since it is first proposal as a fraud detection tool by Hal Varian, and is a common test in the U.S. Securities and Exchange Commision's audits of corporations to find fraud in transaction streams (Westland, 2020a).

Fraudsters adapt to new prevention and detection measures, and various methods have been proposed to help fraud detection be more adaptive and evolve over time (e.g. see Cortes, Pregibon, & Volinsky, 2001; Senator, 2000).

There is a rich literature in hardware, software and administrative systems for fraud control. Though less flexible than data analytic methods, such methods can prevent fraud before it occurs, and thus lower the economic cost of surveillance. E-commerce fraud manifests in several forms, with credit card fraud being most prevalent and having the largest economic impact. Research has investigated credit card fraud detection using a behavior certificate (BC) (Zheng *et al.*, 2018) to determine the legality of transactions based on historical records of the cardholder. Fraud detection has also used disposable domain names that can detect fraud based on IP masking (Laurens, Rezaeighaleh, Zou, & Jusak, 2019) and

detection of fraudulent transactions using a prudential multiple consensus (PMC) models (Carta, Fenu, Recupero, & Saia, 2019). Research has explored the use of blockchains to prevent fraud through implementation of cryptocurrency in payments and smart contracts (Savita & Datta, 2015) and through various antifraud systems to detect e-commerce frauds (Xie *et al.*, 2018). Various research studies have investigated security in data transactions, passwords, networks, images and trading in e-commerce, finding that fraud can be prevented or ameliorated by using RSA (Rivest–Shamir–Adleman) encryption and Fernet cipher algorithms (Dijesh, Babu, & Vijayalakshmi, 2020). Other studies suggest how risks in e-commerce can be detected quickly and accurately without disrupting system performance (Xu & Chu, 2017), by using a security service oriented architecture (SOA) framework that can protect e-commerce from attacks or threats (Suryono, Purwandari, & Budi, 2019) and using a unified framework to analyze the security data in e-commerce. Other studies have looked at security of one-time password (OTP) using ECC (elliptic curve cryptography) with palm vein biometrics to OTP (Dzulfikar, Sensuse, & Noprisson, 2017). It has been suggested that an e-commerce trusted trading framework (ETTF) using blockchain can improve security in e-commerce (Luhach, Dwivedi, & Jha, 2014) and two-way authentication based on visual cryptography and steganography can further protect e-commerce from fraud (Ismanto, Ar, Fajar, Bachtiar, & others, 2019). Research has also studied graphical passwords for e-commerce applications that can improve the security and usability of customers (Qiu & Li, 2017). Mahto and Yadav (2015) proposed a unified framework for securing image data stored in third-party clouds and Sharma, Mathur, and Srivastava (2018) proposed a system that combines text-based steganography, visual cryptography and OTP can avoid identity theft and customer data privacy. All of these systems, though, required substantial upfront investments in hardware, software and administrative systems before they can be effective. They are also difficult to modify or improve based on experience – if they turn out to be ineffective or are hacked; their fraud control value is often substantially reduced.

Bedard *et al.* (2009), Hoitash, Hoitash, and Bedard (2009) and Bedard and Graham (2011) examined detection and severity classification of internal control deficiencies, finding that external auditors, during their Section 404 audit, detect about three-fourths of unremediated internal control deficiencies. Ge *et al.* (2016) looked at a sample of 261 companies that disclosed at least one material weakness in internal control in their Sarbanes-Oxley (SOX) filings, finding that poor internal control is usually related to an insufficient commitment of resources for accounting controls, with the most common account specific material weaknesses occurring in accounts receivable and inventory. SOX 302 disclosures, in contrast, tended to describe internal control problems in complex accounts such as the derivative and income tax accounts. They found that disclosing a material weakness is positively associated with business complexity, e.g. multiple segments, and foreign currency negatively associated with firm size, e.g. market capitalization, and negatively associated with firm profitability metrics, e.g. return on assets. Lin, Pizzini, Vargus, and Bardhan (2011) investigated the role that a firm's internal audit function plays in the disclosure of material weaknesses reported under SOX 404, using data from 214 firms. They found that material weakness disclosures are negatively correlated with the education level of the internal auditors and positively correlated with the practice of grading audit engagements and external-internal auditor coordination. Ashbaugh-Skaife *et al.* (2008) reported that SOX disclosed internal control deficiencies were associated with more complex operations, recent organizational changes, greater accounting risk, more auditor resignations and have fewer resources available for internal control. They also found that firms with SOX disclosed internal control deficiencies had more prior SEC enforcement actions and financial restatements, were more likely to use a single dominant audit firm and had more concentrated institutional ownership. Feng, Li, McVay, and Skaife (2014), Feng, Li, and McVay (2009) and Berger, Li, and Wong (2005) found that internal control deficiencies were correlated with less accurate guidance. In

particular the impact of ineffective internal controls on forecast accuracy was found to be three times larger when the weakness was related to revenues or cost of goods sold. This finding reflects the importance of revenues and cost of goods sold in forecasting earnings. Ashbaugh-Skaife *et al.* (2008) found that firms that report internal control deficiencies have lower quality accruals, as measured by accrual noise and absolute abnormal accruals, when compared to firms not reporting internal control problems. Additionally, firms whose auditors confirm remediation of previously reported internal control deficiencies exhibit an increase in accrual quality relative to firms that do not remediate their control problems. They further found that material weaknesses are correlated with: (1) noise, (2) with accrual noise higher error term variance and (3) with intentional misstatements that bias earnings upward.

3. Dataset, curation and analysis

Acquisition of data for the analysis in this report started with the firm's financials, from the inception of the company in 2016, up to March 18, 2021. Sales and operations data were obtained from the firm's financial statements, and internal accounting datasets for monthly and daily sales. Website statistics were obtained through GA application programming interfaces (API) for all available data between the firm's inception in 2016 and March 18, 2021. This included the period that the fraudulent sites and Uniform Resource Locators (URL) were in operation, between April 7, 2018 and May 13 2020 (this period covered 767 days or 25.56 months), as well as "control" periods before and after that period which were used to determine the firm's baseline operations and web statistics. The complete dataset was extracted from GA APIs over a period of three weeks (allowing for throttling) and tracked the complete web history for the website between "2016-09-08" and "2021-03-18". Figure 1 delineates the steps involved in obtaining, aggregating and analyzing the dataset. I have abbreviated for brevity in Figure 1, but not that "GA" in the figure means GA; optical character recognition (OCR); API; and time series (TS). These activities yielded a total of 7,238,819 useable records have the "curation-cleanup" step.

The specific interpretation and methodological choices in my curation, analysis and calculation involving the firm's financial data have been motivated by:

(1) a choice of the most objective and least subjective methodology that will insulate data and analysis from investigator bias, while providing objective, verifiable and replicable conclusions;

(2) the elimination of effects due to confounding and unobserved variables; and

(3) the choice of the most appropriate methods and prior assumptions that would allow the firm's financial data and the analysis based upon that data, to speak for itself without imposing any investigator bias.

Any comparison of operations between two time periods needs to first remove inflation, seasonality and organic business growth in order to isolate cogent effects. The deflated values are stated in 2016 dollars. Restatement in the period from 2016-2021 dollars used the deflators and reinflators calculated in the technical appendices. Failure to detrend data would have added spurious effects. The current analysis deseasonalized and detrended data to remove extraneous influences of inflation, seasonality and organic growth in the firm's business. I controlled for unobserved covariant predictors, where this was necessary, through mixed-effects models. Interpolations were used where needed, using industry best-practice cross-sectional and TS methods. Firm's business model was neither particularly complex, nor were their operations and revenue flows particularly volatile. In my opinion, the interpolations applied to compute missing data points provided accurate estimates.

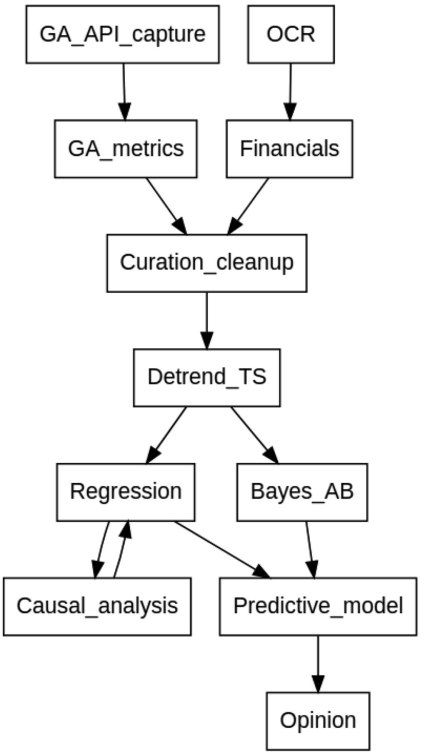

Figure 1.
Workflow schematic of analysis in this research

A model of future sales using GA predictors was constructed and optimized around an ordinary least squares (OLS) regression model. The model used only detrended TS, as in particular time-series autocorrelations would have inflated the variation in sales explained by GA predictors to an $R^2 = 90\%$, whereas the true variation explained by only GA predictors is $R^2 = 60\%$. Detrending and deseasonalizing data provided conservative (i.e. tending towards underreporting the loss of sales due to the actions of fraudulent sites) estimates of loss. The OLS regression model used in this analysis assumed normal residuals. A complete analysis of empirical residuals validated this assumption and provided the justification for the use of a normal inverse gamma prior in the A/B analysis.

The data analysis revealed causation between the firm's sales and web traffic, as measured in their GA statistics, and the existence of fraudulent websites and URLs. Furthermore, causation of the firm's sales decline due to the negative impact of fraudulent sites and URLs on the firm's web traffic is strongly supported by the time sequence of events (Granger, 1969). After the initiation of fraudulent websites and URLs, the firm's business sales, website visits and time on website declined significantly. I tested the results of the OLS regression model of sales using GA predictors to determine whether the regression coefficients represented causality or correlation. I used a Granger causality test - the industry best-practices statistical hypothesis test for ascertaining causal effects in economic studies. The Granger causality test determines whether one TS is useful in forecasting another. Causality in economics can be tested for by measuring the ability to predict the future values of TS using prior values of another TS. In this study, TS of GA predictors was shown to

Granger-cause (Granger, 1969) A TS of sales by applying a series of *t*-tests and *F*-tests on lagged values of GA predictors, with lagged values of sales also included, to show that those GA values provide statistically significant information about future values of sales.

Frequentist and Bayesian A/B testing assumed a normal distribution of data. It assumed a diffuse, noninformative prior, a normal inverse gamma – Normal conjugate family and computed loss as the difference of posterior means, assessing loss estimates on 95% credible intervals. Bayesian A/B test results were highly significant, and all needed methodologies to control potential estimation biases and confounding effects were applied to maintain strict control over the results. The negative impact of fraudulent websites during the period of their existence, on the firm's sales is captured in their negative impact in the recorded GA site metrics for the firm website. Bayesian A/B testing of the sales data during the period when the fraudulent sites existed, versus the other periods when there were no fraudulent sites operating shows that during the time the fraudulent sites were active. The specific factorsanalyzed are defined in Tables 1 and 2 estimates the impact of these factors on the firm's economic performance using regression analysis. The firm's website is the major source of sales revenue. I built a regression model to predict firm sales from their GA statistics, which explains around 60% ($R^2 = 58.03\%$) of the variance in sales, the remaining variance arising from other nonwebsite influences.

Interpolations were used to complete the dataset, using industry best-practice cross-sectional and time-series methods. The firm's business model was neither particularly

GA factor and distribution	Description
Modeled with Inverse Gamma-Normal conjugate family	
SessionDuration (normal)	Time user spent on firm site
sessionsPerUser (normal)	Number of times a particular user visited the firm's site
daysSinceLastSession (normal)	Days since a user last visited firm
avgTimeOnPage (normal)	Average of time user spends on a page
pageviewsPerSession (normal)	Number of page views inside the firm's site for each session
uniquePageviews (normal)	New page views
Modeled with Beta-Binomial Conjugate family	
entranceRate (beta)	Entrances and entrance rate (clicks to landing page)
bounceRate (beta)	Bounces and bounce rate (left after landing page)
percentNewSessions (beta)	New sessions (new users)
exitRate (beta)	Exits

Table 1.
Google analytics factors for Bayesian A/B tests

GA.Metric	Unit.Cntr	Std.error	*t*.stat	*p*.value	B.A.mean	CI.5	CI.95
(Intercept)	−$38,732.09	42071.22	−0.92	0.36	NA	NA	NA
new_vis	$84.44	36.82	2.29	0.03	2472	−4.48	−0.81
users	−$1.71	0.45	−3.80	0.00	−61415	−8.81	6.24
bounces	−$33.99	10.74	−3.17	0.00	1581	−5.58	3.62
bounceRate	−$2.16	0.70	−3.09	0.00	184973	−4.20	−0.87
sessionDuration	−$0.52	0.26	−2.02	0.05	1982081	−7.64	5.09
avgSessionDuration	$0.48	0.26	1.84	0.07	2023928	−7.66	5.17
uniqueDimensionCombinations	$43.70	11.85	3.69	0.00	159	−5.96	3.83
entranceRate	−$0.30	0.19	−1.60	0.12	−113215	−8.19	5.19
timeOnPage	−$0.33	0.12	−2.86	0.01	57039	−9.80	7.37
avgTimeOnPage	$0.79	0.16	4.92	0.00	104429	−9.16	7.00

Table 2.
Regression estimates

complex, nor were their operations and revenue flows particularly volatile. In my opinion, the interpolations applied to compute missing data points provided accurate estimates. The deflated values are stated in 2016 dollars. Restatement in the period from 2016–2021 dollars can be done using the reinflators calculated here.

The following graphs extract trends, seasonalities and random fluctuations from the figure of merit analyzed in this section.

Where GA factors are count data, Bayesian conjugate distribution model was assumed to be beta-binomial with prior hyperparameters (α β) and posterior hyperparameters $\left(\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \alpha \sum_{i=1}^{n} x_i\right)$ for α successes and β failures for a sample $\{x_i\}$ of N_i observations. I also assumed that volume was sufficient (I had ~8m observations at my disposal) that central limit theorem convergence easily allows me to assume the beta-binomial data converges to Inverse Gamma – normal conjugate family. Closed-form posterior probabilities for the beta-binomial can be computed:

$$p_{SS} \overset{SS}{\sim} Beta(\alpha_{SS}, \beta_{SS})$$

$$p_{noSS} \overset{noSS}{\sim} Beta(\alpha_{noSS}, \beta_{noSS})$$

$$Pr(p_{noSS} > p_{SS}) = \sum_{i=0}^{\alpha_{noSS}-1)} \frac{(B(\alpha_{SS} + i, \beta_{SS} + \beta_{noSS})}{(B(\beta_{noSS}) + i)B(1 + i, \beta_{noSS})B(\alpha_{SS}, \beta_{noSS})}$$

central limit theorem convergence was validated through exploratory testing of data and models and is used to support the assumption that the Bayesian model is assumed to be normal-normal with prior hyperparameters (μ, τ) and posterior hyperparameters $\left(\frac{\tau_0 \mu_0 + \tau \sum_{i=1}^{n} x_i}{\tau_0 + n\tau},\ \tau_0 + n\tau\right)$ for sample $\{x_i\}$

4. A/B testing methodologies

A/B tests compare two sample proportions and require that there are two groups and that the data for each participant are dichotomous (Little, 1989). A/B testing has been applied in situations with three different objectives:

(1) Making a binary choice based on a critical value of some sort as an alternative to hypothesis testing or event studies,

(2) Performance ranking of two (or more) alternatives such as ranking marketing or pricing strategies, and

(3) Risk assessment such as identifying transactions, investments, customers, strategies and so forth that face more or less risk.

The context used in this paper applies A/B testing for risk assessments, and falls into the categories of supervised classifiers discussed in the prior literature section, e.g. similar to the usage of classifiers in Hand and Henley (1997), Jha *et al.* (2012), Ahfock *et al.* (2022) and Jha and Westland (2013) as well as more recent neural network based classifiers for fraud detection (e.g. see Westland, 2020b; 2019).

A/B tests, in one form or another, have been an essential part of scientific marketing since the pioneering work of Claude Hopkins (Hopkins, 1923; Schorman, 2008) in the early 20*th* century. The A/B test is standard operating procedure for the analysis of clinical trial data, where study participants are randomly allocated to one of two experimental groups (typically called A and B). A/B tests are common in fields such as biology, psychology and conversion

rate optimization in online marketing. Similar, but less developed protocols are used in finance, accounting and law, where they are called event studies, and control and treatment groups are typically partitioned based on time of the event. Where A/B testing is applied, researchers attempt to statistically infer whether and to what extent the experimental condition has a higher success rate than the control condition. Practitioners can choose between the frequentist and the Bayesian approaches, though because of mathematical complexity, the Bayesian approach is seldom used. Nonetheless Gronau, Raj, and Wagenmakers (2019) argue for the superiority of the Bayesian approach because:

(1) evidence can be obtained in favor of the null hypothesis;

(2) evidence can be updated continually, as the data accumulate; and

(3) expert knowledge can be taken into account.

A/B testing is a "Natural Experiment" involving customer experience which today has standardized on the industry best-practice Bayesian A/B tests. Earlier studies applied frequentist A/B testing (bucket testing or split-run testing) which compared two versions of a subject's response to variant A against variant B, and constructs a Neyman-Pearson hypothesis test assessed on p-values. Frequentist A/B tests are inappropriate in loss calculation, since they cannot generate loss numbers, only a "yes/no" assessment of whether losses have occurred. Bayesian A/B tests are commonly used for understanding user engagement and satisfaction of online features. Large social media sites like LinkedIn, Facebook and Instagram continually employ Bayesian A/B testing to make user experiences more successful and as a way to monetize their services. Bayesian A/B tests are the preferred method for binary comparisons in marketing campaigns, business strategies and operations choices in industry. Bayesian A/B tests do not require the analyst to claim an unreasonable level of prior knowledge of events and their consequences, as for example, does the positing of hypotheses for frequentist A/B tests. Bayesian A/B testing allows the data to speak for itself, free of human biases.

Practitioners of frequentist A/B tests predominantly use $p-value$ and Neyman-Pearson hypothesis significance testing, which fails to meet the standards prescribed in Gronau *et al.* (2019) and Box (1987). Of most concern is the fact that frequentist A/B tests cannot distinguish between absence of evidence and evidence of absence (Keysers, Gazzola, & Wagenmakers, 2020; Robinson, 2018). Evidence of absence means that the data support the hypothesis that there is no effect (i.e. the two conditions do not differ); absence of evidence, however, means that the data are inconclusive (Altman & Bland, 1995). With Neyman-Pearson tests, the data cannot be tested sequentially without necessitating a correction for multiple comparisons that depends on the sampling plan; this problem is delineated in Berger and Wolpert (1988), Wagenmakers (2007) and Wagenmakers *et al.* (2018a, b). Camerer *et al.* (2018) found that poor replicability do to frequentist approaches is especially a problem in social science, finding that replicability varies between 57% and 67%) for studies relying on complementary replicability indicators. In academic research, the low replicability of social science outcomes may be considered a curiosity; but in business and clinical trials, it can mean life or death differences for individuals and firms.

Many researchers in online marketing believe that it is efficient to act as soon as the data provide evidence that is sufficiently compelling; and frequentist A/B test practitioners repeatedly peek at interim results and stop data collection as soon as the $p-value$ is smaller than some predefined α-level (Goodson, 2014; Stolberg, 2006). However, this practice inflates the Type I error rate which in practice invalidates Neyman-Pearson hypothesis testing (Jennison & Turnbull, 1990; Wagenmakers, 2007). Additionally, Neyman-Pearson testing does not allow marketing professions to incorporate detailed expert knowledge. For example,

online advertising campaigns often yield minuscule increases in conversion rates because of poor reliability of statistical decisions (Johnson, Lewis, & Nubbemeyer, 2017). In contrast, the Bayesian framework is conceptually straightforward, incorporates expert knowledge and results in more informed statistical analyses (Lindley, 1993). Limitations in frequentist statistics can be overcome by adopting a Bayesian data analysis approach (Kamalbasha & Eugster, 2021) as described in the following synopsis of the method (Doorn *et al.*, 2021).

Let n_A denote the total number of observations and y_A denote the number of successes for group A. Let n_B denote the total number of observations and y_B denote the number of successes for group B. The commonly used Bayesian A/B testing model is specified as follows:

$$y_A \sim \text{Binomial}(n_A, \theta_A)$$

$$y_B \sim \text{Binomial}(n_B, \theta_B)$$

This model assumes that y_A and y_B follow independent binomial distributions with success probabilities θ_A and θ_B. These success probabilities are assigned independent beta(α, β).

For example, with the data in hand one may find that $\rho = 0.15$, and that the power to detect a minuscule effect was only 0.20. However, power is a predata concept and consequently it remains unclear to what extent the observed data affect our knowledge (Wagenmakers *et al.*, 2015). Moreover, the selection of the minuscule effect is often motivated by Bayesian considerations (i.e. it is a value that appears plausible, based on substantive domain knowledge). A particularly convenient conjugate family of distributions is the Beta distribution – whenever a Beta prior is used and the observed data are binomially distributed, the resulting posterior distribution is also a Beta distribution. Specifically, if the data consist of s successes and f failures, the resulting posterior beta distribution equals $Beta(\alpha+s, \beta+f)$ (Doorn, Meijer, Frampton, Barclay, & Boer, 2020). Beta distributions that encode the relative prior plausibility of the values for θ_A and θ_B. In a Beta distribution, the α values can be interpreted as counts of hypothetical "prior successes" and the β values can be interpreted as counts of hypothetical "prior failures" (Lee & Wagenmakers, 2014):

$$\theta_A \sim Beta(\alpha_A, \beta_A)$$

$$\theta_B \sim Beta(\alpha_B, \beta_B)$$

Data from the A/B testing experiment update the two independent prior distributions to two independent posterior distributions as dictated by Bayes' rule:

$$p(\theta_A|y_A, n_A) = p(\theta_A) \times p(y_A, n_A|\theta_A)$$
$$p(y_A, n_A)$$

$$p(\theta_B|y_B, n_B) = p(\theta_B) \times p(y_B, n_B|\theta_B)$$
$$p(y_B, n_B)$$

where $p(\theta_A)$ and $p(\theta_B)$ are the prior distributions and $p(y_A, n_A|\theta_A)$ and $p(y_B, n_B|\theta_B)$ are the likelihoods of the data given the respective parameters.

Bayesian learning, reflecting the evolution of probability from prior to posterior is brought about by the data. Bayesian A/B models learn from the data, and probabilities increases for parameter values that predict the data well and decreases for parameter values that predict the data poorly (Kruschke, 2013; Doorn *et al.*, 2020 and Wagenmakers, Morey, & Lee, 2016). In practice we are interested in the difference $\delta = \theta_A - \theta_B$ between the success rates of the two experimental groups, as this difference indicates whether the experimental condition shows the desired effect (e.g. more sales).

5. Results

We constructed a regression model of sales using GA metrics. A sequence of searches resulted in a final model with R^2: 0.5803, Adjusted R^2: 0.4849 and p-value: 9.793e−06. Results are summarized in the table below. Table 3 reports the results of Baysian vs. frequentist A/B test statistics. Note that the two sets of statistics are not perfectly comparable, given the differences in structure of the tests. Specifically Bayesian tests provide a full reporting of posterior estimator characteristics, while frequentist statistics provide only p-factors in a Neyman-Pearson setting. Table 4 presents the key empirical findings of the analysis, reporting the probability that a change in a firm's website factor was caused by a change in sales due to competition from fraudulent sites.

The Granger causality test is a statistical hypothesis test for determining whether one TS is useful in forecasting another – a standard econometric definition of "causality". A TS X is said to Granger-cause Y if it can be shown through a series of t-tests and F-tests on lagged values of X that those X values provide statistically significant information about future values of Y.

One retains in this regression all lagged values of x that are individually significant according to their t-statistics, provided that collectively they add explanatory power to the regression according to an F-test (whose null hypothesis is no explanatory power jointly added by the x's). In my tests of the monthly data, the shortest lag was 1 and longest was 6. A lag of 6 months would test for the influence of a change in web traffic on the amount of sales six months later.

GA factor	*p−value* A/B	μ A/B	σ^2 A/B	μ-Low on CI	μ-High on CI	σ^2-Low on CI	σ^2-High on CI	μ for *post−E(loss)*	σ^2 for *post−E(loss)*
Sales	0.204	0.203	0.136	−13.296	9.389	−0.642	0.234	−1.424	0.643
New visitors	0.289	0.287	0.330	−7.008	5.365	−0.544	0.566	1.754	0.332v
Users	0.276	0.276	0.872	−8.426	6.351	−0.181	1.827	−0.895	0.030
Bounces	0.058	0.058	0.791	−3.674	0.168	−0.278	1.494	−1.163	0.056
Bounce rate	0.002	0.001	0.995	−3.185	−1.153	0.449	3.983	0.000	0.001
Average session duration	0.173	0.172	0.999	−8.356	4.466	0.975	5.847	−1.308	0.000
Session duration	0.183	0.180	1.000	−8.668	4.820	0.985	5.804	−2.311	0.000
Unique dimension combinations	0.043	0.042	0.788	−3.441	−0.153	−0.274	1.480	−3.903	0.056
Entrance rate	0.301	0.300	0.992	−10.351	8.365	0.351	3.650	−0.074	0.001
Time on page	0.182	0.181	0.995	−8.339	4.769	0.455	4.001	−2.471	0.001

Table 3. A/B testing Results from both Frequentist and Bayesian A/B Testing

	F-statistic	Prob(>*F*)	Prob(Change in this site statistic caused a change in sales)
new_vis	2.1968627	0.0659050	93.41%
users	1.2412285	0.3087493	69.13%
bounces	1.0421771	0.4147192	58.53%
bounceRate	2.2119510	0.0642794	93.57%
sessionDuration	1.5073274	0.2036900	79.63%
avgSessionDuration	1.5149517	0.2012254	79.88%
uniqueDimensionCombinations	0.5693141	0.7519153	24.81%
entranceRate	1.2733206	0.2939657	70.60%
timeOnPage	2.8037845	0.0242251	97.58%
avgTimeOnPage	1.2170068	0.3203199	67.97%

Table 4. Probability that a change in the the firm's Website statistics was caused by a change in sales due to competition from fraudulent sites

The null hypothesis that x does not Granger-cause y is accepted if and only if no lagged values of x are retained in the regression. This is tested using the Wald test to assess constraints on statistical parameters based on the weighted distance between the unrestricted estimate and its hypothesized value under the null hypothesis, where the weight is the precision of the estimate. The methods used in this study for Bayesian A/B testing including prior elicitation options were based on Kass and Vaidyanathan (1992).

In the Bayesian A/B analysis, (B-blue) represents data from the period in which fraudulent sites were active, and (A-orange) represents data from outside that period B−A==> reduction in traffic due to fraudulent sites; negative implies increased traffic credible interval on (A−B)/B for interval length(s) (0.9, 0.9).

Table 3 reports the findings of this research. Note that Bayesian A/B testing yields a plethora of measurements of the posterior distribution, while frequentist A/B tests, being Neyman-Pearson hypothesis tests, offer only the $p-value$. The $p-value$ for frequentist A/B tests is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that there is no difference between A and B (Hubbard & Lindsay, 2008; Wasserstein & Lazar, 2016). Unfortunately, a precise meaning of p-value is hard to grasp, and misuse is widespread and has been a major topic in metascience (Munafò, Nosek, Bishop, Button, & Chambers, 2017; Wasserstein & Lazar, 2016). While misuse of $p-values$ in scholarly articles may simply be grist for academic debate, the uncertainty surrounding the meaning of $p-values$ in business analytics actually can cost firms money. This is one of the conclusions that one may draw from Table 3.

In Table 3, GA factor refers to the GA metric tracked for the site over the period of the research dataset. The $p-value$ is the reuslt of frequentist A/B testing and is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that there is no difference between A transactions and B transactions. Distribution statistics (μ, σ^2) are given for posterior distributions, direct probabilities that A > B (by percent lift), credible intervals on (A−B)/B and the posterior expected loss. Credible intervals are the Bayesian counterpart to confidence intervals.

The following summarizes the main effects of fraudulent websites on the firm's web traffic discovered through Bayesian A/B testing.

New Visitors: Each new visitor to the firm site is worth $84.44, and the fraudulent site reduced the number of new visitors by 2472 per month. GA differentiates between new and returning users based on visitors' browser cookies.

Users: a user is a visitor who has initiated a session on the firm's website. These rose during the period the fraudulent sites were in existence. One explanation would be confusion spawned by the fraudulent websites, where potential customers looking to buy a roof were confronted with irrelevant but sensationalist information and were curious enough to visit the firm's site. Since the impact of more users appears to slightly decrease the potential sales dollars (perhaps from spurious site visits or searches), both of these could statistics may have been influenced by the existence of the fraudulent sites.

Customer Bounces: *bounces* and *bounceRate* increased during the period that the fraudulent sites were active, suggesting that there were some potential customers that left the firm's site immediately after visiting the landing page, costing firm advertising dollars without generating sales. Each additional bounce cost firm $33.99, and the fraudulent sites compelled additional customer bounces of 1581 per month.

Customer Time on Site: *sessionDuration* and *aveSessionDuration* number of minutes spent on the site during the period that the fraudulent sites were active fell, but these metrics had very small unit contributions to sales, and the estimates of these unit contributions were only barely statistically significant. The fraudulent sites clearly had a negative impact on time spent on the site due to the substantial increase in bounces, i.e. customers visited the landing page, and immediately left the site.

Customer Time on a Page: *timeOnPage* is the absolute time spent looking at the firm's pages during the existence of the fraudulent sites was reduced by 104,429 minutes per month during the period the fraudulent sites were in existence. *avgTimeOnPage*:_is the average time spent looking at the firm's pages during the existence of the fraudulent sites reduced by 57,039 minutes per month during the period the fraudulent sites were in existence. Average time differs from absolute time on the page during the period of the fraudulent site, because of the increase in bounces.

Customer Entrance on Pages other than the Landing Page: *entranceRate* reflects where GA records an entrance for each page that a user begins a new session on. The number of entrances given for a specific page shows how many users began their session with that page. This should not be confused with visiting the landing or entrance (top) Page. Review of the data suggests that landing inside the firm's site occurs mainly with organic searches, and probably reflects the Google indexing of the firm's site. These rose during the period the fraudulent sites were in existence, suggesting that visitors may have visited the site through a random, likely organic Google search rather than an advertising link or referral.

Unique Combinations of Customer Search: *uniqueDimensionsCombinations* or unique dimension combinations counts the number of unique dimension-value combinations for each dimension. For example, if you have the dimensions: a.Region, b.Language and c.Mobile Device Info, then GA counts the number of times it sees the same combination of dimension values for each row in the report. It appears that the existence of fraudulent sites had little effect on this metric, since it is relatively complex and reflects their shopping choices during and after their decision to purchase a metal roof. The unit contribution of this metric is large, but it represents the choices already made by committed firm customers.

Sales: Perhaps the most cogent statistic yielded by Bayesian A/B tests of the empirical data was that showing loss of sales during the existence of the fraudulent websites. Here we can use the posterior distribution to directly compute the expected loss from expected sales during a period. Our calculations based on the above analysis yielded:

(1) 2019-12-01 $544,549 (less than expected)

(2) 2020-10-01 $2,208,775 (more than expected)

(3) 2021-03-01 $3,003,636 (more than expected)

This not only highlighted the existence of fraud, but provided specific figures on the magnitude of that fraud. This analysis developed a model to show that during the 25.56 month fraudulent sites operating period between April 7, 2018 and May 13 2020, their activities reduced visits, reduced time on the firm's site and reduced conversions to the extent that firm lost an estimated $5,474,856 of sales revenue during the period. These direct effects were captured in the GA firm website metrics. I built a regression model and validated that these changes in GA site metrics directly caused the reduction in sales during the period of operation of fraudulent websites.

6. Conclusions and discussion

The regression model used to compute the firm's loss of sales due to changes in the GA metrics during the period the fraudulent sites were active, explained 60% of sales variability (with the rest attributable to word of mouth, repeat customers and so forth). The firm's website was the main driver of the firm's sales during the fraudulent period.

Direct analysis of the firm's sales TS estimated that they lost $5,474,856 of sales revenue during the 25.56 month fraudulent site period between April 7, 2018 and May 13 2020 when fraudsters posted and managed several websites with fabricated domain names and derogatory content.

Indirect analysis from predictions of the firm's sales revenue based on actual GA metrics, and using the regression model developed in this analysis, estimated a 95% prediction interval of [0, $7,787,300] – i.e. prediction limits validate the direct estimate developed in the analysis and show that the loss of $5,474,856 of sales revenue during the 25.56 month fraudulent site period was caused by fraudulent sites an URLs interfering with customer visits to the firm's website.

7. Potential applications of the research findings and managerial implications

The massive transaction volumes of e-commerce retailers – e.g. Amazon averages 1.6m transactions per day – automated fraud detection is of intense interest to e-commerce firms. There simply is no way that human auditors could effectively monitor that volume of sales. Automation, though, is predicated on efficient and effective algorithms. The current research has shown the particular appropriateness of Bayesian A/B testing for assessing the economic impact of fraud and identifying where to investigate fraud.

Though the mathematics and its application for A/B tests have been well understood, their application in fraud detection has to this point been almost nonexistent. This is partly due to most commercial applications of A/B testing using frequentist algorithms, because they are readily derived from Neyman-Pearson hypothesis testing. This has been unfortunate, as Bayesian A/B testing, unlike frequentist approaches not only allows ready computation of the economic consequences (the difference between posterior means) of fraudulent transactions, but also allows methods to precisely measure risk (i.e. the tail value at risk, or the integral of the tail of the posterior distribution over the appropriate loss function).

This research does move forward our understanding of how to manage A/B tests in a real e-commerce environment. Table 5 summarizes these in terms of the qualitative characteristics of Bayesian vs. frequentist statistical approaches. The posterior distribution of each of the GA factors provides a very conservative assessment of posterior expected loss and credible intervals (similar to confidence limits in fiducial inference). In frequentist statistics – the statistics that today are most often used in electronic market A/B testing – the alternative to the posterior distribution is the *p−value* – probabilities of obtaining test results at least as extreme as the result actually observed, under the assumption that there is no difference between A and B (Hubbard & Lindsay, 2008; Wasserstein & Lazar, 2016). Unfortunately, a precise meaning of *p−value* is hard to grasp and misuse is widespread and has been a major topic in metascience (Munafò *et al.*, 2017; Wasserstein & Lazar, 2016). While misuse of *p−values* in scholarly articles may simply be grist for academic debate, the uncertainty surrounding the meaning of *p−values* in business analytics actually can cost firms money. This is the main conclusions that one should draw from this research. Bayesian A/B tests of the data not only yielded a clear delineation of the timing and impact of the IP fraud, but calculated the loss of sales dollars, traffic and time on the firm's website, with precise confidence limits. Frequentist A/B testing identified fraud in bounce rate at 5% significance, and bounce at 10% significance, but was unable to ascertain fraud at the standard significance cutoffs for scientific studies. From a managerial standpoint, being able to only weakly conclude or reject the existence of fraud offered in frequentist *p−values* (particularly from a dataset of ~8m transactions) pales in comparison to the rich set of options for reporting loss and damage to reputation and traffic that is offered by Bayesian A/B testing.

The empirical research in this paper provides a guide for what one might expect as "typical" values for the empirical parameters of the Bayesian vs Frequentist approaches. These are given in Table 6.

Within this particular Bayesian A/B detective control, we also need to consider that there may be improvements to be made in confidence and prediction intervals that measure the

Objective of fraud audit	Bayesian	Frequentist
Controls		
Preventive controls over fraud (passive)	No	No
Detective control (group or individual transactions)	Yes	Yes
Corrective control allowing lost cost, accurate recovery from fraud	No	No
Economics		
Able to identify transaction sets that are fraudulent	Yes	Yes
Objective is maximizing net savings from fraud	Yes	No, *p−values* only estimate the probability that our decision is correct for a group of transactions being fraudulent
Able to apply firm's actual loss function	Yes	No, *p−values* only estimate the probability that our decision is correct for a group of transactions being fraudulent
Able to calculate loss under competing decisions	Yes	No, *p−values* only estimate the probability that our decision is correct for a group of transactions being fraudulent
Able to compute fraud cost	Yes	No, *p−values* only estimate the probability that our decision is correct for a group of transactions being fraudulent
Able to calculate loss under competing decisions	Yes	No, *p−values* only estimate the probability that our decision is correct for a group of transactions being fraudulent
Operations		
Generalist algorithm	Yes	No, requires a hypothesis testing framework
Empirical fraud detection methodology	Yes	Yes
Supports labeling of individual transactions as potentially fraudulent	Yes	Yes, with limitations
Can be scaled up for large transaction volumes	Yes	Yes, though *p−values* for a decision are less and less reliable as the application is scaled up to larger transaction numbers
Simple and low cost to implement	Yes	No, requires a hypothesis testing framework
Highly efficient, low cost transaction processing	Yes	Yes
Many tools available for implementation	No	Yes
Comparison with competitive methods		
Autoencoders	Competitive	Not Competitive
Benford tests	Competitive	Not Competitive
Sarbanes-Oxley tests	Competitive	Competitive for Section 302 tests, but not for Section 404 tests
Supervised Rule-based methods	Competitive	Competitive
Supervised Tree-based algorithms	Competitive	Competitive
Supervised Methods with misclassification	Competitive	Competitive
Unsupervised classification methods	Not competitive, requires labeling	Not competitive, requires labeling

Table 5. Comparative advantages of Bayesian vs Frequentist approaches for fraud auditing

p−value A/B	μ A/B	σ^2 A/B	μ-Low on CI	μ-High on CI	σ^2-Low on CI	σ^2-High on CI	μ for *post−E*(*loss*)	σ^2 for *post−E*(*loss*)
0.1711	0.17	0.7898	−7.4744	4.2387	0.1296	2.8886	−1.1795	0.112

Table 6. Comparative average statistics for the research dataset

accuracy of models in estimating a mean, or predicting a new value, respectively. Intervals allow one to estimate a range of values that can be said with reasonable confidence (typically 95%) contains the true population parameter. As this analysis is focused on the regression model that is used to predict firm sales from GA metrics, I am interested here in computing the prediction interval. I have used the industry standard 95% prediction intervals – i.e. 19 out of 20 times, our answer will be correct, a value that was set as a scientific best-practice early in the 20th century by the statistician Ronald Fisher (Fisher, 1932), the implications of which are studied at length in (Stigler, 2008). Sales clearly cannot be less than zero, and thus lower limits were truncated to zero. In addition to the arbitrary selection of significance equal to 0.05 due to Fisher (which was initially only a suggestion), more recently (Cohen, 2016) has suggested that power of tests should be set at 0.80. In practice, *TypeI* and *TypeII* choices should be dictated by the loss function, but these choices of cut-offs for decision making are too often applied without any thought to loss (or perhaps to avoid decisions about the loss function).

8. Limitations and future research

The research presented in this paper is only one component of a complete fraud management operation in e-commerce. The research provides a method of automated monitoring of groups of transactions to identify ones with a high probability of being fraudulent. They are detective controls that require a detailed follow-up investigation, and implementation of corrective controls to repair the damage done by the fraud. As such, the algorithms developed and tested in this paper should not be seen as a complete solution; rather they are an exceptionally efficient and informative part of automated detective controls.

The regression model used to compute the firm's loss of sales due to changes in the GA metrics during the period the fraudulent sites were active, explained 60% of sales variability (with the rest attributable to word of mouth, repeat customers and so forth). The firm's website was the main driver of the firm's sales during the fraudulent period. But one must also recognize that 40% of sales variability was not explained and that is a limitation that can be improved on in future research.

References

Ahfock, D., Pyne, S., & McLachlan, G.J. (2022). Statistical file-matching of non-Gaussian data: A game theoretic approach. *Computational Statistics and Data Analysis*, *168*, 107387.

Al-Shabi, M. (2019). Credit card fraud detection using autoencoder model in unbalanced datasets. *Journal of Advances in Mathematics and Computer Science*, *33*, 1–16.

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal*, *311*, 485.

Ashbaugh-Skaife, H., Collins, D. W., Kinney, W. R. Jr, & LaFond, R. (2008). The effect of SOX internal control deficiencies and their remediation on accrual quality. *The Accounting Review*, *83*, 217–250.

Bedard, J. C., & Graham, L. (2011). Detection and severity classifications of Sarbanes-Oxley section 404 internal control deficiencies. *The Accounting Review*, *86*, 825–855.

Bedard, J. C., Hoitash, R., & Hoitash, U. (2009). Evidence from the United States on the effect of auditor involvement in assessing internal control over financial reporting. *International Journal of Auditing*, *13*, 105–125.

Berger, A., & Hill, T. P. (2011). A basic theory of Benford's law. *Probability Surveys*, *8*, 1–126.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle*. IMS.

Berger, P. G., Li, F., & Wong, M. F. (2005). The impact of sarbanes-oxley on cross-listed companies. Unpublished Paper.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, *17*, 235–255.

Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit scoring and credit control VII* (pp. 235–255).

Box, J. F. (1987). Guinness, gosset, Fisher, and small samples. *Statistical Science*, 45–52.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Cart. classification and regression trees*, Routledge.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... & Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*, 637–644.

Carta, S., Fenu, G., Recupero, D. R., & Saia, R. (2019). Fraud detection for e-commerce transactions by employing a prudential multiple consensus model. *Journal of Information Security and Applications*, *46*, 13–22.

Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems and Their Applications*, *14*, 67–74.

Cheng, D., Wang, X., Zhang, Y., & Zhang, L. (2020). Graph neural network for fraud detection via spatial-temporal attention. *IEEE Transactions on Knowledge and Data Engineering*.

Chhikara, R. S., & McKeon, J. (1984). Linear discriminant analysis with misallocation in training samples. *Journal of the American Statistical Association*, *79*, 899–906.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, *3*, 261–283.

Cohen, W. W. (1995). Fast effective rule induction. *Machine learning proceedings 1995* (pp. 115–123). Elsevier.

Cohen, J. (2016). A power primer. *Psychological Bulletin*, 1992 (Jul), *112*(1).

Cortes, C., Pregibon, D., & Volinsky, C. (2001). Communities of interest. *International symposium on intelligent data analysis* (pp. 105–114). Springer.

Dijesh, P., Babu, S., & Vijayalakshmi, Y. (2020). Enhancement of e-commerce security through asymmetric key algorithm. *Computer Communications*, *153*, 125–134.

Doorn, A. S. V., Meijer, B., Frampton, C. M., Barclay, M. L., & Boer, N. K. D. (2020). Systematic review with meta-analysis: SARS-CoV-2 stool testing and the potential for faecal-oral transmission. *Alimentary Pharmacology and Therapeutics*, *52*, 1276–1288.

Doorn, J. V., Bergh, D. V. D., Böhm, U., Dablander, F., Derks, K., Draws, T., ... & others (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin and Review*, *28*, 813–826.

Dzulfikar, M. F., Sensuse, D. I., & Noprisson, H. (2017). A systematic literature review of information system adoption model applied to enterprise 2.0. *2017 international conference on information technology systems and innovation (ICITSI)* (pp. 14–19). IEEE.

Feng, M., Li, C., & McVay, S. (2009). Internal control and management guidance. *Journal of Accounting and Economics*, *48*, 190–209.

Feng, M., Li, C., McVay, S. E., & Skaife, H. (2014). Does ineffective internal control over financial reporting affect a firm's operations? Evidence from firms' inventory management. *The Accounting Review*, *90*, 529–557.

Fisher, R. (1932). *Statistical methods for research workers*. Edinburgh: Oliver and boyd. 1925. Google Scholar.

Ge, W., Koester, A., & McVay, S. (2016). The costs and benefits of section 404 (b) exemption: Evidence from small firms' internal control disclosures. Available from: SSRN.

Goodson, M. (2014). Most winning a/b test results are illusory, Whitepaper. Available from: https://tinyurl.com/y9g3m9bq

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.

Gronau, Q. F., Raj, K., & Wagenmakers, E. -J. (2019). Informed bayesian inference for the a/b test. *arXiv preprint arXiv:1905.02068*.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*, 523–541.

Hill, T. P. (1995). Base-invariance implies Benford's law. *Proceedings of the American mathematical society* (Vol. 123, pp. 887–895).

Hoitash, U., Hoitash, R., & Bedard, J. C. (2009). Corporate governance and internal control over financial reporting: A comparison of regulatory regimes. *The Accounting Review*, *84*, 839–867.

Hooi, B., Song, H. A., Beutel, A., Shah, N., Shin, K., & Faloutsos, C. (2016). Fraudar: Bounding graph fraud in the face of camouflage. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 895–904).

Hopkins, C. C. (1923). *Scientific advertising*. Lincolnwood, IL: NTC. (First published 1923). Google Scholar.

Hubbard, R., & Lindsay, R. M. (2008). Why *p* values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology*, *18*, 69–88.

Ismanto, L., Suwito Ar, H., Fajar, A. N., & Bachtiar, S. (2019). Blockchain as E-commerce platform in Indonesia. *Journal of Physics: Conference Series*, IOP Publishing, *1179*(1): 012114.

Jennison, C., & Turnbull, B. W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science*, *5*, 299–317.

Jha, S., & Westland, J. C. (2013). A descriptive study of credit card fraud pattern. *Global Business Review*, *14*, 373–384.

Jha, S., Guillen, M., & Westland, J. C. (2012). Employing transaction aggregation strategy to detect credit card fraud. *Expert Systems with Applications*, *39*, 12650–12657.

Johnson, G. A., Lewis, R. A., & Nubbemeyer, E. I. (2017). Ghost ads: Improving the economics of measuring online ad effectiveness. *Journal of Marketing Research*, *54*, 867–884.

Kamalbasha, S., & Eugster, M. J. (2021). Bayesian a/b testing for business decisions. *Data science–analytics and applications* (pp. 50–57). Springer.

Kass, R. E., & Vaidyanathan, S. K. (1992). Approximate bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*, 129–144.

Keysers, C., Gazzola, V., & Wagenmakers, E. -J. (2020). Using bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*, 788–799.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*, 573.

Laurens, R., Rezaeighaleh, H., Zou, C. C., & Jusak, J. (2019). Using disposable domain names to detect online card transaction fraud. *ICC 2019-2019 IEEE international conference on communications (ICC)* (pp. 1–7). IEEE.

Lee, M. D., & Wagenmakers, E. -J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Li, Z., Song, J., Hu, S., Ruan, S., Zhang, L., Hu, Z., & Gao, J. (2019). Fair: Fraud aware impression regulation system in large-scale real-time e-commerce search platform. In *2019 IEEE 35th international conference on data engineering (ICDE)* (pp. 1898–1903). IEEE.

Lin, S., Pizzini, M., Vargus, M., & Bardhan, I. R. (2011). The role of the internal audit function in the disclosure of material weaknesses. *The Accounting Review*, *86*, 287–323.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*, 22–25.

Little, B. R. (1989). Personal projects analysis: Trivial pursuits, magnificent obsessions, and the search for coherence. In *Personality Psychology* (pp. 15–31). Springer.

Luhach, A. K., Dwivedi, S. K., & Jha, C. K. (2014). Applying SOA to an e-commerce system and designing a logical security framework for small and medium sized e-commerce based on SOA. *2014 IEEE international conference on computational intelligence and computing research* (pp. 1–6). IEEE.

Mahto, D., & Yadav, D. K. (2015). Enhancing security of one-time password using elliptic curve cryptography with biometrics for e-commerce applications. *Proceedings of the 2015 third international conference on computer, communication, control and information technology (C3it)* (pp. 1–6). IEEE.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., & Ioannidis, J.P.A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.

Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, *133*, 113303.

Pumsirirat, A., & Liu, Y. (2018). Credit card fraud detection using deep learning based on auto encoder and restricted Boltzmann machine. *International Journal of Advanced Computer Science and Applications*, *9*(1).

Qiu, L., & Li, J. (2017). Covering the monitoring network: A unified framework to protect e-commerce security. *Complexity*, *2017*.

Quinlan, J. R. (1990). Decision trees and decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*, *20*, 339–346.

Rice, S. C., Weber, D. P., & Wu, B. (2014). Does SOX 404 have teeth? Consequences of the failure to report existing internal control weaknesses. *The Accounting Review*, *90*, 1169–1200.

Ripley, B. D., & Ripley, R. M. (2001). Neural networks as statistical methods in survival analysis. *Clinical Applications of Artificial Neural Networks*, *237*, 255.

Robinson, G. K. (2018). What properties might statistical inferences reasonably be expected to have?—crisis and resolution in statistical inference. *The American Statistician*.

Savita, R., & Datta, U. (2015). Two way authentication in MITM attack to enhance security of e-commerce transactions. *International Journal of Security and Its Applications*, *9*, 265–274.

Schorman, R. (2008). Claude hopkins, earnest calkins, bissell carpet sweepers and the birth of modern Advertising1. *The Journal of the Gilded Age and Progressive Era*, *7*, 181–219.

Senator, T. E. (2000). Ongoing management and application of discovered knowledge in a large regulatory organization: A case study of the use and impact of NASD regulation's advanced detection system (RADS). *Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 44–53).

Sharma, V. K., Mathur, P., & Srivastava, D. K. (2018). Secure electronic fund transfer model based on two level authentication. *2018 second international conference on electronics, communication and aerospace technology (ICECA)* (pp. 1338–1342). IEEE.

Stigler, S. (2008). Fisher and the 5% level. *Chance*, *21*, 12.

Stolberg, M. (2006). Inventing the randomized double-blind trial: The nuremberg salt test of 1835. *Journal of the Royal Society of Medicine*, *99*, 642–643.

Suryono, R. R., Purwandari, B., & Budi, I. (2019). Peer to peer (P2P) lending problems and potential solutions: A systematic literature review. *Procedia Computer Science*, *161*, 204–214.

Wagenmakers, E. -J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin and Review*, *14*, 779–804.

Wagenmakers, E. -J., Beek, T. F., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., . . . & Pinto, Y. (2015). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology*, *6*, 494.

Wagenmakers, E. -J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

Wagenmakers, E. -J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … & others (2018a). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76.

Wagenmakers, E. -J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., … & others (2018b). Bayesian inference for psychology. Part i: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*, *25*, 35–57.

Wasserman, S., & Faust, K. (1994). *Advances in social network analysis: Research in the social and behavioral sciences*. Sage.

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

Webb, E.J., Campbell, D.T., Schwartz, R.D., & Sechrest, L. (1999). *Unobtrusive measures*. Sage Publications, *2*.

Westland, J. C. (2000). Research report: Modeling the incidence of postrelease errors in software. *Information Systems Research*, *11*, 320–324.

Westland, J. C. (2002). The cost of errors in software development: Evidence from industry. *Journal of Systems and Software*, *62*, 1–9.

Westland, J. C. (2004). The cost behavior of software defects. *Decision Support Systems*, *37*, 229–238.

Westland (2017). An empirical investigation of analytical procedures using mixture distributions. *Intelligent Systems in Accounting, Finance and Management*, *24*, 111–124.

Westland, J. C. (2019). The information content of Sarbanes-Oxley in predicting security breaches. *Computers and Security*, *90*, 101687, doi: 10.1016/j.cose.2019.101687.

Westland (2020a). *Audit analytics: Data science for the accounting profession*. Springer Nature.

Westland, J. C. (2020b). Predicting credit card fraud with Sarbanes-Oxley assessments and Fama-French risk factors. *Intelligent Systems in Accounting, Finance and Management*, *27*(2): 95–107.

Xie, W., Zhou, W., Kong, L., Zhang, X., Min, X., Xiao, Z., & Li, Q. (2018). Ettf: A trusted trading framework using blockchain in e-commerce. *2018 IEEE 22nd international conference on computer supported cooperative work in design (CSCWD)* (pp. 612–617). IEEE.

Xu, M., & Chu, Y. (2017). A intelligent risk detection method in online transactions. *2017 12th international conference on intelligent systems and knowledge engineering (ISKE)* (pp. 1–3). IEEE.

Zhang, G., Li, Z., Huang, J., Wu, J., Zhou, C., Yang, J., & Gao, J. (2022). Efraudcom: An e-commerce fraud detection system via competitive graph neural networks. *ACM Transactions on Information Systems (TOIS)*, *40*, 1–29.

Zheng, L., Liu, G., Luan, W., Li, Z., Zhang, Y., Yan, C., & Jiang, C. (2018). A new credit card fraud detecting method based on behavior certificate. *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)* (pp. 1–6). IEEE.

About the author

James Christopher Westland. I am currently Professor in the Department of Information and Decision Sciences at the University of Illinois – Chicago. I have a BA in Statistics and an MBA in Accounting from Indiana University and received my PhD in Computers and Information Systems from the University of Michigan. I have professional experience in the US as a certified public accountant and as a consultant in technology law in the US, Europe, Latin America and Asia. I am the author of numerous academic papers and of seven books: Global Electronic Commerce (MIT Press 2000); Global Innovation Management (Palgrave Macmillan 2nd ed 2017); Red Wired: China's Internet Revolution (Marshall Cavendish, 2010); Structural Equation Modeling (Springer 2015); Financial Dynamics (Wiley 2003); Valuing Technology (Wiley 2002) and Audit Analytics: Data Science for the Accounting Profession (in R. Gentleman's "Use R!"

series @ Springer). I am the Editor-in-Chief of Electronic Commerce Research (Springer) and have served on editorial boards of several other information technology journals including Management Science, ISR, ECRA, IJEC and others. I have served on the faculties at the University of Michigan, University of Southern California, Hong Kong University of Science and Technology, Tsinghua University, University of Science and Technology of China, Harbin Institute of Technology and other academic institutions. In 2012 I received High-Level Foreign Expert status in China under the 1000-Talents Plan and am currently Overseas Chair Professor at Beihang University. I have advised on patent, valuation and technology strategy for numerous technology firms. James Christopher Westland can be contacted at: westland@uic.edu