THE EFFECT OF SELECTIVE DATA OMISSION ON TYPE I ERROR RATES: A SIMULATION STUDY

A Thesis

by

MARK JEFFERY HARDCASTLE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	David Bessler
Co-chair of Committee,	Henry Bryant
Committee Member,	Brent Donnellan
Head of Department,	Parr Rosson

December 2015

Major Subject: Agricultural Economics

ABSTRACT

There do not exist widely accepted guidelines or standards for identification and removal of outlying data in empirical research. There are sometimes significant incentives for researchers to discover particular research results. Researchers have been observed to use flexibility in outlier omission to selectively omit data in search of statistically significant findings. The degree to which this practice can affect the credibility of research findings is unknown. This study uses Monte Carlo simulation to estimate the propensity of certain types of selective outlier omission to inflate type I error rates in regression models.

Simulations are designed to analyze posttest only control group design with no underlying intervention effect, such that any statistically significant findings represent type I errors. Omission of observations is simulated in an exploratory manner, such that observations are omitted and regressions are run iteratively until either a type I error is made or until a maximum trimming threshold is reached, whichever occurs first. Omission of observations based on z-score thresholds, a common research practice in some disciplines, is simulated. Additionally, omission from only of one tail of data—simulating the removal of only "disconfirming" observations—is analyzed. Simulations are performed using a variety of sample sizes and with samples drawn from several underlying population distributions. In all simulations, type I error rates are inflated; type I error rates are found to range from 7.86% to 100%, compared to the expected 5% in the absence of data omission.

ACKNOWLEDGEMENTS

I would like to thank my Chair and Co-chair, Dr. Bessler and Dr. Bryant, and my committee member, Dr. Donnellan, for their thorough and indispensable guidance. I would also like to thank Dr. Joseph Simmons of the Wharton School of the University of Pennsylvania, who graciously provided me with the code that he used to perform similar simulations to those conducted in this thesis.

TABLE OF CONTENTS

Page

1. INTRODUCTION

In their analysis of the economics literature, Ioannidis and Doucouliagos (2013) concluded that "the credibility of the economics literature is likely to be modest or even low." This finding was based on a variety of factors, including a lack of standardization in and resultant flexibility of research design which contributes to the creation of false research claims (Ioannidis and Doucouliagos 2013). Leamer (1983) famously demonstrated that simple changes to empirical economic model specifications could very easily be used to reverse the sign and alter the magnitude of the observed relationship between two variables. If a researcher has an incentive to produce specific research results and design flexibility that allows for results to be modified, he or she might be expected to use the flexibility to get the desired results, regardless of the accuracy of the resultant research conclusions (Ioannidis 2005).

The types of retrospective observational analyses that allow for the flexibility that Leamer (1983) examined compose a smaller portion of the economics literature than they did when he authored the critique: Hamermesh (2013) found that, in three top journals, the proportion of experimental studies increased tenfold between 1983 and 2011, from .8% to 8.2%, and the prevalence of quasi-experimental studies has also increased (Angrist and Pishke 2010). This is nearly certain to increase the credibility of economics research: Angrist and Pishke (2010) refer to a "credibility revolution" in the field of economics spurred by this greater use of experimental and quasi-experimental research design. But even with studies designed as "gold standard" randomized controlled trials, a large degree of flexibility in research design still exists, and this

flexibility could be exploited by researchers who are willing to change their design specifications in search of particular results (Simmons et al. 2011). Simmons et al. (2011) demonstrated that, under certain assumptions of flexibility of research design (which the authors suggest might be conservative), a researcher conducting an experiment could erroneously discover a non-existent relationship between two variables in over half of simulations.

Researchers often have a variety of incentives that drive them toward a preference for a particular research outcome: financial conflicts of interest, pre-conceived hopes for or expectations of outcomes, or a belief that a certain result may be more likely to be published than another are all factors that can influence a researcher to do what he or she can to find a particular result (Krimsky and Rothberg 1998; Stanley et al. 2008; Statzner and Resh 2010). Several methods exist to exploit flexibility in research design and reporting in order to achieve the outcome that a researcher might desire; collectively, these methods are sometimes referred to as "questionable research practices," or QRPs (Simmons et al. 2011).

The objective of this study is to analyze, in depth, the consequences of exploiting one of these flexible research design components. Researchers are often not constrained in determining when and how to omit apparently outlying data observations, which could be used selectively to influence research findings (Bakker and Wicherts 2014b). A review of statistics textbooks performed by Bakker and Wicherts (2014b) found that, in the majority of texts, exclusion of any outliers was only recommended in cases where data recording errors were made. They additionally found that a number of texts

described seemingly outlying but non-erroneous observations as important elements from the population being studied that should not be excluded (for example, Freedman et al. 2007, as cited in Bakker and Wicherts 2014b). Outlier omission, however, is often not performed only when errors are present in the data: researchers often simply use exclusion rules that eliminate all observations beyond a threshold percentile, value, or z-score (Bakker and Wicherts 2014b; Simmons et al. 2011). The effect of these sorts of exclusion rules were examined in simulations run by Bakker and Wicherts (2014b), who found that they can dramatically increase type I error rates.

This study uses Monte Carlo simulations to examine a model that assumes that a researcher is attempting to discover a specific result through *selective* data omission, estimating the impact that such omission will have on type I error rates under a variety of assumptions. An underlying assumption is not just that data will be omitted, but that omission will take place *only* when it helps to cross a $p = .05$ significance threshold; data will systematically be excluded at different thresholds until a maximum threshold value is met or a statistically significant coefficient is observed, whichever happens first. This is designed to simulate the potential impacts of the flexibility that researcher have in deciding how to exclude outliers when this flexibility is exploited. The simulations run are described in detail in section 3.

This thesis additionally analyzes the context in which selective data omission might take place: first, the incentives that researchers have to commit QRPs are explored. Whether and to what degree researchers face disincentives to publishing false findings is additionally examined. Next, the existing evidence for the degree to which

researchers have flexibility to influence their research findings is analyzed. The empirical evidence for the prevalence of QRPs is then investigated. A review of the consequences of QRP use follows. The design of and results from the simulation performed for this study are then described, followed by a discussion of the results and the methods that have been proposed to limit the prevalence and impact of QRPs, including selective outlier omission. Finally, the limitations of the study and its place in the broader QRP-related literature are discussed.

2. LITERATURE REVIEW

2.1 Incentives faced by researchers

Researchers often have strong incentives to arrive at particular research conclusions: direct financial incentives of researchers or their funders can cause them to actively seek to achieve a certain result (Krimsky and Rothenberg 1998). Financial conflicts of interest have been found to be common and underreported in a variety of disciplines, including economics (Papanikolaou et al. 2001; Carrick-Hagenbarth and Epstein 2012).

Researchers often also have pre-conceived expectations that cause them to prefer particular research findings; the socio-economic characteristics of researchers have been found to explain a large portion of the variation in research findings in empirical economics (Stanley et al. 2008). This "confirmation bias" could cause researchers to selectively ignore interpretations of the data that run counter to their ideological or other interests (Rosenthal 1966, as cited in Camfield, et al. 2014; Nickerson 1998).

Even in the absence of direct financial or personal reasons to prefer a particular result, there exist strong reputational incentives to achieve results that will be published and widely cited (Statzner and Resh 2010). In the field of economics, the St. Louis Federal Reserve's Research Papers in Economics (RePEc) rankings of researchers and their output are perhaps the most widely used. These rankings largely depend on the number of papers published and the number of citations that they receive (Zimmerman 2012). Additionally, the number of publications is a common criterion for hiring and promotion decisions (Statzner and Resh 2010). The salary of academic economists has

been found in two separate studies to correlate significantly with their number of publications in top journals (O'Keefe and Wang 2014) and with the quantity of published papers in all journals (Hamermesh and Pfann 2012). These strong incentives to publish, particularly in top journals, are likely to influence how economists conduct research: some results are nearly certainly more likely to be published and widely cited than others, giving researchers an incentive to produce those preferred results.

There also exists pressure to create only results that are found to be statistically significant using null hypothesis statistical testing (NHST; Bakker et al. 2012). Publication is often so dependent on achieving statistically significant positive results that Bakker et al. (2012) described science as being treated by researchers as a "game," with the goal of producing statistically significant results. Among a large assortment of disciplines, NHST has become the dominant way for researchers to assess and communicate their findings (Gigerenzer 2004). Multiple analyses have shown that the predominance of NHST has increased over time and among several disciplines, including economics (Ziliak and McCloskey 2004; Fiddler et al. 2006). It has been demonstrated repeatedly that positive results are more likely to be published than negative results, and that negative results which are accepted by journals take significantly longer to be published (Stern and Simes 1997). Furthermore, the proportion of studies with positive results has been *increasing* over time, with the frequency of positive results growing by 22% between 1990 and 2007 (Fanelli 2011). The growth in proportion of positive results was observed across a variety of disciplines, including economics (Fanelli 2011). These findings point to the existence of strong incentives for

researchers to produce statistically significant findings using NHST, perhaps regardless of the accuracy of the research claims being made (Head et al. 2015).

The preceding paragraphs describe evidence about the incentives that researchers have to produce particular findings. Also important are the disincentives to producing inaccurate or misleading research claims. Certainly, outright research fraud that is discovered can carry large consequences (Lacetera and Zirulia 2011). But false results made through honest mistakes or through certain dishonest research practices are difficult to detect in economics. McCullough et al. (2008) examined publishing requirements in economics and determined that "most economics journals provide no mechanism whereby false results can be discovered". The authors discovered additionally that, even in journals that require submitters to publish the data on which their research conclusions are based, those who failed to meet the requirements did not suffer any penalties (McCullough et al. 2008). When data are available, they are very rarely requested; Hamermesh (2007) examined patterns of data requests for papers published in two journals, Industrial and Labor Relations Review and Journal of Human Resources, finding that data requests were never made for over 60% of papers. Even cases where mistakes are discovered to have been made and an author's work is retracted can carry relatively light consequences: on average, a researcher who suffers the retraction of one of his or her articles will receive 12.5% fewer citations on his or her subsequent work, and citations on the retracted work itself are still likely to be made even five years post-retraction (Lu et al. 2013).

2.2 Opportunities for researchers to influence research conclusions

The desire to achieve particular research findings would be of little importance if researchers did not have the ability to influence these findings. However, there exists a large variety of QRPs that allow researchers the flexibility to change their models and data in ways that can assist them to achieve their desired results (Simmons et al. 2011). Among other practices, these can include changing the statistical specifications being used, flexibly deciding when it is appropriate to stop data collection, deciding whether to attempt to get a result published, and considering whether and which data should be excluded, all based on whether the results of any of these manipulations would help the reported results to conform to the researcher's desired outcomes.

As mentioned, Leamer (1983) demonstrated that changes in model specifications could be used to dramatically influence research findings, to the point of reversing research conclusions. Simmons et al. (2011) simulated a case where a researcher could choose between two different correlated dependent variables, finding that simply choosing the specification that results in the highest p-value resulted in type I errors 9.5% of the time at the 5% significance level. This simulation is likely to represent a conservative estimate of the type I error chance that would result were this method employed in a real-world setting: a researcher might commonly have more than two dependent variables and a host of independent variables to select from (Simmons et al. 2011; Leamer 1983).

Flexible stopping rules for data collection are another potentially exploited QRP: researchers have the ability to run statistical tests while a study is ongoing and, if a significant result is observed, to cease data collection and report the result (Simmons et al. 2011). Similarly, if a researcher collects data and a statistically significant result is *not* observed, he or she would sometimes have the flexibility to simply add more observations (Simmons et al. 2011). A researcher could add one more observation at a time and, by adding an additional observation a sufficient number of times, never fail to be successful in observing a statistically significant result whether or not there exists an actual relationship in the population being studied (Wagenmakers 2007).

Deciding whether and how to omit outliers after data collection could also be used selectively to influence research findings (Simmons et al. 2011). There are no clear guidelines for identifying outlying data that it is appropriate to exclude, and a researcher could choose among and justify any number of outlier exclusion rules (Simmons et al. 2011; Bakker and Wicherts 2014b). This flexibility could allow a researcher to exclude observations in the way that makes the data most likely to appear to provide evidence for his or her preferred research conclusions (Simmons et al. 2011).

2.3 Evidence for the prevalence of QRPs

Empirically, it has been observed that the intersection of strong incentives to achieve positive, statistically significant findings and the considerable researcher degrees of freedom that could be exploited leads to a large portion of researchers engaging in QRPs in pursuit of positive results (John et al. 2012). Much of the evidence that suggests the prevalence of QRPs broadly does not provide an indication of which QRPs are causing the results. Because of this, the degree to which observed evidence for QRPs is influenced by selective outlier omission often cannot be determined. This section

therefore discusses the evidence for the degree that all QRPs take place, but focuses on evidence for the prevalence of selective outlier omission when it is available.

Direct surveys of researchers are a source of considerable evidence for the propensity of researchers to engage in questionable research practices. John et al. (2012) conducted a survey of psychology researchers, finding that more than half reported using at least one questionable research practice. Large proportions of respondents admitted to making a decision on when to stop data collection based on whether their results were significant, selectively reporting studies and results with positive findings (John et al. 2012). 38% admitted to having decided whether to exclude outliers based on the effects of doing so (John et al. 2012). LeBel et al. (2013) asked the authors of psychology publications to disclose otherwise undisclosed methods that were used in their reported research: of the 46.4% of contacted researchers who elected to reply non-anonymously, almost all admitted to failing to disclose some relevant research practices and 11.2% failed to report excluded observations. Bailey et al. (2001) surveyed accounting researchers and estimated from respondents that almost 4% of the top publications in the field were "seriously tainted" by "intentional violations that would affect the truthfulness of research reports." The survey respondents also reported the suspected prevalence of violations by others in the field, believing on average that upwards of 20% of the research was affected by intentional and serious QRPs (Bailey et al. 2001).

Fanelli (2009) conducted a meta-analysis of surveys of academics and found that, across disciplines, an average of approximately 2% of researchers admitted to having fabricated or manipulated their data at least once, "a serious form of misconduct by any

standard." Survey respondents suspected that about 14% of their colleagues had performed these data manipulations (Fanelli 2009). Up to 72% of researchers admitted to these and other QRPs (Fanelli 2009).

Replicability of studies is often used as an imperfect proxy for researcher honesty: while failing to replicate a study is not proof of deliberate questionable research practices (honest mistakes, statistical flukes, and lack of external validity of internally valid findings are possible explanations), consistent ability to replicate studies would certainly provide evidence that research practices are generally honest (Hamermesh 2007). Replication falls into three primary categories: *pure replication* uses the same data and analysis that were used in the study being replicated; *statistical replication* uses the same model, population, and specifications of a study but a different sample; *scientific replication* uses a non-identical model to analyze the same *idea* expressed in a paper, but using different populations or research specifications (Hamermesh 2007).

By the metric of replicability, a large portion of the economics literature that has been analyzed has fared poorly: Dewald et al. (1986) famously attempted to perform pure replication attempts on a number of studies published in the Journal of Money, Credit, and Banking. They found that nearly two thirds of authors contacted would not or could not provide the data that they had used to reach their research conclusions, in spite of a publication requirement to make their data available upon request (Dewald et al. 1986). The authors were able to access the data that were used in some of the studies, and a majority of these were not able to be replicated, although most failures to replicate were due to coding errors rather than deliberate QRPs (Dewald et al. 1986). After this

and similar controversies, the Journal of Money, Credit, and Banking and many other economics journals eventually instituted mandatory data and code submission (McCullough, 2007). Even with this more stringent data sharing requirement and two decades after the replication attempt by Dewald et al. (1986), the vast majority of authors submitting empirical papers to the journal still did not submit the required data and code, making pure replication of the majority of papers submitted impossible to attempt (McCullough 2007). A similar analysis of studies published in the Federal Reserve Bank of St. Louis *Review* found that fewer than 10% of studies analyzed were able to be replicated (McCullough et al. 2008). Replication failures extend far beyond the field of economics: in their famous review of large-scale replication attempts of cancer drug research findings, Begley and Ellis (2012) found that replication succeeded in only 11% of analyzed studies.

Several meta-analyses have been conducted in search of results that would suggest the exploitation of QRPs in a variety of ways. As mentioned, Stanley (2008) conducted a meta-analysis of research in economics found that research outcomes depend partly on socio-economic characteristics of the researchers, a result that is inconsistent with perfectly objective research. Rosenthal (1978) found that, during data collection, roughly two thirds of observation errors that are made are made in a way that favors the hypothesis of the person making the error. There is evidence that outliers are excluded from analysis more often when doing so helps to confirm a researcher's theory (Rosenthal 1994). Bakker and Wicherts (2014a) analyzed studies where data was available and outlier exclusion was reported and found no evidence that those studies

involved selective outlier omission, but discovered a large number of studies where data exclusion took place but was not reported. Of psychological papers that use NHST, 96% report positive findings, while reviews of the literature demonstrate that the studies do not have sufficient statistical power to so often achieve positive results, although this could be consistent with publication bias rather than with other QRPs (Bakker et al. 2012). Hróbjartsson et al. (2012) compared blinded to non-blinded trials and found that both statistical significance and effect sizes were inflated for the non-blinded trials; this suggests that researchers have a propensity to influence the results of their experiments when they have the ability to do so, whether intentionally or unintentionally. Chan et al. (2004) analyzed medical studies and found that over half of those that were examined had failed to report all of their findings, and that the results that were not reported usually were those that would have made the benefits of the intervention being studied appear to be smaller or the negative effects of the intervention appear to be larger than the reported results would suggest. There is similar evidence that economists with financial conflicts of interest are more likely to reach research conclusions that would assist them financially (Carrick-Hagenbarth and Epstein 2012).

Evidence for "p-hacking," or selectively carrying out statistical analyses until a statistically significant regression coefficient is discovered, can be assessed by analyzing the distribution of p-values occurring in the literature: if a substantial amount of p-hacking occurs, then it is expected that a large number of p-values very near the $p = .05$ significance threshold will be observed (Head et al. 2015). Head et al. (2015) performed a systematic review analyzing p-curves that represent literature from a variety of

disciplines and found significant evidence that p-hacking occurs based on a discontinuity in the p-curves: there was a much higher concentration of p-values just below .05 than anywhere else, which would not be expected to occur in the absence of p-hacking. Determining whether to drop outliers based on the effect on coefficients is a form of p-hacking, although it is unknown to what degree the results of Head et al. (2015) depend on outlier exclusion rather than other QRPs.

Perhaps the strongest evidence for the existence of QRPs is provided by the cases in which individual scientists have been discovered to have performed fraudulent research. Lacetera and Zirulia (2011) examined scientific fraud in several very high-profile cases, concluding that "examples abound of scientists who falsified, fabricated, or plagiarized findings and were still able to publish and get recognition from them." The rate of outright detection of fraud, however, is quite low: in the social sciences, it was found to be only .002% of papers, many of which were retracted due to errors, often detected by the original authors themselves, rather than fraud (Lu et al. 2013). The true prevalence of fraudulent research is certain to be higher than is suggested only by the number of cases that are detected (Lu et al. 2013), but detection of fraud certainly provides evidence that researchers respond to the incentives that exist to exploit questionable research methods (Lacatera and Zirulia 2011).

Strong evidence exists that the use of QRPs is widespread: admissions of QRP usage by surveyed researchers, a high rate of replication failures, statistical evidence that comes from systemic reviews, and discovered cases of researcher fraud all provide evidence that QRPs are not uncommon among several disciplines. A few of these pieces

of evidence, such as some surveys of researchers, provide evidence specific to selective outlier omission (Fanelli 2009), but many do not: when a replication failure occurs, for example, the reasons for the failure are often not forthcoming (McCullough et al. 2008). Even if the prevalence of any particular QRP cannot be estimated, the evidence suggests a willingness among researchers to commit QRPs, including selective outlier omission.

2.4 Scientific consequences of QRPs

Engaging in QRPs increases the odds that research conclusions will be reached in error (Ioannidis 2005). Additionally, in cases where there exist statistically significant relationships between variables that are being tested, QRPs can inflate the observed effect sizes, resulting in an overstatement of the analytical significance of results (Bakker et al. 2012). Ioannidis (2005) conducted simulations to evaluate the consequences of contemporary research practices, finding that "for most study designs and settings, it is more likely for a research claim to be false than true." While a number of factors that were unrelated to QRPs were mentioned as contributors to this result, flexibility in research design that allows researchers to engage in QRPs was identified as one of the reasons for the preponderance of false research claims (Ioannidis 2005).

The costs involved in generating false research claims are large: every false research claim involves researchers spending time and money to create erroneous results. Chalmers and Glasziou (2009) examined research in biomedical sciences and estimated that 85% of research effort is wasted (as cited by Ioannidis 2014). Dissemination of false information potentially leads others to accept false research claims, which can result in poor policy outcomes, medical choices, or other decisions: in

the case of medicine, decisions made based on false research conclusions can cause significant negative health impacts, including death (Goldacre 2014). Similarly, research claims that are made in error in economics can lead to undesirable policy prescriptions with potentially large-scale financial consequences (Carrick-Hagenbarth and Epstein 2012). False research claims can persist for decades before they are corrected, being used to inform decision makers and other researchers for the duration (Pashler and Harris 2012). Even papers that were discovered to have reached their conclusions in error and were retracted have been found to continue being cited, with "half or more of the future citations [continuing] to accept the original claims." (Lu et al. 2013).

When a variety of evidence is available, systematic meta-analysis is often used as a tool to combine and analyze the available data (Stanley 2001). Statistical flukes and errors that are made randomly can be corrected for by using meta-analysis, but meta-analysis can fail to produce accurate results when the results of the literature are systematically biased (Goldacre 2014). This is a commonly criticized consequence of publication bias in particular (Kotiaho and Tomkins 2002), but meta-analyses can be made inaccurate by other QRPs if they cause distortions that systematically skew results in one direction (Goldacre 2014).

QRPs, then, have the potential to frequently result in erroneous research conclusions which can persist for decades, and the consequences of decisions based on false research claims can be large. Researchers are required to spend time and effort sorting out which research claims are made in error, and failure to do so can result in future research being derivative of the outcomes of erroneous past research. It is also

possible that meta-analysis, a powerful tool to analyze data even in the presence of random errors, can be rendered inaccurate by systematic QRPs.

3. THEORY & METHODOLOGY

This study examines in depth the propensity for one QRP—selectively omitting what is deemed to be outlying data when it will help to achieve a researcher's "desired" outcomes—to inflate type I error rates. Simulations have been performed by Bakker and Wicherts (2014b), who simulated the effect of selective outlier omission on type I error rates using normally distributed data and data from actual psychological datasets, identifying outlying data based on z-score thresholds. This study further analyzes the effects of selectively omitting outliers from samples drawn from chi-squared and logistic distributions with various parameters. Additionally, this study analyzes selective omission of the extreme data that a researcher would wish to omit: that is, rather than only analyzing z-score trimming, this study examines trimming only the extreme values in one tail of the collected data. Given the observed variability in outlier omission rules (Simmons et al. 2011) and the significant evidence that data is excluded without disclosure (Bakker and Wicherts 2014a; LeBel et al. 2013), this form of one-sided omission could be occurring. Additionally, larger sample sizes than those used in Bakker and Wicherts (2014b) are analyzed: while they used sample sizes of up to 500 observations, this study additionally examines sample sizes of 1,000 and 5,000, which might be on the order of those used in large-scale studies in economics and finance.

The simulation attempts to model the following behavior: a researcher is performing an experiment using posttest-only analysis. Two groups are randomly drawn from the same population and an intervention is performed on one group, while another group serves as a control. The intervention is ineffective, such that the outcome variable

for either group is the same in expectation; any statistically significant intervention effects that are observed will represent a type I error. The researcher is attempting to find evidence that the intervention significantly affects the outcome variable. Exploratory analysis is used, such that data are identified as outliers and excluded only when doing so will result in the observation of a statistically significant intervention effect. Different amounts of data exclusion are analyzed until a statistically significant intervention effect is observed or the researcher determines that a maximum allowable amount of exclusion has occurred.

The statistical test used is an ordinary least squares regression with the following model:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

where:

$$x_i = \begin{cases} 1 \; if \; intervention \; is \; received \\ 0 \; if \; intervention \; is \; not \; received \end{cases}$$

and y_i is drawn at random from the distribution selected to be used in the simulation. $\hat{\beta}_1$is the estimated sample coefficient in any given trial. The magnitude and statistical significance of $\hat{\beta}_1$ is the variable of interest: it represents the observed "treatment effect" of the intervention. Because the samples are drawn from the same distribution, the true population value of β_1 is zero, and any rejection of the null hypothesis, that $\beta_1 = 0$, is a type I error.

In expectation, without implementing any QRPs, a statistically significant treatment coefficient $\hat{\beta}_1$ occurs at the $p = .05$ level in 5% of cases by common

convention. The regression is performed. If $\hat{\beta}_1$ is not found to be significant at the p = .05 level, then one percent of the most extreme values from one tail of the intervention group in the sample are omitted and the regression is repeated on the trimmed sample. Data will continue to be trimmed and the regression will continue to be run with the modified data until either a statistically significant $\hat{\beta}_1$ coefficient is observed or a maximum trimming threshold of 5% is reached, whichever occurs first. This process is repeated with trimming rules that are based on z-score thresholds rather than percentile thresholds, excluding results that exceed the z = 3, z = 2.5, and z = 2 thresholds, again proceeding to exclude progressively more until a statistically significant result has been observed or until all three z-score thresholds have been tested.

This is simulated in Stata (the code used to generate the results is available in appendices A and B) by generating a "treatment" sample with *n* observations with a value of 1 assigned to the treatment dummy variable and a "control" sample with *n* observations with a value of 0 assigned to the treatment dummy variable. The outcome variable for each group is generated from the same random population distribution, using a variety of different population distributions and sample sizes to analyze the effects of selective trimming under different assumptions: the standard normal distribution is used to analyze the effects of performing the omission procedure on a symmetric distribution. The logistic distribution with mean = 0 and scale = 1 is used because the distribution has greater kurtosis than the standard normal distribution; simulating both a logistic distribution and a standard normal distribution allows for comparing the effects of selective trimming on samples taken from populations with symmetric distributions with

different tail thickness. Underlying population data in real datasets is often skewed, so asymmetric chi-squared distributions of various skewness are analyzed. The effects of selective trimming on samples taken from chi-squared distributions with 4, 10, and 50 degrees of freedom is assessed to evaluate the importance of varying levels of skewness on resultant type I error rates. Sample sizes of 100, 500, 1000, and 5000 are used because they are in the range of sample sizes for datasets analyzed in economics.

For the simulations testing one-tailed exclusion, x_i is regressed on y_i; if a statistically significant coefficient on x_i is observed, then no further analysis is performed. If it is not, then the regression is re-run with a new dataset that excludes observations in the treatment group with the smallest 1% of observed outcomes y_i if left-tailed exclusion is being analyzed, or the largest 1% of observed outcomes y_i if right-tailed exclusion is being analyzed. If the resulting coefficient on x_i is statistically significant, then it is recorded; if not, the regression is re-run with a new dataset that excludes 2% rather than 1% of the most extreme values, again omitting from only the chosen tail. This process is repeated either until either 5% of the tail has been trimmed or until the coefficient $\hat{\beta}_1$ is observed to be statistically significant, whichever occurs first.

The simulations testing trimming rules based on z-scores are broadly similar: x_i is regressed on y_i, and if a statistically significant coefficient $\hat{\beta}_1$ is observed then no other analyses will be performed. Otherwise, a new dataset containing all observations in the control group and only those observations in the treatment group that have z-scores less than 3 is created. The regression is re-run on the truncated data. This is repeated, creating datasets that exclude observations with z-scores greater than 2.5 and 2, until

either a statistically significant coefficient $\hat{\beta}_1$ has been observed or all three z-score exclusion rules have been implemented, whichever occurs first.

Each of these two types of simulation is performed on each possible combination of distribution choice and sample size, for a total of 40 simulations (5 distributions, 4 sample sizes, and 2 simulated exclusion rules). In each simulation, 20,000 iterations of the described process are performed. Each time a type I error is observed to occur, the trimming threshold that was used to generate the type I error is recorded. Additionally, if a type I error is not observed prior to data exclusion but is observed after data exclusion, the difference in coefficient size resulting from the data exclusion is recorded.

4. RESULTS

All of the results that follow are shown in detail in Tables 1 through 6 in Appendix D and summarized in Figures 1 through 12 in Appendix C.

4.1 Normal distribution

None of the results sampling from normally distributed data are sensitive to the mean or standard deviation of the distribution. Because of this, all reported results for data drawn from normal distributions are taken from the standard normal distribution.

For normally distributed data and using only one-tailed trimming, the probability of type I error was observed to be highly dependent on sample size. As expected, trimming from the upper tail results in a negative coefficient $\hat{\beta}_1$ and trimming from the lower tail results in positive coefficient $\hat{\beta}_1$, but the probability of type I error does not depend on which tail is trimmed. The probability of type I error was observed to range from 15.18% for $n = 100$ to 100% for $n = 5000$.

Trimming based on z-score thresholds results in much smaller but still inflated type I error rates, from 7.8% to 8.1%. Type I error rates were observed to consistently increase very slightly as sample size increased.

4.2 Logistic distribution

The results for randomly generated data taken from the logistic distribution with mean = 0 and scale = 1 and with outliers selectively trimmed from one tail are similar to the results from the standard normal distribution: type I error rates depend significantly on sample size and do not depend on which tail was being excluded. The thicker tails associated with the logistic relative to the standard normal distribution causes data

exclusion to result in a somewhat larger probability of type I error; for n = 1,000, for example, the percentage of type I error rates for samples taken from the logistic distribution is approximately 78%, while that of the normal distribution with the same specifications is approximately 70%. Type I error rates for the logistic range from 16.25% for n = 100 to 100% for n = 5,000.

The results based on selective z-score exclusion are of similar magnitude for data drawn from the standard normal and logistic distributions. Type I error rates in simulations using samples drawn from logistic distributions were observed to be slightly larger than similar simulations using data sampled from standard normal distributions, ranging from 8.2% to 8.5%. There is no consistent relationship between sample size and probability of type I error.

4.3 Chi-squared distributions

Results for omissions performed on data sampled from chi-squared distributions, as expected, are highly dependent on which tail of the distribution is being trimmed. Because the distributions are skewed rightward, eliminating observations from the right tail excludes significantly more extreme values than eliminating observations from the left tail, resulting in much higher type I error rates. Type I error rates have a positive relationship with sample size for each of the three distributions analyzed. When the left side of the samples are omitted, type I error rates are lower for samples drawn from more skewed distributions. Similarly, when the right tail of the samples are excluded, type I error rates are greater for data sampled from more skewed distributions. The lowest observed type I error rate in any of these simulations involving one-sided data exclusion,

9.5%, was observed for data sampled from the chi-squared distribution with 4 degrees of freedom and n = 100. Type I error rates were still observed to be up to 100% for n = 5000 and underlying distributions with 10 or 50 degrees of freedom, and 92.87% for data drawn from the distribution with 4 degrees of freedom.

In the simulations wherein data is trimmed based on z-score thresholds, type I error rates are slightly smaller than in simulations where data is trimmed selectively from the right tail. Type I error rates were observed to be larger, in all cases, for skewed distributions than for symmetric distributions, ranging from 10.2% for samples drawn from a chi-squared distribution with 50 degrees of freedom and n = 100 to 100% for samples drawn from chi-squared distributions with 4 or 10 degrees of freedom and n = 5000.

4.4 Coefficient sizes

In all cases, the one-sided exclusion of outliers was observed to have the expected effects: eliminating data from the right tail of a sample was found to systematically reduce observed $\hat{\beta}_1$ coefficients, and eliminating data from the left tail of a sample was found to increase observed $\hat{\beta}_1$ coefficients. The change in $\hat{\beta}_1$ coefficient size from systematic trimming depends negatively on sample size: the largest changes were consistently observed for n = 100, while the smallest were observed for n = 5,000. As expected, excluding more extreme data results in larger changes in coefficient size: eliminating the right tail of data drawn from a chi-squared distribution was observed to cause the largest change in coefficient size, while eliminating the left tail of data from the same distribution was observed to cause the smallest change in coefficient size in all

cases. The largest average change in coefficient size is .1002 standard deviations, resulting from trimming the right tail of data taken from a chi-squared distribution with 4 degrees of freedom and $n = 100$, while the smallest observed average change is .03313 standard deviations for data taken from the same distribution and for $n = 5{,}000$.

When exclusion is performed based on z-score thresholds, changes in coefficient sizes are quite small for symmetric distributions and are negatively related to sample size. The largest observed change for symmetric distributions is a change of .056 standard deviations, and the smallest is only .00064 standard deviations. For right-skewed distributions, the changes are on the same order of magnitude of those observed when samples are trimmed from the right tail only: they range from .035 standard deviations at the smallest to .13 standard deviations at the largest.

5. DISCUSSION

5.1 General discussion of results

For simulated one-sided trimming rules, shown in Tables 1 and 2, the probability of type I error ranges from 9.5% to 100%, depending on the simulation specification: with sufficient sample size and a sufficiently skewed distribution, a researcher could *always* discover a statistically significant relationship where one does not exist by systematically omitting 5% of treatment observations. In all cases, systematically trimming one side of any distribution results in dramatically larger type I error rates: a type I error rate of 9.5%, the smallest observed, is almost double the purportedly expected type I error rate of 5%. In the simulation with the *smallest* observed type I error rates, systematically eliminating just 1% of the data was observed to cause type I error rates to increase to 5.59%.

The probability of type I error resulting from systematic data exclusion from only one tail depends critically on whether and to what degree the population from which the data is drawn is skewed. Eliminating observations from the right tail of a sample taken from a heavily right-skewed distribution was observed to have the largest effect on both coefficient size and on type I error rates. Simmons et al. (2011) claim that it might be the case that researchers will be more likely to engage in QRPs when they could more easily be justified to themselves or to others; excluding the most extreme data from a sample that is heavily lopsided might seem to be justifiable. Eliminating observations from the left side of right-skewed distributions was observed to have the smallest effect on type I error rates, and might rarely take place: identifying observations on the left side of a

sample taken from a heavily right-skewed distribution as outliers would likely be difficult for a researcher to justify. This would restrict the ability for researchers to influence the sign of their findings: eliminating data from only the right side of right-skewed data was always observed to result in negative treatment coefficients, so a researcher hoping to find positive treatment coefficients and acting in a way that seems somewhat justifiable would not be able to exploit this QRP. However, inasmuch as justification for or even disclosure of outlier omission often does not take place, it is possible that these manipulations could still occur.

Excluding data based on z-score exclusion rules was found by Bakker and Wicherts (2014b) to be a common and accepted practice that researchers justified in a number of studies. For symmetric distributions, this study found that performing this practice inflates type I error rates, but by a much smaller amount than single-tailed data exclusion: simulated type I error rates range from 7.8% to 8.6%. But for right-skewed distributions, the results with the highest z-scores will primarily be those in the right tail of the distribution. Because of this, excluding data based on commonly accepted z-score thresholds was observed to result in only a slightly smaller type I error chance than selectively excluding up to 5% of the right tail: for $n = 100$ and samples drawn from chi-squared distributions with 4 degrees of freedom, for example, type I error rates were found to be 21.33% when z-score trimming rules were used and 22.67% when up to 5% of the largest of observations were omitted. With $n = 5000$ and using data from the same distribution, type I error rates occurred in 100% of simulations with either trimming rule.

Overall, this study suggests that the lack of defined standards for the treatment of outliers has significant potential to generate false research findings, consistent with the conclusions of Bakker and Wicherts (2014b). For the most conservative assumptions of trimming, based on z-score thresholds and with a sample size of $n = 100$, the smallest observed type I error chance was 7.8%, over half again as large as the expected 5% type I error chance in the absence of outlier exclusion. As mentioned, several simulations found that, under certain assumptions, false positives could be generated in 100% of simulations. This was found to occur even in some simulations where z-score trimming thresholds were used, as is commonly accepted practice (Bakker and Wicherts 2014b).

5.2 Discussion of solutions

Solutions to the issue of selective data omission, many of which extend to QRPs more generally, broadly fall into two categories: creating and enforcing rules that provide disincentives for researchers to engage in QRPs, and conducting follow-up research that limits the consequences of individual inaccurate or exaggerated research claims. In their review of replication attempts in the social science literature, Pashler and Harris (2012) determine that "there is every reason to believe that the great majority of errors that do enter the literature will persist uncorrected indefinitely, given current practices." This suggests the importance both of ensuring that research claims are not made in error, as the errors will persist, and of modifying academic practices to make correction of errors a greater priority.

5.2.1 Limiting or disincentivizing QRP usage

In the simulations conducted for this study, type I errors in excess of 5% were only observed when data were selectively omitted: in the absence of outlier exclusion, type I error rates were, as predicted, found to be roughly 5%. Directly preventing or disincentivizing researchers from exploiting this QRP, then, is a potential solution. However, as mentioned by Bakker and Wicherts (2014b), there are circumstances when omitting outlying data is desirable—an ideal solution would permit data to be excluded in these cases, while preventing questionable data omission. A number of solutions exist that aim to provide disincentives for engaging in this and other QRPs while still allowing researchers to omit data when it is appropriate.

Simmons et al. (2011) proposed a requirement that any researcher who reports statistical results that are based on data with omitted observations must also report the same results with the full, non-trimmed data. This requirement, if observed, would make it clear when an author's results—both in terms of magnitude and statistical significance—are the result of omitting observations, while still allowing researchers to exclude data when they can provide a justification for doing so (Simmons et al. 2011). This proposal, Simmons et al. (2011) warned, would not be likely to entirely prevent the practice: researchers might simply fail to disclose that data were omitted. If doing so were a violation of a publication requirement, however, researchers might be deterred from performing unreported data exclusion (Simmons et al. 2011).

An alternative solution addresses the issue of ex-post flexibility in determining data exclusion rules. As mentioned, there is no commonly accepted practice for deciding

which data are outliers worthy of exclusion: systematic reviews have found that researchers have employed a wide variety of exclusion rules, with some eliminating no outliers and others excluding up to 10% of observed data (Simmons et al. 2011; Bakker and Wicherts 2014b). Additionally, 38% of surveyed psychological researchers admitted to determining whether to exclude data after observing the impact of doing so (John et al. 2012). These facts, taken together, demonstrate enormous flexibility in data exclusion that is often exploited. Pre-registration of the studies that will take place and committing to research methods in advance has been proposed as a solution to issues of research flexibility: if researchers were compelled to commit to criteria for outlier identification and removal ex-ante, then this would reduce the problematic flexibility in determining how to exclude data ex-post (Bakker and Wicherts 2014b; Wagenmakers et al. 2012).

A review of economics and business journals conducted by Karabag and Berggren (2012) found that a large majority of economics and business journals do not have explicitly stated policies regarding academic dishonesty. Furthermore, the review found that, in cases where academic dishonesty had been discovered, the consequences for dishonest authors were inconsistent and were often not made public (Karabag and Berggren 2012). These findings suggest a lenient incentive structure for researchers who are deciding whether to engage in potentially dishonest behaviors, including systematic data exclusion: if there are neither explicit rules nor known consequences regarding the behavior, then the disincentives from engaging in it are weak. Journal policies or discipline-wide guidelines that explicitly restrict and provide consequences for unwarranted data exclusion would possibly disincentivize the behavior.

Using positive results under NHST as an important criterion for publication has been met with criticism: Ziliak and McCloskey (2004) argue that, even in the absence of QRPs, NHST is not always appropriate. They and other critics of NHST point to the trend of researchers often making no determination of analytical or "economic" significance of their results, with statistical significance being the dominant evaluative tool (Altman 2004; Ziliak and McCloskey 2004). While the focus of these authors is not on QRPs, the incentives to engage in QRPs with the intention of crossing statistical significance thresholds are certainly larger the more important that statistical significance is deemed to be.

Increasing the frequency with which replication attempts are made has also been suggested as a means for reducing the proportion of false research findings: replication is sometimes seen as "a threat that might keep potential cheaters honest" (Hamermesh 2007). A higher probability of a replication attempt being made on a researcher's work represents a higher probability of the discovery of false results: if a researcher knows that results that depend on inappropriate data omission might be discovered to be questionable through replication attempts, then he or she might be less likely to engage in the dubious behavior.

5.2.2 Limiting the consequences of inaccurate research findings

As mentioned, mandatory disclosure requirements that require researchers to report in detail whether statistical findings depend on data omission might prevent them from excluding data inappropriately (Simmons et al. 2011). Disclosure requirements of this sort have the additional benefit of providing more information to those who are

interpreting research findings (LeBel et al. 2013). A reader of an academic publication, for example, would better be able to judge the robustness of a research finding if any data omission were fully disclosed. The impact and dissemination of research findings are likely to depend on how robust the results are perceived to be: disclosure of any practice that might make results questionable, including data omission, would therefore be likely to reduce the negative potential of spurious results.

In the previous section, replication was discussed as a means for disincentivizing researchers from engaging in questionable research practices. Replication is often additionally suggested as a method for reducing the impact of research conclusions that are made in error (Hamermesh 2007). All three discussed types of replication can be used to limit the impact of false research claims.

Pure replication attempts can be used to discover and analyze any data exclusions that are being performed prior to data analysis: if a result depends on undisclosed data exclusion, then this would be revealed through a pure replication attempt on the original data. The ability for pure replication to take place depends on the data on which the author's research claims depend being publicly available, which is often a publishing requirement (Hamermesh 2007). As mentioned, even in cases where it is a requirement to make the data and/or code from which research conclusions are drawn publicly available, the requirement is often shirked with no consequences for the violating researchers (Dewald et al. 1986; McCullough 2007). If mandatory data submission were a more common requirement and/or the requirement were more strictly enforced, then there would be improved ability to make pure replication attempts and to discover

otherwise undisclosed data exclusion, in addition to other QRPs. Even with pure replication attempts, however, it is possible that undisclosed data exclusion would be missed: if a researcher trims his or her data prior to making it available, then a successful replication attempt could be made without detecting that the undisclosed exclusion took place (Camfield and Palmer-Jones 2013).

Statistical and scientific replication can be used to attempt to verify a study's findings regardless of the availability of the data analyzed for the study (Hamermesh 2007). If a study makes false research claims for any reason, including the potential reason of a result depending on inappropriate data exclusion, then these claims might be contradicted by statistical or scientific replication attempts (Schmidt 2009). Replication, in this way, "can provide a useful check on the spread of incorrect results" (Duvendack et al. 2015).

As mentioned, Hamermesh (2007) found that pure replication attempts very rarely take place, with no data requests taking place for over 60% of papers with freely available data. Similarly, economists rarely have the incentives to make statistical or scientific replication attempts (Mirowski and Sklivas 1991). Replication has been described as having a "lack of popularity among economists" (Camfield and Palmer-Jones 2013) for a variety of reasons, including small professional rewards and sometimes even negative consequences for replication attempts. Attempts by journals to increase the amount of replication that takes place have largely failed (Hamermesh 2007), but increasing the number of replications that take place remains as a possible method for reducing the frequency and impact of QRPs.

Meta-analysis has been suggested as another potential method to mitigate the effects of QRPs (Head et al. 2015). Empirically, there is some evidence that the impact of QRPs on coefficient sizes has been small enough that the impact on meta-analytic reviews would not be dramatic (Head et al. 2015). The findings of this study are not entirely consistent with this finding: as mentioned in Section 4.4, in simulations where only one side of each sample is omitted, pre- and post-trimming differences in statistically significant results were observed to be as much as .100 standard deviations. In simulations where outlier identification is based on z-score thresholds, coefficient size differences were observed to be relatively small for large symmetric distributions, but potentially large in all other cases, ranging from .035 to .127 standard deviations. Additionally, meta-analytic review can depend on replications having taken place (Duvendack 2015). While they remain uncommon in economics, meta-analyses are sometimes conducted (Stanley 2001) and can conceivably be used to reduce the apparent effect of QRPs (Head et al. 2015), possibly including data exclusion.

5.3 Limitations

The data exclusion rule analyzed in this study is one of a number that could be used: Simmons et al. (2011) found that, in studies that reported excluding outliers, a large number of outlier detection and exclusion rules were employed. Exclusion that takes place in practice, including unreported exclusion, could involve exclusion of a much larger or smaller percentage of observations than this study analyzed. Additionally, the simulations in this study were designed to represent "exploratory" data exclusion, where data is continuously omitted until either the 5% maximum trimming

threshold or a statistically significant result is achieved. It is unknown to what degree the results would generalize to selective data exclusion that takes place without following the exact algorithms employed in the simulations used in this study.

The simulations additionally are based on assumptions about the distributions of the outcome variable for the underlying population: they are drawn from the standard normal distribution, the logistic distribution with mean = 0 and scale = 1, and chi-squared distributions with various degrees of freedom. It is likely that the results based on data drawn from these distributions would generalize to selective trimming that is undertaken on data drawn from similarly distributed populations. However, it is unknown how the results of this study would generalize if the same data exclusion rules were applied to populations which do not follow the distributional assumptions of the simulations.

Treatment effects in the study are always assumed to be zero. If there were an underlying treatment effect, then the effect of selective data omission would not be the same: in the cases where, for example, a positive treatment effect exists and data exclusion is done in a way that makes detection of a positive treatment effect more likely, then the exclusion would *reduce* the chance of a type II error. The increase in probability of type I error in the absence of an underlying treatment effect is the only possibility that was explored.

6. CONCLUSIONS

While replication and meta-analysis can be effectively used to reduce the impact of research conclusions that are made in error, in several cases they may not offer solutions to the empirical problems that economics presents. Economic experiments frequently involve expensive interventions which can and sometimes do make statistical or scientific replication attempts prohibitively expensive; while it might be normal in medicine for several studies to analyze the same intervention (Goldacre 2014), the same can rarely be said for economics. Empirically, as described, replication attempts are rarely made in economics (Hamermesh 2007). Also as described, it has been demonstrated across several fields that false research claims, even those that are retracted, are still cited and disseminated (Lu et al. 2013). These facts underscore the importance of reducing the type I error rate: if funding and incentives for replication are scarce and false results cause lasting harm, then preventing false positives from entering the literature as much as possible is likely to be the most effective strategy for limiting the dissemination of erroneous economic research claims.

Wagenmakers et al. (2011) point to "the relative ease with which an inventive researcher can produce statistically significant results even when the null hypothesis is true." The results of this study are broadly consistent with this claim. The literature-wide consequences of any QRP are likely to depend on a combination of the flexibility in and consequences of its use. The literature demonstrates that whether and how a researcher is able to identify and omit outliers is extremely flexible (Simmons et al 2011; Bakker and

Wicherts 2014b). The results of this study provide evidence that selective outlier omission has significant potential to generate false positive results.

REFERENCES

Altman, Morris. “Statistical Significance, Path Dependency, and the Culture of Journal Publication.” *The Journal of Socio-Economics* 33, no. 5 (2004): 651–63.

Angrist, Joshua, and Jörn-Steffen Pischke. “The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics.” National Bureau of Economic Research, 2010. http://www.nber.org/papers/w15794.

Bailey, Charles N., James R. Hasselback, and Julia N. Karcher. “Research Misconduct in Accounting Literature: A Survey of the Most Prolific Researchers’ Actions and Beliefs.” *Abacus* 37, no. 1 (2001): 26–54.

Bakker, Marjan, Annette van Dijk, and Jelte M. Wicherts. “The Rules of the Game Called Psychological Science.” *Perspectives on Psychological Science* 7, no. 6 (2012): 543–54.

Bakker, Marjan, and Jelte M. Wicherts. “Outlier Removal and the Relation with Reporting Errors and Quality of Psychological Research.” *PLoS ONE* 9, no. 7 (July 29, 2014b): e103360. doi:10.1371/journal.pone.0103360.

———. “Outlier Removal, Sum Scores, and the Inflation of the Type I Error Rate in Independent Samples T Tests: The Power of Alternatives and Recommendations.” *Psychological Methods* 19, no. 3 (2014a): 409.

Begley, C. Glenn, and Lee M. Ellis. “Drug Development: Raise Standards for Preclinical Cancer Research.” *Nature* 483, no. 7391 (2012): 531–33.

Camfield, Laura, Maren Duvendack, and Richard Palmer-Jones. “Things You Wanted to Know about Bias in Evaluations but Never Dared to Think.” *IDS Bulletin* 45, no. 6 (2014): 49–64.

Camfield, Laura, and Richard Palmer-Jones. “Three ‘Rs’ of Econometrics: Repetition, Reproduction and Replication.” *Journal of Development Studies* 49, no. 12 (2013): 1607–14.

Carrick-Hagenbarth, Jessica, and Gerald A. Epstein. “Dangerous Interconnectedness: Economists’ Conflicts of Interest, Ideology and Financial Crisis.” *Cambridge Journal of Economics* 36, no. 1 (2012): 43–63.

Chalmers, Iain, and Paul Glasziou. “Avoidable Waste in the Production and Reporting of Research Evidence.” *The Lancet* 374, no. 9683 (2009): 86–89.

Chan, An-Wen, Asbjørn Hróbjartsson, Mette T. Haahr, Peter C. Gøtzsche, and Douglas G. Altman. “Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles.” *Jama* 291, no. 20 (2004): 2457–65.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. “Replication in Empirical Economics: The Journal of Money, Credit and Banking Project.” *The American Economic Review* 76, no. 4 (1986): 587–603.

Duvendack, Maren, Richard W. Palmer-Jones, and W. Robert Reed. “Replications in Economics: A Progress Report.” *Scholarly Comments on Academic Economics* 12, no. 2 (2015): 164–91.

Fanelli, Daniele. “Negative Results Are Disappearing from Most Disciplines and Countries.” *Scientometrics* 90, no. 3 (2011): 891–904.

Fanelli, Daniele, others. “How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data.” *PloS One* 4, no. 5 (2009): e5738.

Fidler, Fiona, Mark A. Burgman, Geoff Cumming, Robert Buttrose, and Neil Thomason. “Impact of Criticism of Null-Hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology.” *Conservation Biology* 20, no. 5 (2006): 1539–44.

Freedman, David, Robert Pisani, and Roger Purves. *Statistics (4th Edn)*. Norton, New York, 2007.

Freedman, Leonard P., Iain M. Cockburn, and Timothy S. Simcoe. “The Economics of Reproducibility in Preclinical Research.” *PLoS Biol* 13, no. 6 (June 9, 2015): e1002165. doi:10.1371/journal.pbio.1002165.

Gigerenzer, Gerd. “Mindless Statistics.” *The Journal of Socio-Economics* 33, no. 5 (2004): 587–606.

Goldacre, Ben. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. Macmillan, London, 2014.

Hamermesh, Daniel S. “Six Decades of Top Economics Publishing: Who and How?” *Journal of Economic Literature* 51, no. 1 (2013): 162-172

———. “Viewpoint: Replication in Economics.” *Canadian Journal of Economics/Revue Canadienne D’économique* 40, no. 3 (2007): 715–33.

Hamermesh, Daniel S., and Gerard A. Pfann. “Reputation and Earnings: The Roles of Quality and Quantity in Academe.” *Economic Inquiry* 50, no. 1 (2012): 1–16.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. “The Extent and Consequences of P-Hacking in Science.” *PLoS Biol* 13, no. 3 (2015): e1002106.

Hróbjartsson, Asbjørn, Ann Sofia Skou Thomsen, Frida Emanuelsson, Britta Tendal, Jørgen Hilden, Isabelle Boutron, Philippe Ravaud, and Stig Brorson. “Observer Bias in Randomised Clinical Trials with Binary Outcomes: Systematic Review of Trials with Both Blinded and Non-Blinded Outcome Assessors.” *BMJ* 344 (2012): e1119.

Ioannidis, John, and Chris Doucouliagos. “What’S To Know About The Credibility Of Empirical Economics?” *Journal of Economic Surveys* 27, no. 5 (2013): 997–1004.

Ioannidis, John PA. “How to Make More Published Research True,” 2014. http://dx.plos.org/10.1371/journal.pmed.1001747.

———. “Why Most Published Research Findings Are False.” *Chance* 18, no. 4 (2005): 40–47.

John, Leslie K., George Loewenstein, and Drazen Prelec. “Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling.” *Psychological Science*, 2012, 0956797611430953.

Karabag, Solmaz Filiz, and Christian Berggren. “Retraction, Dishonesty and Plagiarism: Analysis of a Crucial Issue for Academic Publishing, and the Inadequate Responses

from Leading Journals in Economics and Management Disciplines." *Journal of Applied Economics and Business Research* 2, no. 4 (2012): 172–83.

Kotiaho, Janne S., and Joseph L. Tomkins. "Meta-Analysis, Can It Ever Fail?" *Oikos*, 2002, 551–53.

Krimsky, Sheldon, and L. S. Rothenberg. "Financial Interest and Its Disclosure in Scientific Publications." *Jama* 280, no. 3 (1998): 225–26.

Lacetera, Nicola, and Lorenzo Zirulia. "The Economics of Scientific Misconduct." *Journal of Law, Economics, and Organization* 27, no. 3 (2011): 568–603.

Leamer, Edward E. "Let's Take the Con out of Econometrics." *The American Economic Review*, 1983, 31–43.

LeBel, Etienne P., Denny Borsboom, Roger Giner-Sorolla, Fred Hasselman, Kurt R. Peters, Kate A. Ratliff, and Colin Tucker Smith. "PsychDisclosure.org Grassroots Support for Reforming Reporting Standards in Psychology." *Perspectives on Psychological Science* 8, no. 4 (July 1, 2013): 424–32. doi:10.1177/1745691613491437.

Lu, Susan Feng, Ginger Zhe Jin, Brian Uzzi, and Benjamin Jones. "The Retraction Penalty: Evidence from the Web of Science." *Scientific Reports* 3 (2013). http://www.nature.com/srep/2013/131106/srep03146/full/srep03146.html?message-global=remove.

McCullough, B. D. "Got Replicability? The _Journal of Money, Credit and Banking_ Archive." *Econ Journal Watch* 4, no. 3 (2007): 326–37.

McCullough, Bruce D., Kerry Anne McGeary, and Teresa D. Harrison. “Do Economics Journal Archives Promote Replicable Research?” *Canadian Journal of Economics/Revue Canadienne D’économique* 41, no. 4 (2008): 1406–20.

O’Keefe, Suzanne, and Ta-Chen Wang. “Publishing Pays: Economists’ Salaries Reflect Productivity.” *The Social Science Journal* 50, no. 1 (2013): 45–54.

Papanikolaou, George N., Maria S. Baltogianni, Despina G. Contopoulos-Ioannidis, Anna-Bettina Haidich, Ioannis A. Giannakakis, and John PA Ioannidis. “Reporting of Conflicts of Interest in Guidelines of Preventive and Therapeutic Interventions.” *BMC Medical Research Methodology* 1, no. 1 (2001): 3.

Pashler, Harold, and Christine R. Harris. “Is the Replicability Crisis Overblown? Three Arguments Examined.” *Perspectives on Psychological Science* 7, no. 6 (2012): 531–36.

Rosenthal, Robert. “Experimenter Effects in Behavioral Research.,” 1966. http://doi.apa.org/psycinfo/1967-09647-000.

———. “How Often Are Our Numbers Wrong?” *American Psychologist* 33, no. 11 (1978): 1005.

———. “Science and Ethics in Conducting, Analyzing, and Reporting Psychological Research.” *Psychological Science* 5, no. 3 (1994): 127–34.

Schmidt, Stefan. “Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences.” *Review of General Psychology* 13, no. 2 (2009): 90.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. “False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science*, 2011, 0956797611417632.

Stanley, Tom D. “Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection*.” *Oxford Bulletin of Economics and Statistics* 70, no. 1 (2008): 103–27.

———. “Wheat from Chaff: Meta-Analysis as Quantitative Literature Review.” *Journal of Economic Perspectives*, 2001, 131–50.

Statzner, Bernhard, and Vincent H. Resh. “Negative Changes in the Scientific Publication Process in Ecology: Potential Causes and Consequences.” *Freshwater Biology* 55, no. 12 (2010): 2639–53.

Stern, Jerome M., and R. John Simes. “Publication Bias: Evidence of Delayed Publication in a Cohort Study of Clinical Research Projects.” *Bmj* 315, no. 7109 (1997): 640–45.

Wagenmakers, Eric-Jan. “A Practical Solution to the Pervasive Problems Ofp Values.” *Psychonomic Bulletin & Review* 14, no. 5 (2007): 779–804.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han LJ van der Maas, and Rogier A. Kievit. “An Agenda for Purely Confirmatory Research.” *Perspectives on Psychological Science* 7, no. 6 (2012): 632–38.

Ziliak, Stephen T., and Deirdre N. McCloskey. “Size Matters: The Standard Error of Regressions in the American Economic Review.” *The Journal of Socio-Economics* 33, no. 5 (2004): 527–46.

Zimmermann, Christian. “Academic Rankings with RePEc.” *Econometrics* 1, no. 3 (2013): 249–80.

APPENDIX A

CODE FOR SIMULATIONS USING ONE-TAILED OUTLIER OMISSION WITH PERCENTILE THRESHOLDS

```
#delimit ;
set more off;
clear;
set seed 3;

// make variables

generate treatment = 0;
generate double baseline = .;
generate double trimmed = .;
local success = 0;
local success_pos = 0;
local success_neg = 0;
local successtrim = 0;
local successtrim_pos = 0;
local successtrim_neg = 0;
local total_pre_pos_size = 0;
local total_pre_neg_size = 0;
local total_post_pos_size = 0;
local total_post_neg_size = 0;
local post_effect_neg = 0;
local post_effect_pos = 0;
local pre_effect_pos = 0;
local pre_effect_neg = 0;
local percentile = 0;
local percentile_cutoff = .05;
local count = 0;
local count_percent = 0;
local success1 = 0;
local success2 = 0;
local success3 = 0;
local success4 = 0;
local success5 = 0;
local unif = 0;
local logi = 0;
local iterations=20000;

//This is the sample size. It is changed for different simulations.

local n = 5000;

while `iterations' > 0{;

       local count = 1;
       local trims = 1;
       local trimcount = 1;

       local i = `n' * 2;
       set obs `i';
```

```
/*This generates the dataset. It creates n "treatment" variables and
n "control" variables taken from the specified distribution. The
distribution that is being used is the only one that is not preceded
by a line comment, "//". For logistic distributions, the third and forth
lines represent a calculation of a random number sampled from a logistic
distribution and the fifth line inserts that random number into the
dataset. */

while `i' > 0{;
	replace baseline = rnormal(0, 1) in `i';
	//replace baseline = rchi2(4) in `i';
	//local unif = runiform();
	//local logi = -ln((1 - `unif')/`unif');
	//replace baseline = `logi' in `i';
	replace treatment = 1 in `i' if `i' <= `n';
	replace treatment = 0 in `i' if `i' > `n';
	local i = `i' - 1;
};


regress baseline treatment;

local p = 2*ttail(e(df_r),abs(_b[treatment]/_se[treatment]));

//if p is statistically significant, then stop there. Otherwise,
//try the trimming routine;

if (`p' <= .05){;

	//log whether the observed coefficient is positive or negative

	if (_b[treatment] > 0){;
		local success_pos = `success_pos' + 1;
	};
	else{;
		local success_neg = `success_neg' + 1;
	};
	local success = `success' + 1;
};

else{;

	/*log the insignificant effect size. This can be compared to
	the significant one if the significance threshold is obtained */

	if (_b[treatment] > 0){;
		local pre_effect_pos = _b[treatment];
	};
	else{;
		local pre_effect_neg = _b[treatment];
	};
	replace trimmed = baseline if treatment == 1;
	replace trimmed = . if treatment == 0;
	sort trimmed;
	replace trimmed = baseline if treatment == 0;
	while (`percentile' < `percentile_cutoff'){;
		local percentile = `percentile' + .01;
		di "Percentile: " `percentile';
		local count_percent = `count' / `n';
```

```
di "count_percent = " `count_percent';

/*This loop performs the data trimming. "trimmed" is a
variable containing all of the control observations
and only the treatment observations that are included
after trimming. The first line within the loop is
commented out when right-tailed trimming is considered,
and the third line commented out when left-tailed
trimming is considered. */

while (`count_percent' <= `percentile'){;
       //replace trimmed = . in `count';
       local uppertrim = `n' - `count' + 1;
       replace trimmed = . in `uppertrim';
       local count = `count' + 1;
       local count_percent = `count' / `n';
};
regress trimmed treatment;
local p = 2*ttail(e(df_r),
abs(_b[treatment]/_se[treatment]));

if (`p' <= .05){;

       //if trimming works, then log the new effect
       //size. A comparison between statistical and
       //analytical significance can be drawn

       if (`percentile' == .01){;
              local success1 = `success1' + 1;
       };
       if (`percentile' == .02){;
              local success2 = `success2' + 1;
       };
       if (`percentile' == .03){;
              local success3 = `success3' + 1;
       };
       if (`percentile' == .04){;
              local success4 = `success4' + 1;
       };
       if (`percentile' == .05){;
              local success5 = `success5' + 1;
       };

       local successtrim = `successtrim' + 1;
       local trimcount = `trims' + 1;

       if (_b[treatment] > 0){;
              local post_effect_pos = _b[treatment];

              local total_post_pos_size =
              `total_post_pos_size' + 0 +
              `post_effect_pos';

              local total_pre_pos_size =
              `total_pre_pos_size' + 0 +
              `pre_effect_pos';

              local successtrim_pos =
              `successtrim_pos' + 1;
```

```
                        };
                        else{;

                                local post_effect_neg = _b[treatment];

                                local total_post_neg_size =
                                `total_post_neg_size' + 0 +
                                `post_effect_neg';

                                local total_pre_neg_size =
                                `total_pre_neg_size' + 0 +
                                `pre_effect_neg';

                                local successtrim_neg =
                                `successtrim_neg' + 1;
                        };

                        //break the loops
                        local count_percent = `n' + 1;
                        local percentile = .05;
                    };
                    if (`count_percent' < `n'){;
                        local count_percent = `count' / `n';
                    };
                    di "count_percent = " `count_percent';
                    local trimcount = `trimcount' + 1;

                };
                local percentile = 0;
        };
        local iterations = `iterations' - 1;
        di "iteration " `iterations';
};
if (`successtrim_pos' > 0){;
        local total_pre_pos_size = `total_pre_pos_size' / `successtrim_pos';
        local total_post_pos_size = `total_post_pos_size' / `successtrim_pos';
};
if (`successtrim_neg' > 0){;
        local total_pre_neg_size = `total_pre_neg_size' / `successtrim_neg';
        local total_post_neg_size = `total_post_neg_size' / `successtrim_neg';
};
di "observations with p < .05 total without trimming: " `success';

di "observations with p < .05, effect negative without trimming: "
`success_neg';

di "observations with p < .05, effect positive without trimming: "
`success_pos';

di "observations with p < .05 total with trimming " `successtrim';

di "observations with p < .05, effect negative with trimming: "
`successtrim_neg';

di "observations with p < .05, effect positive with trimming: "
`successtrim_pos';

/* The average coefficient size is expected to be zero. This only shows
the average coefficient size for the trials for which statistically
```

```
significant results were not observed pre-trimming but were observed
post-trimming. This is done so that the difference in coefficient size
for the coefficients that are altered can be calculated. */

di "pre-trimming average positive effect size: " `total_pre_pos_size';
di "post-trimming average positive effect size: " `total_post_pos_size';
di "pre-trimming average negative effect size: " `total_pre_neg_size';
di "post-trimming average negative effect size: " `total_post_neg_size';

di "success on trim 1: " `success1';
di "success on trim 2: " `success2';
di "success on trim 3: " `success3';
di "success on trim 4: " `success4';
di "success on trim 5: " `success5';

break;
```

APPENDIX B

CODE FOR SIMULATIONS USING OUTLIER OMISSION WITH Z-SCORE THRESHOLDS

```
#delimit ;
set more off;
clear;
set seed 3;

// make variables

generate treatment = 0;
generate double baseline = .;
generate double trimmed = .;
generate double z_score = .;
local success = 0;
local success_pos = 0;
local success_neg = 0;
local successtrim = 0;
local successtrim_pos = 0;
local successtrim_neg = 0;
local total_pre_pos_size = 0;
local total_pre_neg_size = 0;
local total_post_pos_size = 0;
local total_post_neg_size = 0;
local post_effect_neg = 0;
local post_effect_pos = 0;
local pre_effect_pos = 0;
local pre_effect_neg = 0;
local percentile = 0;
local percentile_cutoff = .05;
local count = 0;
local count_percent = 0;
local treatment_mean = 0;
local treatment_sd = 0;
local success2 = 0;
local success25 = 0;
local success3 = 0;
local iterations= 20000;

//This is the sample size. It is changed for different simulations.

local n = 100;

while `iterations' > 0{;

       local count = 1;
       local trims = 1;
       local trimcount = 1;

       local i = `n' * 2;
       set obs `i';

       /*This generates the dataset. It creates n "treatment" variables and
```

```
n "control" variables taken from the specified distribution. The
distribution that is being used is the only one that is not preceded
by a line comment, "//". For logistic distributions, the third and forth
lines represent a calculation of a random number sampled from a logistic
distribution and the fifth line inserts that random number into the
dataset. */

while `i' > 0{;
	replace baseline = rnormal(0, 1) in `i';
	//replace baseline = rchi2(4) in `i';
	//local unif = runiform();
	//local logi = -ln((1 - `unif')/`unif');
	//replace baseline = `logi' in `i';
	replace treatment = 1 in `i' if `i' <= `n';
	replace treatment = 0 in `i' if `i' > `n';
	local i = `i' - 1;
};

local i = `n';

while `i' > 0{;
	replace trimmed = baseline in `i';
	local i = `i' - 1;
};


local i = `n' * 2;
summarize trimmed;

while `i' > 0{;
	local treatment_mean = baseline in `i';

	replace z_score = (`treatment_mean' - r(mean))/r(sd)
	in `i' if `i' <= `n';

	local i = `i' - 1;
};

regress baseline treatment;

local p = 2*ttail(e(df_r),abs(_b[treatment]/_se[treatment]));

//if p is statistically significant, then stop there. Otherwise,
//try the trimming routine

if (`p' <= .05){;

	//log whether the observed coefficient is positive or negative

	if (_b[treatment] > 0){;
		local success_pos = `success_pos' + 1;
	};
	else{;
		local success_neg = `success_neg' + 1;
	};
	local success = `success' + 1;
};

if (`p' > .05){;
```

```
/*log the insignificant effect size. This can be compared to
the significant one if the significance threshold is obtained */

if (_b[treatment] > 0){;
       local pre_effect_pos = _b[treatment];
};
else{;
       local pre_effect_neg = _b[treatment];
};
replace trimmed = baseline;
replace trimmed = . if abs(z_score) > 3 & z_score != .;
regress trimmed treatment;
local p = 2*ttail(e(df_r),abs(_b[treatment]/_se[treatment]));
if (`p' <= .05){;
       local success3 = `success3' + 1;
       if (_b[treatment] > 0){;
                     local post_effect_pos = _b[treatment];

                     local total_post_pos_size =
                     `total_post_pos_size' + 0 +
                     `post_effect_pos';

                     local total_pre_pos_size =
                     `total_pre_pos_size' + 0 +
                     `pre_effect_pos';

                     local successtrim_pos =
                     `successtrim_pos' + 1;

              };
              else{;

                     local post_effect_neg = _b[treatment];

                     local total_post_neg_size =
                     `total_post_neg_size' + 0 +
                     `post_effect_neg';

                     local total_pre_neg_size =
                     `total_pre_neg_size' + 0 +
                     `pre_effect_neg';

                     local successtrim_neg =
                     `successtrim_neg' + 1;

              };

};
if (`p' > .05){;
       replace trimmed = baseline;

       replace trimmed = . if abs(z_score) > 2.5 &
       z_score != .;

       regress trimmed treatment;

       local p = 2*ttail(e(df_r),
       abs(_b[treatment]/_se[treatment]));
```

```
if (`p' <= .05){;
       local success25 = `success25' + 1;
       if (_b[treatment] > 0){;
              local post_effect_pos = _b[treatment];

              local total_post_pos_size =
              `total_post_pos_size' + 0 +
              `post_effect_pos';

              local total_pre_pos_size =
              `total_pre_pos_size' + 0 +
              `pre_effect_pos';

              local successtrim_pos =
              `successtrim_pos' + 1;

       };
       else{;

              local post_effect_neg = _b[treatment];

              local total_post_neg_size =
              `total_post_neg_size' + 0 +
              `post_effect_neg';

              local total_pre_neg_size =
              `total_pre_neg_size' + 0 +
              `pre_effect_neg';

              local successtrim_neg =
              `successtrim_neg' + 1;
       };
};

if (`p' > .05){;
       replace trimmed = baseline;

       replace trimmed = . if abs(z_score) > 2
       & z_score != .;

       regress trimmed treatment;

       local p = 2*ttail(e(df_r),
       abs(_b[treatment]/_se[treatment]));

       if (`p' <= .05){;
              local success2 = `success2' + 1;
              if (_b[treatment] > 0){;

                     local post_effect_pos =
                     _b[treatment];

                     local total_post_pos_size =
                     `total_post_pos_size' + 0 +
                     `post_effect_pos';

                     local total_pre_pos_size =
                     `total_pre_pos_size' + 0 +
```

```
                                        `pre_effect_pos';

                                        local successtrim_pos =
                                        `successtrim_pos' + 1;

                                    };
                                    else{;

                                        local post_effect_neg =
                                        _b[treatment];

                                        local total_post_neg_size =
                                        `total_post_neg_size' + 0 +
                                         `post_effect_neg';

                                        local total_pre_neg_size =
                                        `total_pre_neg_size' + 0 +
                                        `pre_effect_neg';

                                        local successtrim_neg =
                                        `successtrim_neg' + 1;

                                    };
                                };
                            };
                        };

        };
        local iterations = `iterations' - 1;
        di "iteration " `iterations';
};
if (`successtrim_pos' > 0){;
        local total_pre_pos_size = `total_pre_pos_size' / `successtrim_pos';
        local total_post_pos_size = `total_post_pos_size' / `successtrim_pos';
};
if (`successtrim_neg' > 0){;
        local total_pre_neg_size = `total_pre_neg_size' / `successtrim_neg';
        local total_post_neg_size = `total_post_neg_size' / `successtrim_neg';
};

di "observations with p < .05 total without trimming: " `success';

di "observations with p < .05, effect negative without trimming: "
`success_neg';

di "observations with p < .05, effect positive without trimming: "
`success_pos';

di "observations with p < .05 total with trimming " `successtrim';

di "observations with p < .05, effect negative with trimming: "
`successtrim_neg';

di "observations with p < .05, effect positive with trimming: "
`successtrim_pos';

/*The average coefficient size is expected to be zero. This only shows the
average coefficient size for the trials for which statistically significant
results were not observed pre-trimming but were observed post-trimming. This is
```

```
done so that the difference in coefficient size for the coefficients that are
altered can be calculated.
*/

di "pre-trimming average positive effect size: " `total_pre_pos_size';
di "post-trimming average positive effect size: " `total_post_pos_size';
di "pre-trimming average negative effect size: " `total_pre_neg_size';
di "post-trimming average negative effect size: " `total_post_neg_size';

di "success when trimming rule is z > 3: " `success3';
di "success wen trmming rule is z > 2.5: " `success25';
di "success when trimming rule is z > 2: " `success2';

break;
```

APPENDIX C

FIGURES

Figures 1 through 8 show the dependency of type I error rates on the maximum one-tailed trimming threshold used, following the iterative data exclusion process described in Section 3, for various sample sizes. Figures 1 through 4 show the results for simulations wherein data is only excluded from the left tail of analyzed samples. Figures 5 through 8 show results from simulations wherein data is excluded from the right tail of analyzed samples.

Figures 8 through 12 show the dependency of type I error rates on the z-score trimming threshold used, following the iterative process described in Section 3, for various sample sizes.

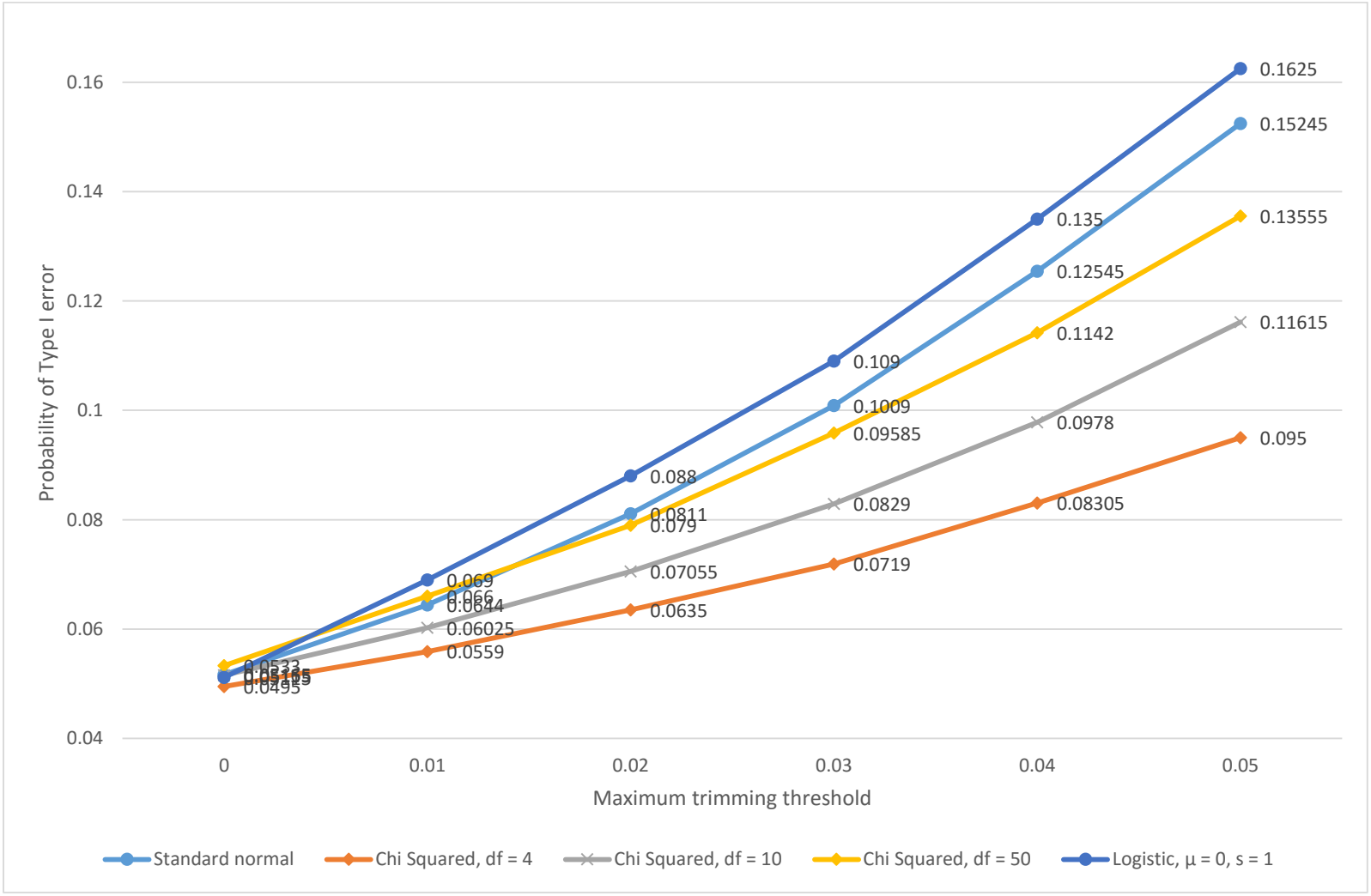

Figure 1: Cumulative probability of type 1 error, n = 100, left tail trimming

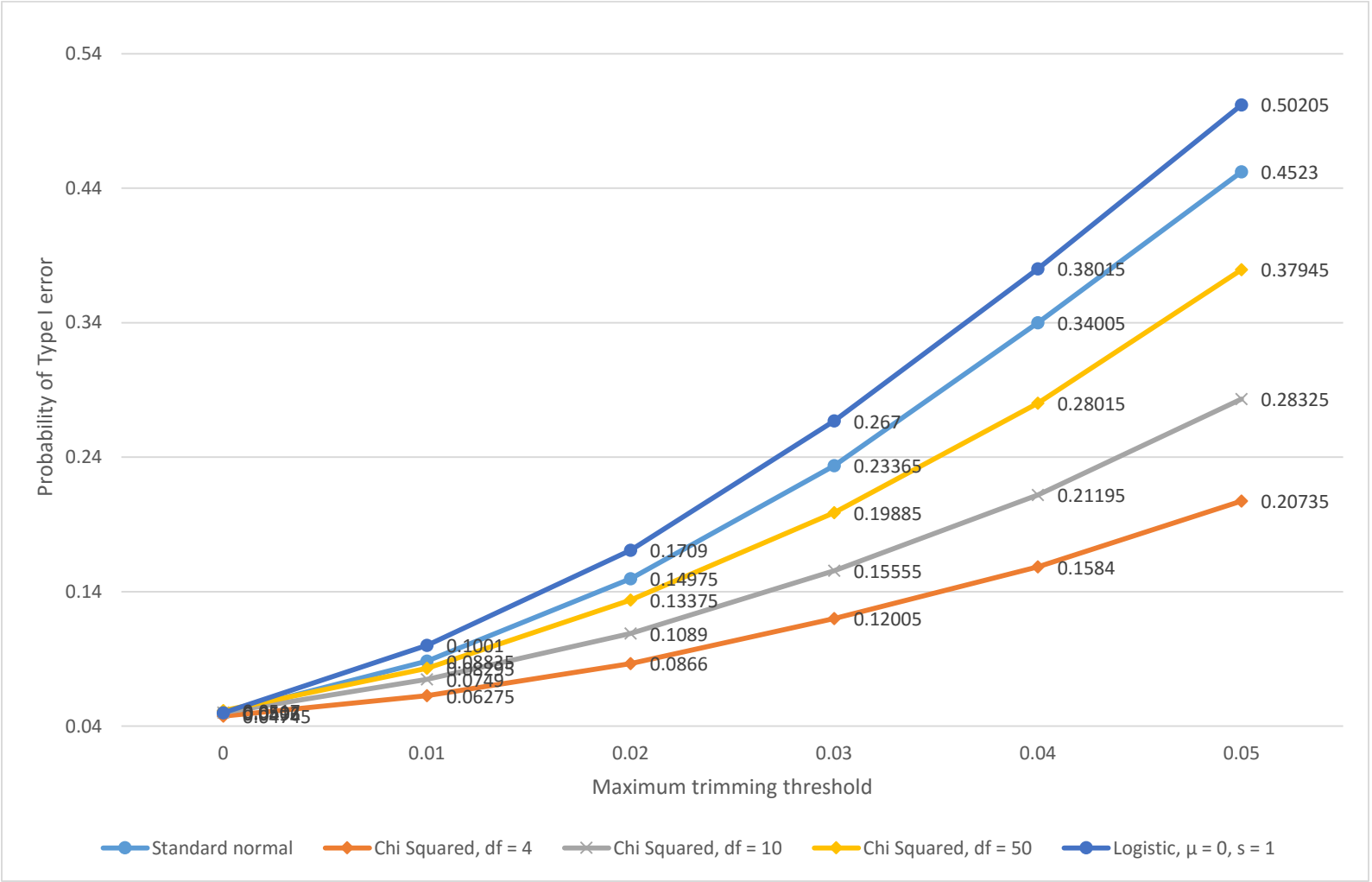

Figure 2: Cumulative probability of type 1 error, n = 500, left tail trimming

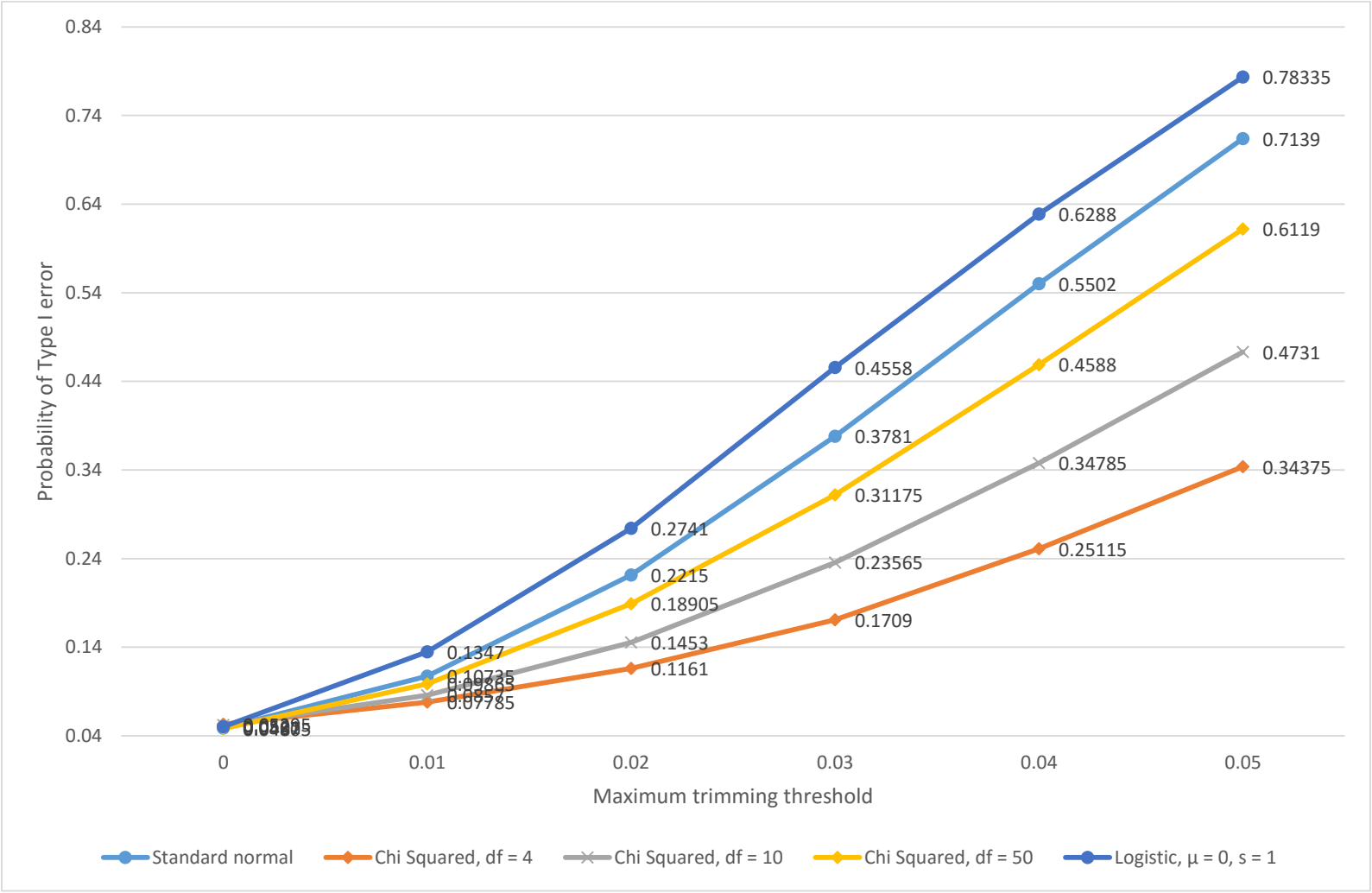

Figure 3: Cumulative probability of type 1 error, n = 1,000, left tail trimming

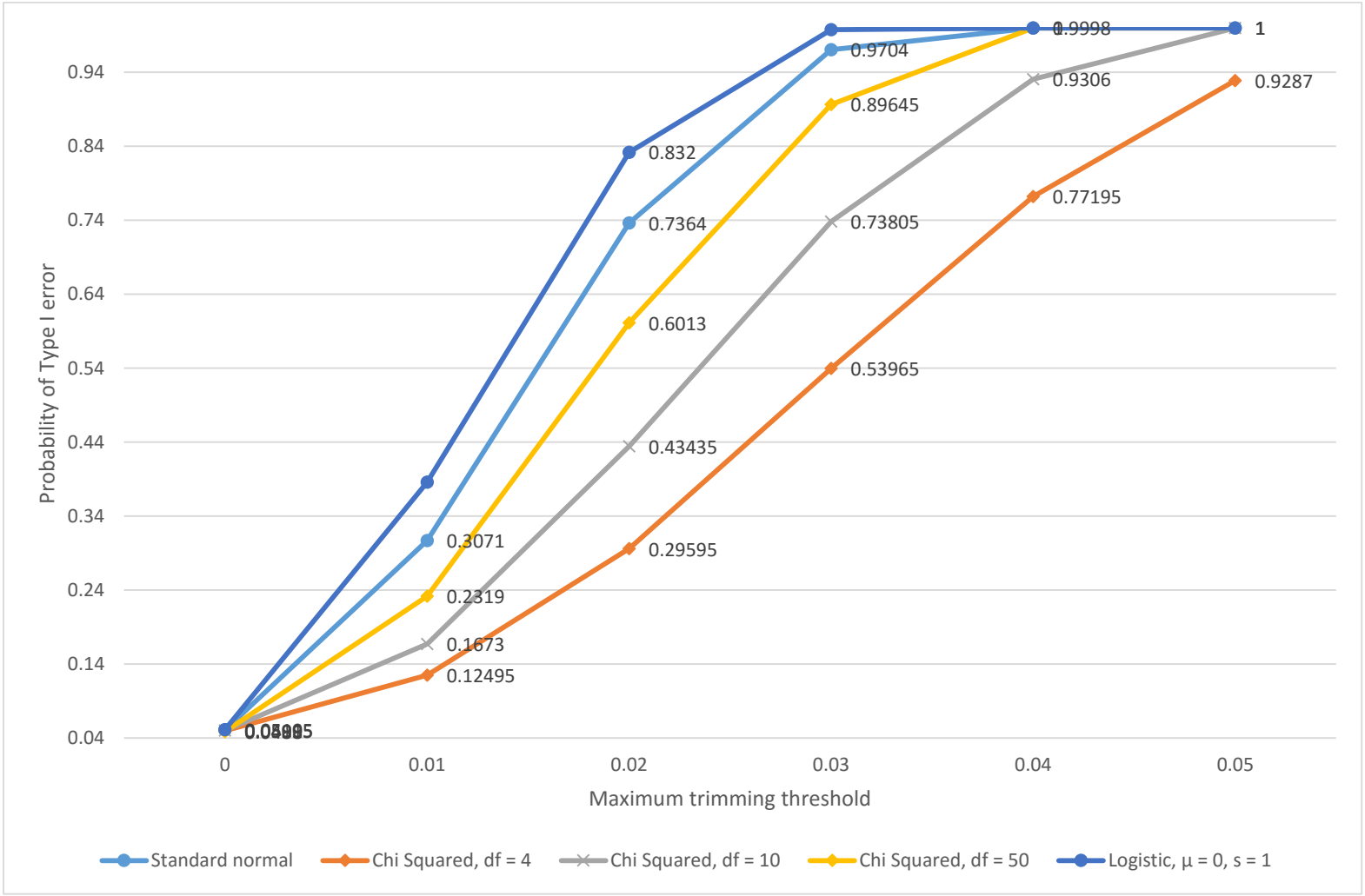

Figure 4: Cumulative probability of type 1 error, n = 5,000, left tail trimming

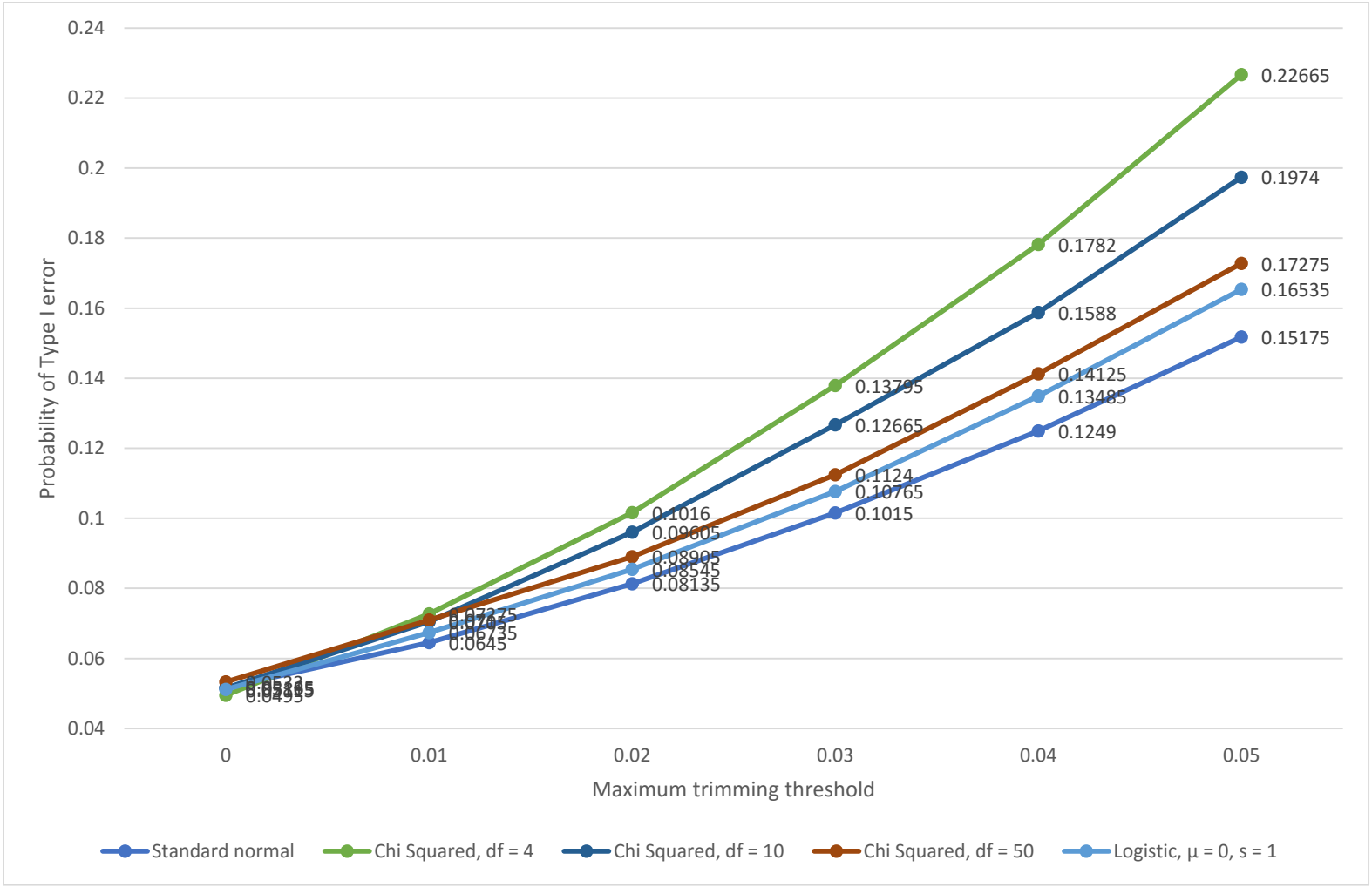

Figure 5: Cumulative probability of type 1 error, n = 100, right tail trimming

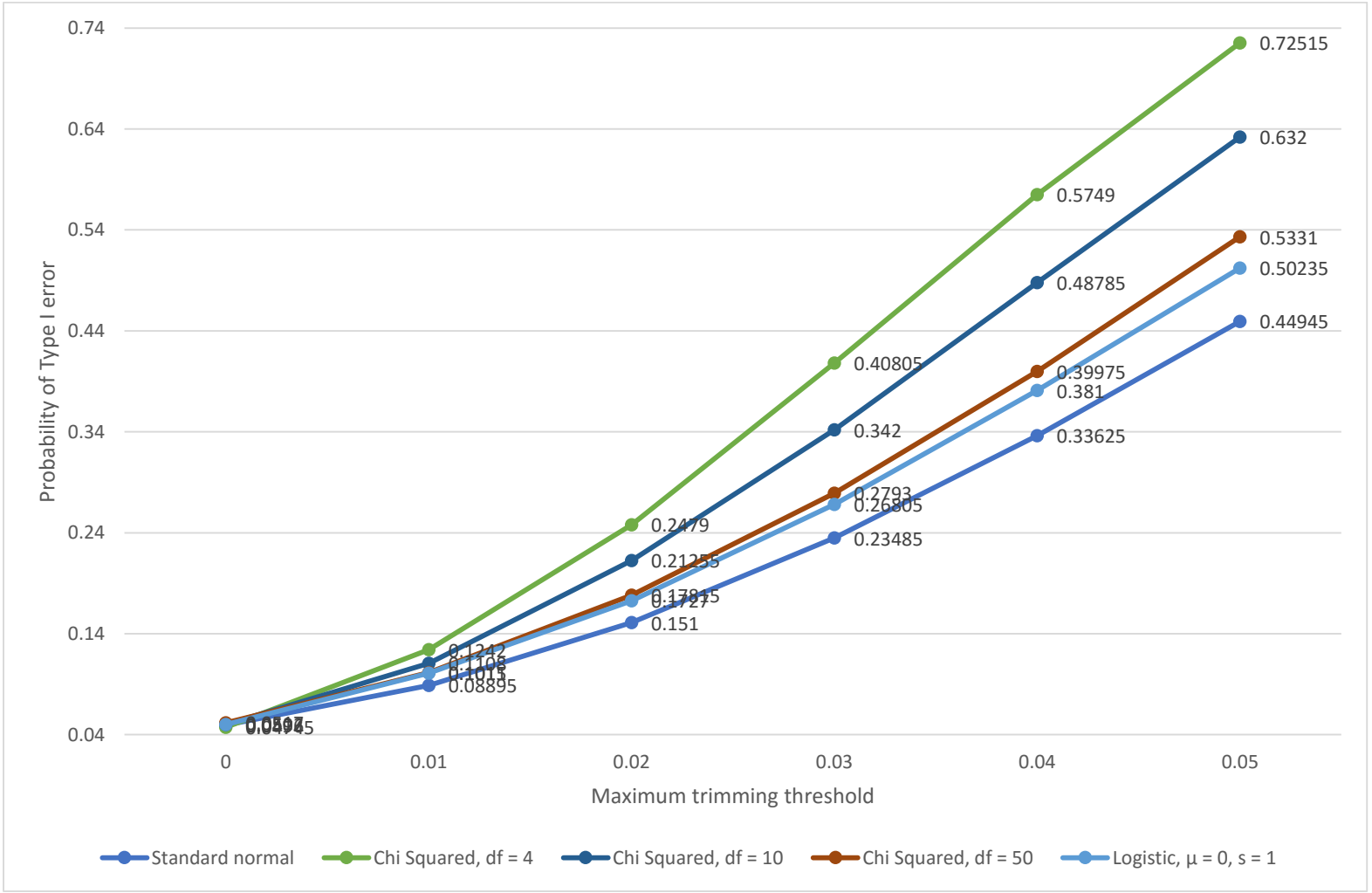

Figure 6: Cumulative probability of type 1 error, n = 500, right tail trimming

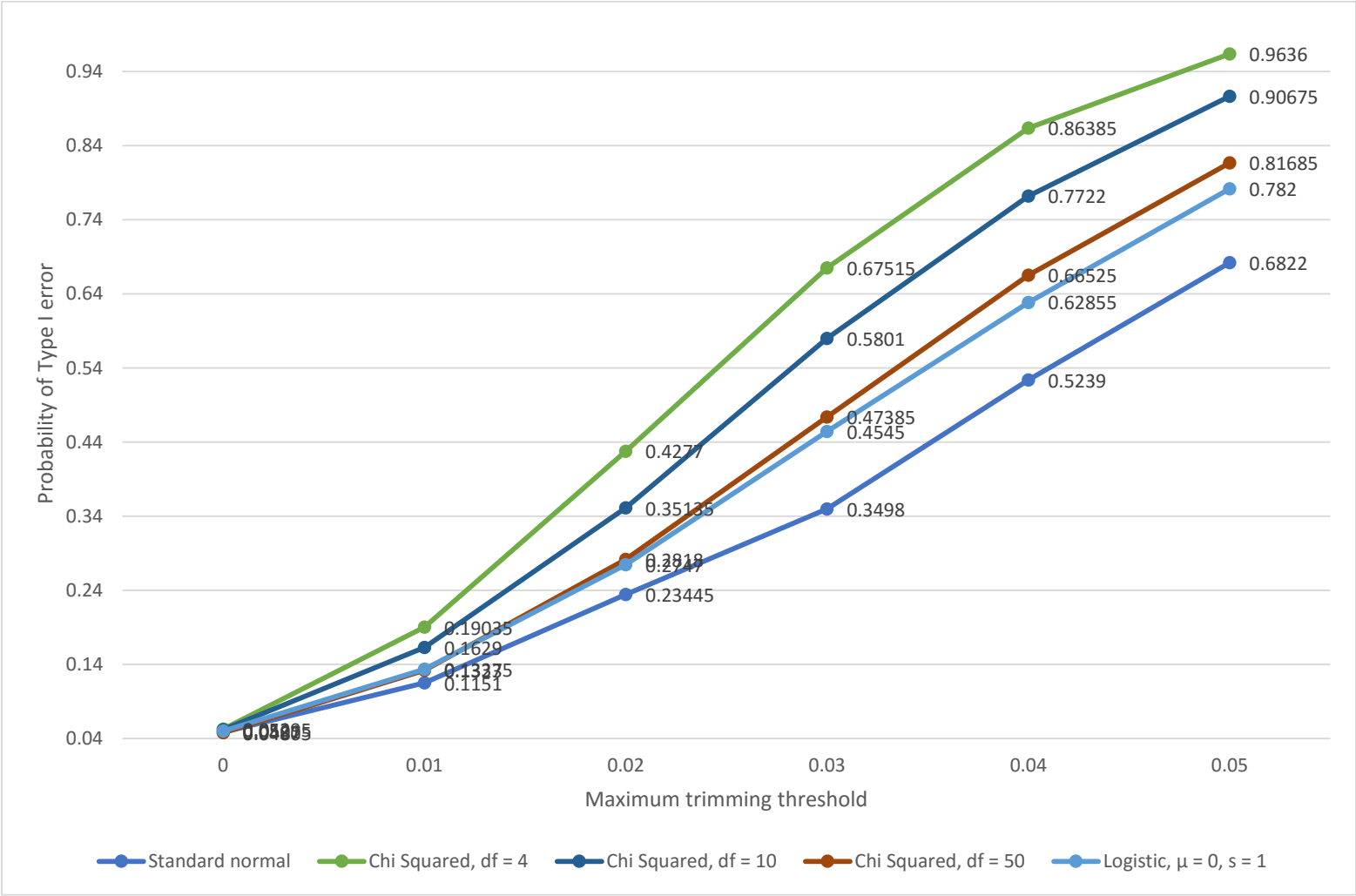

Figure 7: Cumulative probability of type 1 error, n = 1,000, right tail trimming

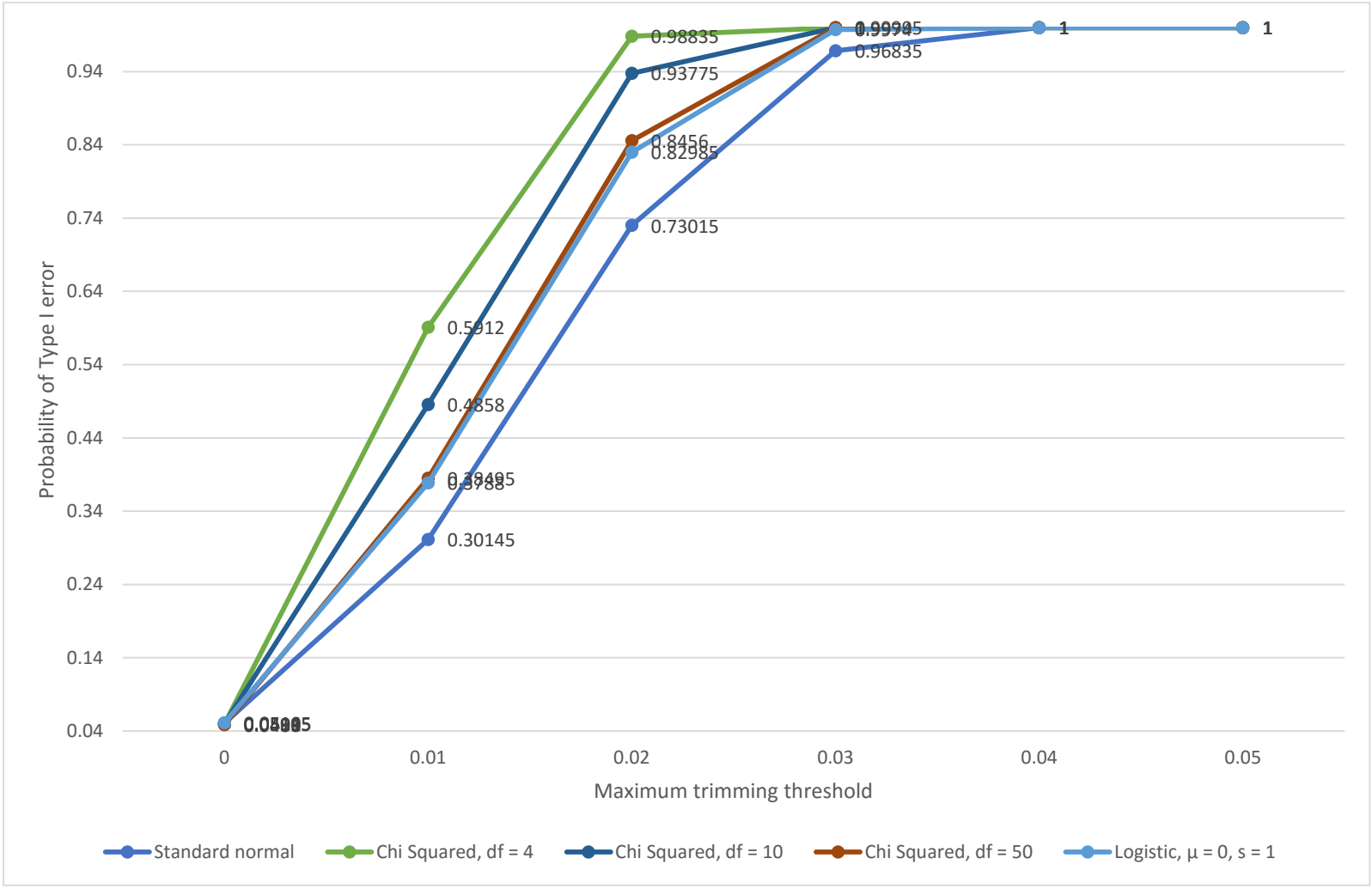

Figure 8: Cumulative probability of type 1 error, n = 5,000, right tail trimming

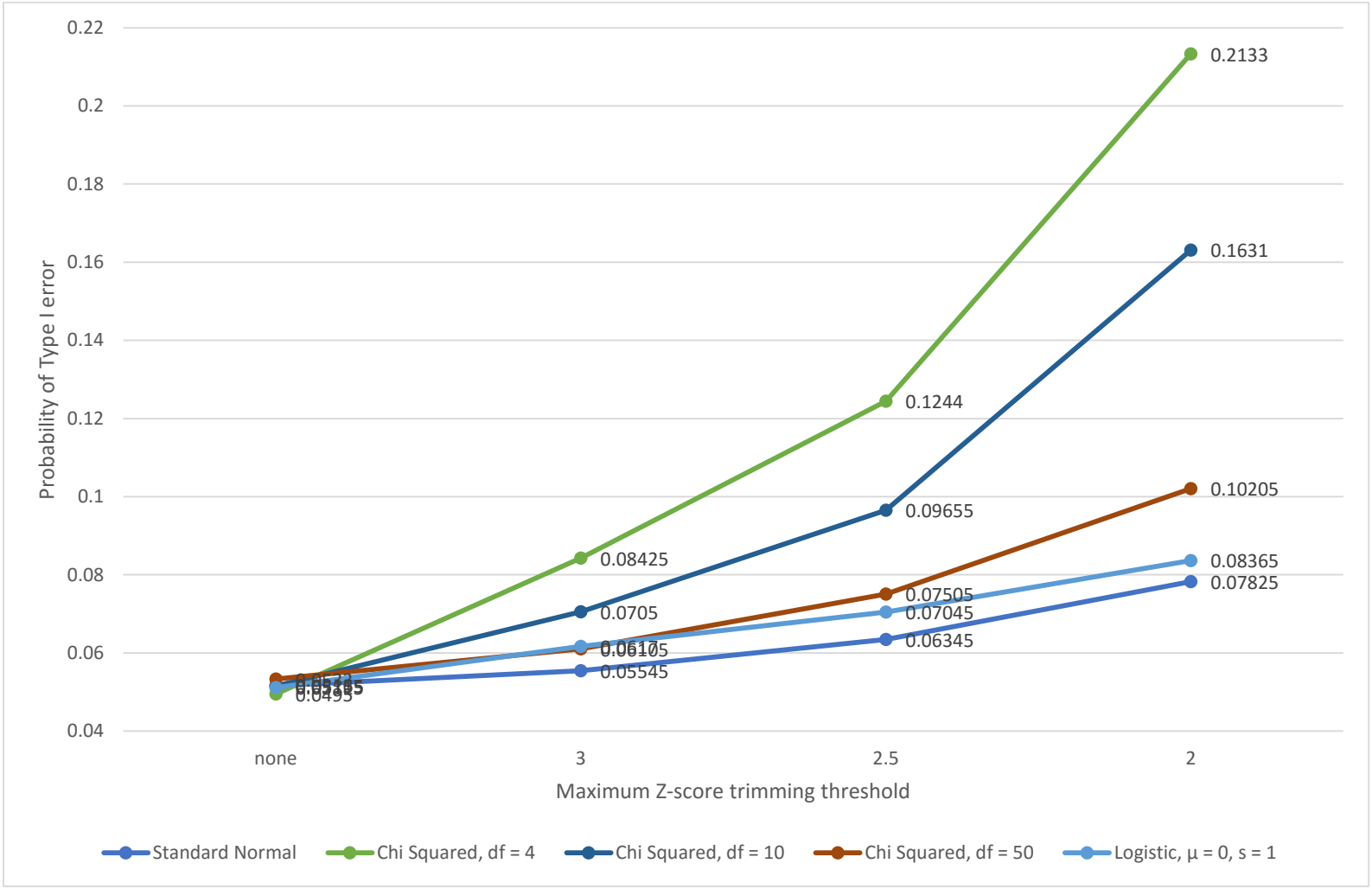

Figure 9: Cumulative probability of type 1 error, n = 100, z-score threshold trimming

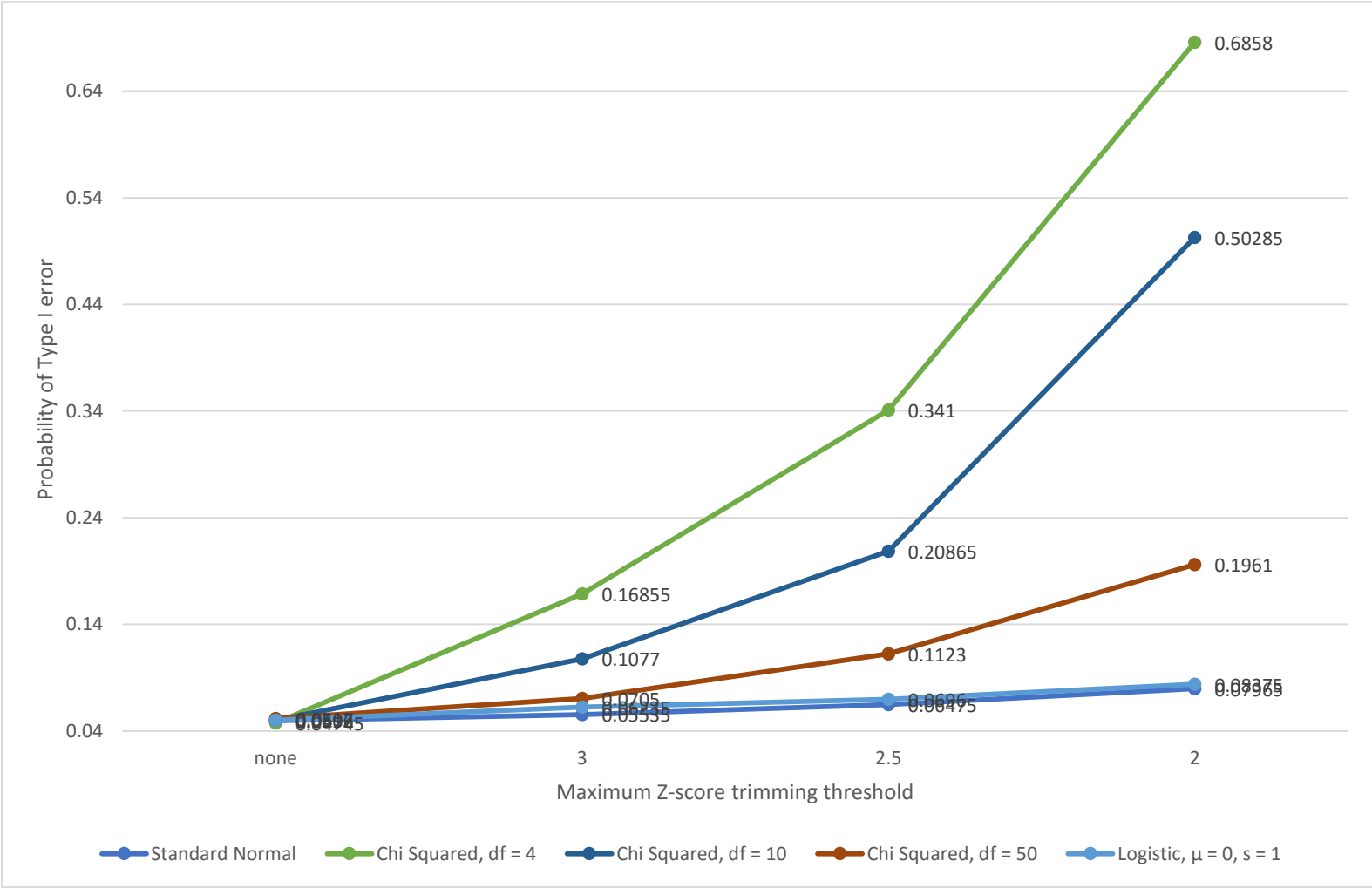

Figure 10: Cumulative probability of type 1 error, n = 500, z-score threshold trimming

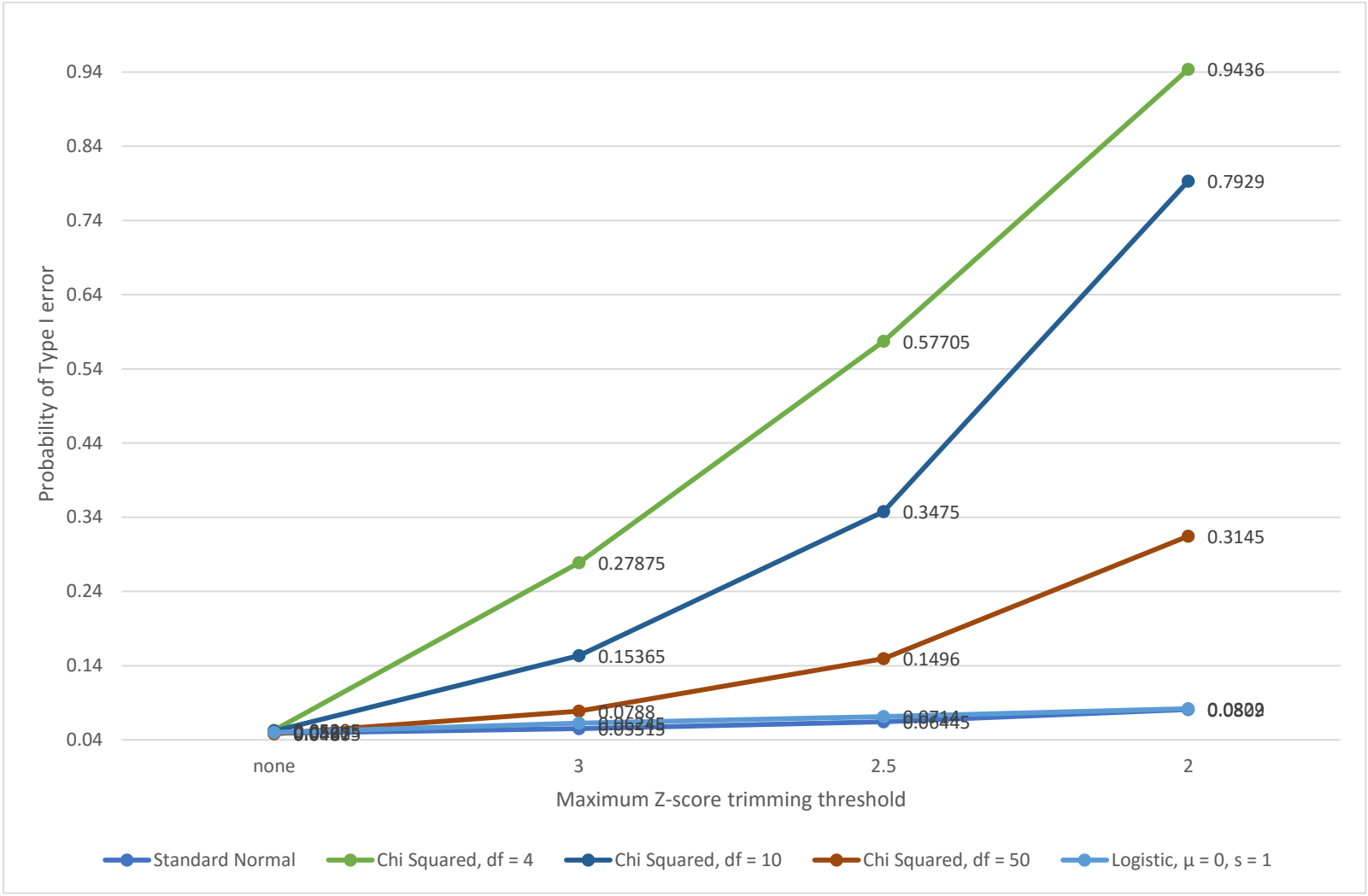

Figure 11: Cumulative probability of type 1 error, n = 1000, z-score threshold trimming

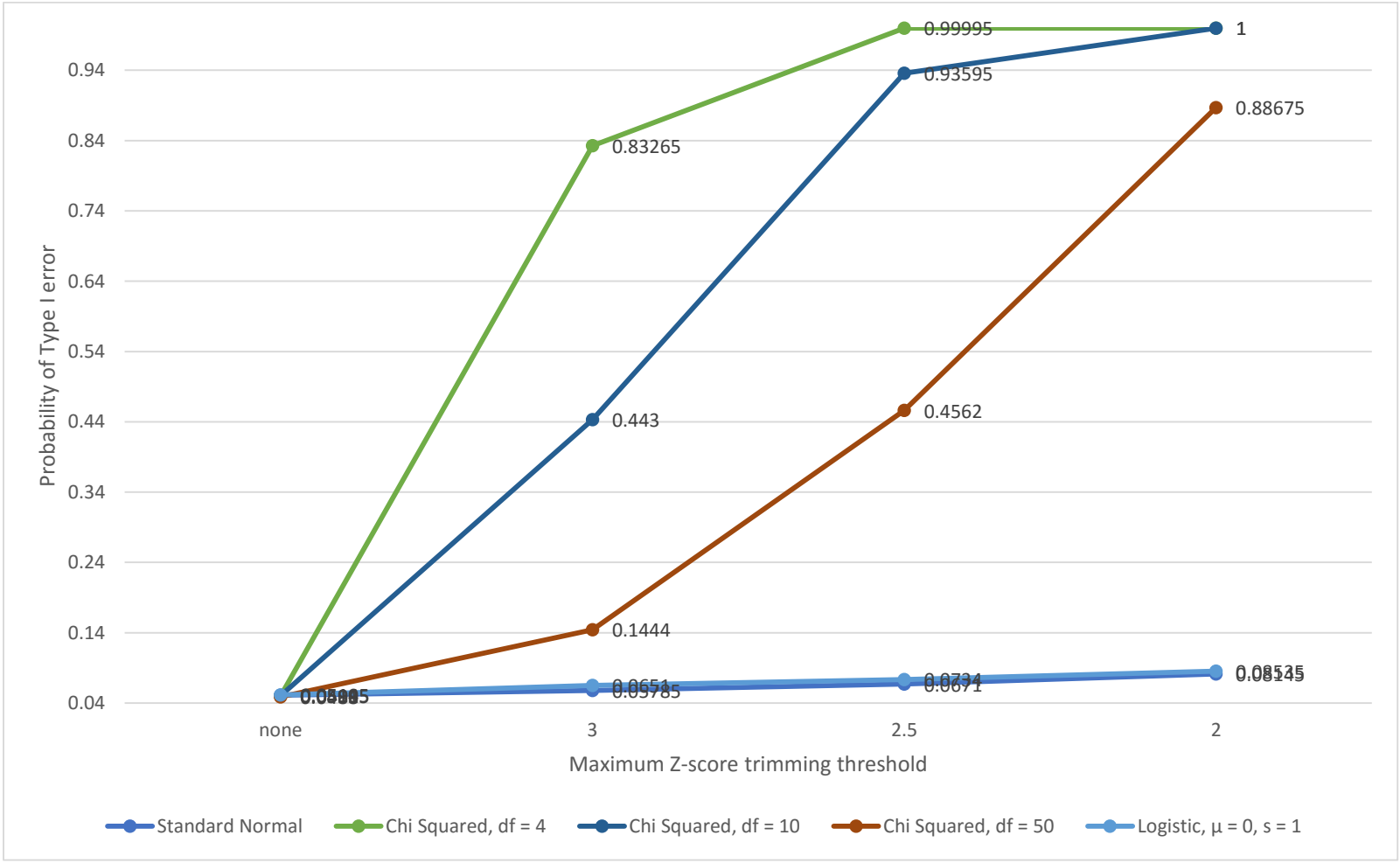

Figure 12: Cumulative probability of type 1 error, n = 5000, z-score threshold trimming

APPENDIX D

TABLES

TABLE 1: Type I error rates for z-score threshold trimming

n	Population Distribution	Z-score Trimming Threshold		
		3	2.5	2
100	**Standard Normal**	0.05545	0.06345	0.07825
100	**χ^2, df = 4**	0.08425	0.1244	0.2133
100	**χ^2, df = 10**	0.0705	0.09655	0.1631
100	**χ^2, df = 50**	0.06105	0.07505	0.10205
100	**Logistic, $\mu = 0, s = 1$**	0.0617	0.07045	0.08365
500	**Standard Normal**	0.05535	0.06475	0.07965
500	**χ^2, df = 4**	0.16855	0.341	0.6858
500	**χ^2, df = 10**	0.1077	0.20865	0.50285
500	**χ^2, df = 50**	0.0705	0.1123	0.1961
500	**Logistic, $\mu = 0, s = 1$**	0.06235	0.0696	0.08375
1000	**Standard Normal**	0.05515	0.06445	0.0809
1000	**χ^2, df = 4**	0.27875	0.57705	0.9436
1000	**χ^2, df = 10**	0.15365	0.3475	0.7929
1000	**χ^2, df = 50**	0.0788	0.1496	0.3145
1000	**Logistic, $\mu = 0, s = 1$**	0.06245	0.0714	0.0822
5000	**Standard Normal**	0.05785	0.0671	0.08145
5000	**χ^2, df = 4**	0.83265	0.99995	1
5000	**χ^2, df = 10**	0.443	0.93595	1
5000	**χ^2, df = 50**	0.1444	0.4562	0.88675
5000	**Logistic, $\mu = 0, s = 1$**	0.0651	0.0734	0.08535

Table 1: type I error rates at 5% significance threshold resulting from selective trimming based on Z-scores exclusion as described in section 3.

TABLE 2: Type I error rates for left-tailed trimming

n	Population Distribution	Percentile Trimming Threshold				
		1%	2%	3%	4%	5%
100	**Standard Normal**	0.0644	0.0811	0.1009	0.12545	0.15245
100	**χ^2, df = 4**	0.0559	0.0635	0.0719	0.08305	0.095
100	**χ^2, df = 10**	0.06025	0.07055	0.0829	0.0978	0.11615
100	**χ^2, df = 50**	0.066	0.079	0.09585	0.1142	0.13555
100	**Logistic, $\mu = 0, s = 1$**	0.069	0.088	0.109	0.135	0.1625
500	**Standard Normal**	0.08835	0.14975	0.23365	0.34005	0.4523
500	**χ^2, df = 4**	0.06275	0.0866	0.12005	0.1584	0.20735
500	**χ^2, df = 10**	0.0749	0.1089	0.15555	0.21195	0.28325
500	**χ^2, df = 50**	0.08295	0.13375	0.19885	0.28015	0.37945
500	**Logistic, $\mu = 0, s = 1$**	0.1001	0.1709	0.267	0.38015	0.50205
1000	**Standard Normal**	0.10735	0.2215	0.3781	0.5502	0.7139
1000	**χ^2, df = 4**	0.07785	0.1161	0.1709	0.25115	0.34375
1000	**χ^2, df = 10**	0.0857	0.1453	0.23565	0.34785	0.4731
1000	**χ^2, df = 50**	0.09865	0.18905	0.31175	0.4588	0.6119
1000	**Logistic, $\mu = 0, s = 1$**	0.1347	0.2741	0.4558	0.6288	0.78335
5000	**Standard Normal**	0.3071	0.7364	0.9704	1	1
5000	**χ^2, df = 4**	0.12495	0.29595	0.53965	0.77195	0.9287
5000	**χ^2, df = 10**	0.1673	0.43435	0.73805	0.9306	1
5000	**χ^2, df = 50**	0.2319	0.6013	0.89645	0.9998	1
5000	**Logistic, $\mu = 0, s = 1$**	0.3859	0.832	0.9975	1	1

Table 2: type I error rates at 5% significance threshold resulting from selective trimming, excluding only the left tail of the sample as described in section 3.

TABLE 3: Type I error rates for right-tailed trimming

n	Population Distribution	Percentile Trimming Threshold				
		1%	2%	3%	4%	5%
100	**Standard Normal**	0.0645	0.08135	0.1015	0.1249	0.15175
100	**χ^2, df = 4**	0.07275	0.1016	0.13795	0.1782	0.22665
100	**χ^2, df = 10**	0.0705	0.09605	0.12665	0.1588	0.1974
100	**χ^2, df = 50**	0.071	0.08905	0.1124	0.14125	0.17275
100	**Logistic, $\mu = 0, s = 1$**	0.06735	0.08545	0.10765	0.13485	0.16535
500	**Standard Normal**	0.08895	0.151	0.23485	0.33625	0.44945
500	**χ^2, df = 4**	0.1242	0.2479	0.40805	0.5749	0.72515
500	**χ^2, df = 10**	0.1108	0.21255	0.342	0.48785	0.632
500	**χ^2, df = 50**	0.1015	0.17815	0.2793	0.39975	0.5331
500	**Logistic, $\mu = 0, s = 1$**	0.1011	0.1727	0.26805	0.381	0.50235
1000	**Standard Normal**	0.1151	0.23445	0.3498	0.5239	0.6822
1000	**χ^2, df = 4**	0.19035	0.4277	0.67515	0.86385	0.9636
1000	**χ^2, df = 10**	0.1629	0.35135	0.5801	0.7722	0.90675
1000	**χ^2, df = 50**	0.1323	0.2818	0.47385	0.66525	0.81685
1000	**Logistic, $\mu = 0, s = 1$**	0.13375	0.2747	0.4545	0.62855	0.782
5000	**Standard Normal**	0.30145	0.73015	0.96835	1	1
5000	**χ^2, df = 4**	0.5912	0.98835	1	1	1
5000	**χ^2, df = 10**	0.4858	0.93775	1	1	1
5000	**χ^2, df = 50**	0.38495	0.8456	0.99995	1	1
5000	**Logistic, $\mu = 0, s = 1$**	0.3788	0.82985	0.9974	1	1

Table 3: type I error rates at 5% significance threshold resulting from selective trimming, excluding only the right tail of the sample as described in section 3.

TABLE 4: Coefficient size differences for z-score threshold trimming

n	Population Distribution	Difference in coefficient size when coefficient is positive		Difference in coefficient size when coefficient is negative	
		Coefficient Difference	Difference in standard deviations	Coefficient Difference	Difference in standard deviations
100	**Standard Normal**	0.0491497	0.0491497	-0.05400332	-0.05400332
100	**χ^2, df = 4**			-0.35873345	-0.12683142
100	**χ^2, df = 10**			-0.4814947	-0.10766548
100	**χ^2, df = 50**	0.2523274	0.02523274	-0.7600572	-0.07600572
100	**Logistic, μ = 0, s = 1**	0.09110917	0.050231118	-0.092821	-0.05117489
500	**Standard Normal**	0.02198541	0.02198541	-0.01898185	-0.01898185
500	**χ^2, df = 4**			-0.26623059	-0.09412672
500	**χ^2, df = 10**			-0.4054659	-0.09066493
500	**χ^2, df = 50**	0.0388016	0.00388016	-0.54777162	-0.05477716
500	**Logistic, μ = 0, s = 1**	0.04289969	0.023651839	-0.03851046	-0.02123192
1000	**Standard Normal**	0.01472989	0.01472989	-0.01379454	-0.01379454
1000	**χ^2, df = 4**			-0.20922835	-0.07397339
1000	**χ^2, df = 10**			-0.35669386	-0.07975917
1000	**χ^2, df = 50**	0.02659144	0.002659144	-0.5170907	-0.05170907
1000	**Logistic, μ = 0, s = 1**	0.02817934	0.015536084	-0.02716737	-0.01497815
5000	**Standard Normal**	0.00635428	0.00635428	-0.00672346	-0.00672346
5000	**χ^2, df = 4**			-0.12980737	-0.04589383
5000	**χ^2, df = 10**			-0.18213386	-0.04072636
5000	**χ^2, df = 50**			-0.3495587	-0.03495587
5000	**Logistic, μ = 0, s = 1**	0.01201063	0.006621807	-0.0122375	-0.00674688

Table 4: Differences in coefficient size, pre- and post-trimming for samples trimmed based on z-score thresholds. Any time that excluding data changed an insignificant result to a significant one at the 5% significance threshold, the difference in coefficient size before trimming and after trimming was logged. This reports the averages of these differences. Blank cells indicate cases where trimming never resulted in a positive coefficient.

TABLE 5: Coefficient size differences for left-tailed trimming

n	Population Distribution	Difference in coefficient size when coefficient is positive		Difference in coefficient size when coefficient is negative	
		Coefficient Difference	Difference in standard deviations	Coefficient Difference	Difference in standard deviations
100	**Standard Normal**	0.07523123	0.07523123		
100	**χ^2, df = 4**	0.13344476	0.047179847		
100	**χ^2, df = 10**	0.2588158	0.057872972		
100	**χ^2, df = 50**	0.6626796	0.06626796		
100	**Logistic, μ = 0, s = 1**	0.14183847			
500	**Standard Normal**	0.0785607	0.0785607		
500	**χ^2, df = 4**	0.13484771	0.047675865		
500	**χ^2, df = 10**	0.26150791	0.058474946		
500	**χ^2, df = 50**	0.69996186	0.069996186		
500	**Logistic, μ = 0, s = 1**	0.1489591	0.082125456		
1000	**Standard Normal**	0.06281515	0.06281515		
1000	**χ^2, df = 4**	0.13688872	0.048397471		
1000	**χ^2, df = 10**	0.26453965	0.059152864		
1000	**χ^2, df = 50**	0.63108952	0.063108952		
1000	**Logistic, μ = 0, s = 1**	0.11399361	0.062847971		
5000	**Standard Normal**	0.03543871	0.03543871		
5000	**χ^2, df = 4**	0.08844666	0.031270617		
5000	**χ^2, df = 10**	0.14816295	0.033130243		
5000	**χ^2, df = 50**	0.34285365	0.034285365		
5000	**Logistic, μ = 0, s = 1**	0.06567853	0.036210471		

Table 5: Differences in coefficient size, pre- and post-trimming for samples with trimmed left tails based on percentile thresholds. Any time that excluding data changed an insignificant result to a significant one at the 5% significance threshold, the difference in coefficient size before trimming and after trimming was logged. This reports the averages of these differences. Blank cells indicate cases where trimming never resulted in a negative coefficient.

TABLE 6: Coefficient size differences for right-tailed trimming

n	Population Distribution	Difference in coefficient size when coefficient is positive		Difference in coefficient size when coefficient is negative	
		Coefficient Difference	Difference in standard deviations	Coefficient Difference	Difference in standard deviations
100	**Standard Normal**			-0.07534788	-0.07534788
100	**χ^2, df = 4**			-0.2835173	-0.10023850
100	**χ^2, df = 10**			-0.4092423	-0.09150936
100	**χ^2, df = 50**			-0.8240382	-0.08240382
100	**Logistic, μ = 0, s = 1**			-0.14517944	-0.08004162
500	**Standard Normal**			-0.07845424	-0.07845424
500	**χ^2, df = 4**			-0.236644	-0.08366628
500	**χ^2, df = 10**			-0.37876655	-0.08469477
500	**χ^2, df = 50**			-0.84304128	-0.08430412
500	**Logistic, μ = 0, s = 1**			-0.14844211	-0.08184042
1000	**Standard Normal**			-0.0631318	-0.0631318
1000	**χ^2, df = 4**			-0.18155345	-0.06418883
1000	**χ^2, df = 10**			-0.28373422	-0.0634449
1000	**χ^2, df = 50**			-0.63694708	-0.06369470
1000	**Logistic, μ = 0, s = 1**			-0.11456275	-0.06316175
5000	**Standard Normal**			-0.03552508	-0.03552508
5000	**χ^2, df = 4**			-0.1136682	-0.04018777
5000	**χ^2, df = 10**			-0.17217563	-0.03849964
5000	**χ^2, df = 50**			-0.36642148	-0.03664214
5000	**Logistic, μ = 0, s = 1**			-0.06601046	-0.03639347

Table 6: Differences in coefficient size, pre- and post-trimming for samples with trimmed right tails based on percentile thresholds. Any time that excluding data changed an insignificant result to a significant one at the 5% significance threshold, the difference in coefficient size before trimming and after trimming was logged. This reports the averages of these differences. Blank cells indicate cases where trimming never resulted in a positive coefficient.