s:fi
RESEARCH

Swiss Finance Institute
Research Paper Series
N°17-37

Anomalies and False Rejections

Tarun CHORDIA
Emory University

Amit GOYAL
University of Lausanne and Swiss Finance Institute

Alessio SARETTO
University of Texas at Dallas

Anomalies and False Rejections

Tarun Chordia **Amit Goyal** **Alessio Saretto***

September 2019

Abstract

We use information from over two million trading strategies that are randomly generated using real data, and from strategies that survive the publication process to infer the statistical properties of the set of strategies that could have been studied by researchers. Using this set, we compute t-statistic thresholds that control for multiple hypothesis testing when searching for anomalies, at 3.84 and 3.38 for time-series and cross-sectional regressions, respectively. We estimate the expected proportion of false rejections that researchers would produce if they failed to account for multiple hypothesis testing to be 45.3%.

*Tarun Chordia is from Emory University, Amit Goyal is from the Swiss Finance Institute at the University of Lausanne, and Alessio Saretto is from University of Texas at Dallas. Part of this research was done while Alessio Saretto was visiting the Olin Business School at the Washington University in St. Louis. We thank Hank Bessembinder, Svetlana Bryzgalova, Ing-Haw Cheng, Jason Donaldson, Baoqing Gan, Ruslan Goyenko, Campbell Harvey, Ohad Kadan, Mina Lee, Yan Liu, Jeff Pontiff, Kalle Rinne, Olivier Scaillet, Christopher Simmons, Peter Westfall, Michael Wolf, Lu Zhang, Hao Zhou, and seminar participants at the 2017 FRA conference, 2017 Luxembourg Asset Management Summit, 2017 Inquire Europe Autumn Seminar, 2017 Lone Star Conference, 2017 SFS Asia Cavalcade, 2018 NBER Summer Institute, 2017 TAU Finance Conference, 2018 Telfer Accounting and Finance Conference, 2018 ITAM Finance Conference, 2018 ABFER Conference, 2018 University of South Australia Conference, 2018 SFM Conference, Australian National University, Caltech, Case Western University, Chinese University of Hong Kong, Lancaster University, Texas Tech University, Tinbergen Institute Amsterdam, University of Queensland, University of New South Wales, University of San Diego, University of Technology Sydney, and University of Texas at Dallas for helpful discussions. Amit Goyal thanks Swiss National Fund grant 100018_182198 for financial support. Alessio Saretto thanks the generous support of the Cyber Infrastructure for Research Department at the University of Texas at Dallas. All errors are our own.

Researchers in finance, as in other academic fields, have long been worried about the possibility that a substantial number of reported discoveries might be false.[1] The problem might be particularly severe for empirical asset pricing studies that investigate cross-sectional predictability in stock returns. Part of the difficulty might be the way in which these strategies are discovered or the absence of well-specified research practices (Harvey 2017).

Nevertheless, there exists a fundamental statistical issue: when many hypotheses are tested, some will appear significant at conventional thresholds from classical hypothesis testing (CHT) even if they are true under the null. We refer to such rejections as false rejections or false discoveries under CHT. Although independently carried out by many researchers, the predictability studies share a common question (i.e., are (expected) stock returns predictable?) and, almost invariably, at least one common dataset (i.e., CRSP). Accounting for multiple hypotheses testing (MHT) on a common question and dataset, thus, might offer not only a way to determine how severe is the false rejection problem but also a way to deal with it.

Despite a number of recent studies that propose the application of MHT methods, the magnitude of the false discovery problem and, closely related, the thresholds that should be applied to test statistics to reduce false discoveries, remain a matter of debate. For example, drawing from a meta-study of 296 published discoveries, Harvey, Liu, and Zhu (2016) suggest an MHT t-statistic threshold higher than three, but warn that there are very good reasons why this number might be too low for studies that are not grounded in theory. Using the Benjamini and Yekutieli (2001) procedure to obtain an MHT threshold, they calculate that 80 of the 296 discoveries (i.e., 27%) are rejected under CHT but not under MHT. On the other hand, in a set of randomly generated trading strategies and using a bootstrap approach (but not formal MHT methods), Yan and Zheng (2017) settle on a rather low statistical threshold that implies a very small number of false discoveries and thousands of informative signals.

[1] Leamer (1983, p. 43) complains of the "fumes which leak from our computing centers" and argues that it is critical that the fragility of results be studied in a systematic way to establish or alter beliefs. In fact, Harvey, Liu, and Zhu (2016) argue that "... most claimed research findings in financial economics are likely false."

The conflicting evidence stems from three sources: the method used to determine an MHT adjusted threshold, the set of trading strategies that are studied (Harvey, Liu, and Zhu 2016 and Yan and Zheng 2017 differ fundamentally on both counts), and the power of the empirical tests.

The problem of choosing an appropriate MHT method (we discuss various MHT methods in Section 1 of the paper) can be clarified by weighing the underlying assumptions in a simulation context. Since asset pricing studies rely on datasets where the data exhibit serial and cross-sectional dependence, a method that is robust to those features might be preferable. Using simulation evidence detailed in Section 3, we suggest the adoption of the procedure advanced by Romano and Wolf (2007) and Romano, Shaikh, and Wolf (2008), RSW henceforth.

Even after having identified a well-suited MHT method, resolving the issue of which (correct) set of trading strategies to study is relatively more complicated. As Harvey, Liu, and Zhu (2016) note, in an ideal setting, one would apply MHT to all the strategies that researchers have attempted to test (say set $\mathcal{S}^{\mathcal{R}}$), and not just those that eventually become public (say set $\mathcal{S}^{\mathcal{P}}$) because they end up in the tails of the distribution of t-statistics, (i.e., those for which the null was rejected and, thus, made it to a circulated paper). The set $\mathcal{S}^{\mathcal{P}}$ does not contain all the strategies that are obviously false (for which the t-statistic is lower than even the conventional thresholds). Therefore, the application of MHT on the set $\mathcal{S}^{\mathcal{P}}$ produces a low threshold for statistical significance. A low MHT threshold implies that a small proportion of strategies that would be rejected by CHT are likely false. Since we do not know which strategies are false in the population, we refer to strategies for which the null is rejected under the CHT threshold but not under the MHT threshold as 'Classical but not Multiple' (CNM) rejections.

One feasible alternative is to consider a large set of randomly generated strategies (say set $\mathcal{S}^{\mathcal{E}}$ that the econometrician draws) that mechanically contains most of the strategies that researchers have attempted but that were not significant. Unfortunately, the set $\mathcal{S}^{\mathcal{E}}$ also

contains many strategies that researchers would never study because their economic foundation is not immediately justifiable. Assuming that economic reasoning gives researchers an advantage in filtering out strategies for which the null of no predictability is true, the set $\mathcal{S}^{\mathcal{E}}$ is then biased in the opposite direction: it produces MHT adjusted thresholds that are too high and, consequently, a very high estimate of CNM rejections. We generate such a set (Section 2 gives details on the construction of random trading strategies) and, applying the RSW procedure, we find that 98% of the signals with t-statistics of long-short portfolios abnormal returns and Fama and MacBeth (1973) (FM) slope coefficients higher than the CHT threshold of 1.96 do not cross the higher MHT threshold implied by our RSW procedure.

Finally, even if we had access to the correct set $\mathcal{S}^{\mathcal{R}}$ of strategies, we could still have a problem because the proportion of CNM rejections may not be close to the unobservable true rate of false rejections (i.e., proportion of strategies that are falsely rejected by CHT). In situations where the power is not very high, the MHT procedure would not reject all hypotheses that are false under the null, and would falsely reject some hypotheses (based on its tolerance level). In such scenarios, the proportion of CNM rejections would not be very close to the proportion of false rejections under CHT. Low signal-to-noise ratios and limited number of observations (in the time series) make power a significant problem in many finance applications. For example, Andrikogiannopoulou and Papakonstantinous (2019) and Harvey and Liu (2019) show that limited power hinders the application of some MHT methods in mutual fund performance evaluation.

In summary, the proportion of false rejections under CHT is an unobservable quantity that cannot be estimated directly from the data. It can only be indirectly inferred by making explicit assumptions about the data generating process and about the ability of researchers to filter out strategies that are economically unsound.

We propose a statistical model (details are in Section 4 of the paper) that allows us to use the information contained in the sets $\mathcal{S}^{\mathcal{E}}$ and $\mathcal{S}^{\mathcal{P}}$ to extract the statistical properties of the set $\mathcal{S}^{\mathcal{R}}$, and hence indirectly infer the proportion of false rejections under CHT. Signals

are drawn from a mixture distribution wherein some signals are informative and others are not. The critical assumption is that the probability of an informed signal in the set $\mathcal{S}^{\mathcal{R}}$ is proportional, by a factor larger than one, to the equivalent probability in the set $\mathcal{S}^{\mathcal{E}}$. The intuition behind this assumption is that researchers can use economic reasoning to select better signals than those randomly selected by the econometrician. Hence, researchers are endowed in the model with a higher ability to draw informative signals.

The key parameters in the model are the noise level in the signals, σ_η, the researchers' superior informative signal drawing ability, Ω, and the proportion of informative signals in the econometrician's set, π. In contrast to other model parameters that can be directly set to quantities that are observable in the data, the calibration of these parameters requires particular attention. Our "identification" strategy for Ω and π relies on our knowledge that the two estimates of CNM rejections that are available to us are biased in opposite directions: A high fraction of CNM rejections in the set $\mathcal{S}^{\mathcal{E}}$ implies a low π, and a low proportion of CNM rejections in the set $\mathcal{S}^{\mathcal{P}}$ implies a high Ω. As it is very difficult to construct an estimate of CNM rejections for published discoveries (in the set $\mathcal{S}^{\mathcal{P}}$) that is free of any bias, we rely heavily on the Harvey, Liu, and Zhu (2016) estimate. The parameters are, therefore, calibrated so that the model matches, among other quantities, the fraction 98% (27%) of CNM rejections in the set $\mathcal{S}^{\mathcal{E}}$ ($\mathcal{S}^{\mathcal{P}}$).[2] The noise in the signal affects how much of individual stocks' alpha can be ported into a portfolio alpha. Hence, we calibrate σ_η by asking the model to produce the same ratio of (significant) portfolio alphas to stock alphas. Given that our simulation has long time-series (500 observations), our design helps us to side-step the power problem alluded to earlier.

Once calibrated, the statistical model allows us to quantify the proportion of false dis-

[2]Harvey, Liu, and Zhu (2016) and Chen (2019) recover the set $\mathcal{S}^{\mathcal{R}}$ by imposing some distributional assumptions and matching the properties of the observed t-statistics in the set $\mathcal{S}^{\mathcal{P}}$. Relative to these studies, instead of matching t-statistics corresponding to a certain frequency (i.e., a certain quantile), we match the frequencies corresponding to certain t-statistics (i.e., fraction of t-statistics larger than a threshold). Since we rely on CNM proportions, our approach has the disadvantage that we may not exactly match the tails of the distribution of t-statistics in the set $\mathcal{S}^{\mathcal{P}}$. However, since we also rely on the set $\mathcal{S}^{\mathcal{E}}$, our approach has the advantage that it avoids under-sampling insignificant strategies.

coveries (our positive contribution) and to provide MHT adjusted thresholds for finance applications that study predictability of trading strategies using standard finance datasets (our normative contribution). We estimate the fraction of false discoveries under CHT to be 45.3%. Our estimate of false rejections is based only on the assumption that published strategies have statistically significant alpha and FM coefficient t-statistics under CHT. In typical published research, authors need a theory/motivation/story to have a better chance of getting published. These additional qualitative hurdles are not taken into consideration in our purely statistical framework. We interpret our estimate as the proportion of false discoveries that one should expect if finance researchers do not adopt MHT adjustments and, instead, rely only on CHT thresholds.

To put this number in perspective, we turn to Linnainmaa and Roberts (2018), who report that, of the 36 strategies that they study, 20 strategies have an insignificant Fama and French (1993) three-factor alpha in the pre-discovery period. This represents a false rejection rate of 55.6%. Our estimate is in the same ball-park as theirs with differences being explained by different approaches, strategies, and sample period.

Turning to the normative goal of our paper, we find that, conditional on the assumptions that we make, the MHT adjusted thresholds for t-statistics of time-series alphas and cross-sectional FM regression slopes are 3.84 and 3.38, respectively, for a six-factor model that includes Fama and French (2015) factors and a momentum factor.

The interpretation of our thresholds warrants a few considerations. First, our estimates are for a specific context of empirical asset pricing studies that use a six-factor model and are concerned with abnormally profitable trading strategies. More research is needed to calculate thresholds and develop an understanding of how to use them in the context of corporate finance applications (such as the ones that typically rely on panel regressions). Second, strong priors about the validity of a signal might require special considerations. Harvey (2017) presents a Bayesian approach to deal with such situations. Third, our statistical approach (indeed any MHT procedure) addresses only the multiplicity problem that naturally arises

in the research process, and hence is of little use in aligning researchers' personal incentives with the scientific discovery process. We refer again to Harvey (2017) for a discussion of the publication process and the incentives therein.

We conduct many robustness checks (details in Section 5 of the paper) that consider alternative assumptions about the data generating process, different factor models, and alternative ways of constructing random strategies. Our estimates of MHT adjusted thresholds and the proportion of false rejections are not overly sensitive to these choices.

The same set of robustness checks, however, highlight an important limitation of our study. Our estimation of false rejections depends critically on the assumptions underlying our statistical framework that allows us to characterize the set of strategies studied by the researcher and on the two estimates of CNM rejections (27% and 98%) that we use. Lower (higher) CNM rejections in either set will lower (raise) our estimate of false rejections. For instance, if the proportion of CNM rejections in the set $\mathcal{S}^{\mathcal{P}}$ was much higher then this would intuitively imply that the researcher is not very good at screening out null strategies. This would, therefore, increase the estimate of false rejections. In fact, we calculate our own measure of CNM rejections on a set of public strategies used by Chen and Zimmerman (2018) and find it to be as high as 50%. Re-calibrating the model, we find, as expected, that the fraction of false rejections increases correspondingly.

We also note that our paper is focused mainly on false (and CNM) rejections. An equally important issue might be the proportion of false non-rejections (i.e., Type II errors). Our simulations allow us to bypass the issue of power as we know exactly which signals are informative. Nevertheless, in real data, balancing Type I and II errors remains a valid concern. See Harvey and Liu (2019) for a detailed analysis of how to tradeoff between size and power in an MHT procedure.

One of the by-products of our investigation is the analysis of dual hurdles of time-series alpha and FM cross-sectional slope coefficient for a signal. There are economic advantages to each approach (see, for example, Fama and French 2008). Statistically, imposing dual hurdles

improves the size of the tests—the probability of a strategy having statistically significant alpha *and* FM coefficient by luck is lower than having either of them. While these size improvements apply to both CHT and MHT, we show that MHT methods that incorporate both the testing approaches have very good size properties (very low false rejection percentage). Therefore, we recommend the application of the afore-mentioned MHT thresholds for *both* alpha and FM coefficients. Note, however, that the application of dual hurdles may be too strict in the sense that it imposes a tighter control on the proportion of false discoveries than that by each of the two individual hurdles (at the expense of low power). Ideally, we would prefer an MHT procedure that ensures the desired (and not stricter) control from dual hurdles. Absent such a procedure, we offer some heuristics on how to adjust the dual MHT thresholds.

Our research echoes the increasing skepticism about the validity of many research findings in a variety of fields. While the findings on the lack of replicability in medical research by Ioannidis (2005) are widely cited, the economics profession has also made an effort to tackle this problem. Leamer (1978, 1983) famously complains about specification searches in empirical research and asks researchers to take the "con" out of econometrics. Dewald, Thursby, and Anderson (1986), McCullogh and Vinod (2003), and Chang and Li (2018) also report disappointing results from replication of economics papers. The use of replication in finance is less widespread with McLean and Pontiff (2016) being a notable recent exception.

Our paper also joins the list of the growing finance literature that studies the proliferation of discoveries of abnormally profitable trading strategies and/or pricing factors and its relation to data-snooping biases in finance. See Lo and MacKinlay (1990), Foster, Smith, and Whaley (1997), Conrad, Cooper, and Kaul (2003), and Karolyi and Kho (2004) for early work emphasizing statistical biases in hypothesis testing. The question of whether the profitability of published strategies survives the test of time is studied in Schwert (2003), Chordia, Subrahmanyam, and Tong (2014), McLean and Pontiff (2016), Linnainmaa and Roberts (2018), and Hou, Xue, and Zhang (2019). Towards the turn of the century, more

formal statistical approaches were developed and applied to the problem of evaluating multiple strategies. See, for example, Sullivan, Timmermann, and White (1999), White (2000), and Romano and Wolf (2005). The MHT approach has been more recently applied to financial settings in Barras, Scaillet, and Wermers (2010), Harvey, Liu, and Zhu (2016), Harvey and Liu (2019), and emphasized in the presidential address of Harvey (2017). Chen and Zimmermann (2018) offer a different Bayesian perspective, by estimating the publication bias in the context of a model that describes the impact of the publication process on experiments designed by researchers. In Section 4.2 we relate our results to Chen (2019) who studies 156 strategies and concludes that no major adjustment to statistical thresholds is required.

Before closing this introduction, we would like to clarify that the goal of our paper is neither to find outperforming strategies (in contrast to Yan and Zheng 2017) nor to construct better factors. To discover better predictive signals, other approaches such as factor analysis or machine learning might be more constructive.[3] Rather, our goal is to provide a careful estimate of the proportion of the false rejections that are likely to obtain because of the use of CHT rather than MHT, and to provide MHT adjusted thresholds for finance researchers studying profitability of trading strategies using standard finance datasets.

1. Multiple hypotheses testing

Consider a multiple testing situation in which a researcher explores S hypotheses using a particular data set. S_0 (S_1) of these hypotheses are true (false) under the null, H_0. Each hypothesis is evaluated at a confidence level α (for example, 5%).[4] A total of R hypotheses are rejected. The breakup of rejections and non-rejections is organized in the following table:

[3]See, for example, Freyberger, Neuhierl, and Weber (2018), Gu, Kelly, and Xiu (2018), and Kozak, Nagel, and Santosh (2018).

[4]The use of symbol α to denote both the significance level as well as the abnormal returns from a factor model is standard. We hope that this does not cause any confusion and the usage is clear from the context.

	H_0 not rejected	H_0 rejected	Total
H_0 True	T_0	F_1	S_0
H_0 False	F_0	T_1	S_1
	$S-R$	$R=F_1+T_1$	S

Overall, F_1 hypotheses are falsely rejected, and F_1/R is the false discovery proportion (FDP). F_0 hypotheses should have been rejected but are not, and $F_0/(S-R)$ is the false non-discovery proportion (FNDP). As S grows, so does F_1, making the probability of making at least one false discovery (i.e., the rate of Type I error) increase rapidly. For example, at a conventional significance level of 5%, if all strategies are true under the null and mutually independent, the probability of making at least one false discovery is $1-0.95^{10}=40\%$ when ten hypotheses are tested, and over 99% when 100 hypotheses are tested.

MHT methods control how large is the number of false rejections, F_1. There are three broad approaches in the statistics literature that differ in the statistical definition of what is actually controlled: family-wise error rate (FWER), false discovery rate (FDR), and false discovery proportion (FDP). In this section, we discuss the basic intuition behind these approaches relegating the implementation details to Appendix A.

1.1 Family-wise error rate (FWER)

The most basic idea in MHT is to control the FWER, defined as the probability of rejecting at least one of the true null hypotheses:

$$\text{FWER} = \text{Prob}(F_1 \geq 1).$$

A testing method is said to control the FWER at a significance level α if FWER $\leq \alpha$. There are four main approaches to controlling FWER. (i) Bonferroni (1936), (ii) Holm (1979), (iii) Bootstrap reality check of White (2000), and (iv) StepM method of Romano and Wolf (2005). The Bonferroni correction does not require any assumptions, the second and third methods assume independence between hypotheses, and the fourth method is valid under arbitrary

dependence in data. MHT methods that control FWER are particularly strict because they try to keep the number of false discoveries at the lowest level, i.e., one false discovery.

All these methods owe their origin to the Bonferroni (1936) adjustment. The logic behind Bonferroni is that if each hypothesis is tested at some significance level α^*, then each one hypothesis gets erroneously rejected with probability α^*. Since S hypothesis are tested, the expected number of false rejections is $E(F_1) = S \times \alpha^*$. Since $\text{Prob}(F_1 \geq 1) \leq E(F_1)$, to guarantee that FWER $\leq \alpha$, one should set $\alpha^* = \alpha/S$. As the number S of hypotheses being tested increases, the correction becomes more and more stringent, leading to very high t-statistic thresholds for rejection. To varying degrees, all procedures that control FWER suffer from the same problem.

One might be willing to tolerate a larger number of false rejections, in order to discover a larger number of true rejections. For example, the k-FWER control allows a researcher to allow F_1 to be as large as k, as opposed to one. Formally:

$$k\text{-FWER} = \text{Prob}\ (F_1 \geq k).$$

Therefore a possible solution to relax the stringent constraints imposed by (1-)FWER is to adopt a k-FWER procedure.

1.2 False discovery proportion (FDP)

An alternative to controlling the "number" of false rejections F_1 is to control the "proportion" F_1/R of false rejections (FDP). Formally, a multiple testing procedure is said to control FDP at proportion γ and significance level α if

$$\text{Prob}\ (\text{FDP} \geq \gamma) \leq \alpha.$$

Since FDP is a random variable, this condition essentially imposes a restriction on the left tail of its distribution. As such, FDP control guarantees that, in any application, the realized FDP cannot (statistically) exceed a particular threshold γ.

Many algorithms have been proposed to control FDP (see Farcomeni 2008 for a general review). We implement the RSW method proposed by Romano and Wolf (2007) and Romano, Shaikh, and Wolf (2008). This procedure runs a sequence of k-FWER procedures with $k = 1, 2$, etc. Say that, at a generic step k, the total number of rejections is R_k. Since step k corresponds to k-FWER, we can be sure that at most k false rejections have taken place so far. This means that one additional rejection in the next step will make the FDP equal to $k/(R_k+1)$. To control the FDP at the desired significance level γ, we undertake the next step only if $k/(R_k + 1) \leq \gamma$, or equivalently only if $R_k \geq k/\gamma - 1$. The last inequality determines the stopping condition of the algorithm.

One desirable property of the RSW algorithm is that it is based on the resampling method. Resampling methods have been studied by Romano and Wolf (2005), for instance. As long as the resampling method preserves the dependence structure in the data, it also allows FDP control under the same arbitrary dependence structure without requiring any distributional assumptions on the underlying data. The main disadvantage of the RSW algorithm is that, because of its recursive nature and because it relies on resampling the data, it is computationally burdensome.

In our implementation, we rely on a bootstrap procedure that maintains the cross-sectional and the time-series structure in the data. We provide the details of the bootstrap in the appendix Section A.5. We note here that our approach is inspired by Kosowski, Timmermann, Wermers, and White (2006) and Fama and French (2010), and used recently by Yan and Zheng (2017). The main idea is to bootstrap the cross-section of strategy returns through time thereby preserving the cross-sectional dependence structure in strategy returns.

1.3 False discovery rate (FDR)

An alternative to controlling the tail behavior of the FDP in a particular application is to control its expected value, the false discovery rate (FDR). Formally, a multiple testing method is said to control FDR at level δ if

$$\text{FDR} \equiv \text{E(FDP)} \leq \delta,$$

where δ is a tolerance level imposed by the researcher. Since procedures that control the tail behavior of FDP are stricter than those that control the mean behavior, γ and α of 5% in FDP control typically imply an FDR control of δ less than 5%.

The two main FDR control methods that we consider are the Benjamini and Hochberg (1995) method and its extension by Benjamini and Yekutieli (2001). We label these methods BH and BHY, respectively. We note that BH assumes independent tests while BHY accounts for (partial) dependence in the data. As is often the case, more general assumptions imply a loss of power. Therefore, BHY is less powerful than BH; BHY typically implies higher thresholds than those by BH.

In BH and BHY, one orders the hypotheses by their p-values so that the first hypothesis is the most significant. Say that j is the generic index of an ordered hypotheses (i.e., j is the j-th most significant hypothesis and has p-value of p_j). If we rejected j hypotheses, the expected number of false rejections would be $S \times p_j$ and, therefore, $(S \times p_j)/j$ would be the realized FDR. The procedure stops at j^* such that the corresponding FDR is less than δ.

One very desirable property of FDR control methods (especially the BH and BHY algorithms) is that they are computationally very easy to implement. The main disadvantage is that in a given application (i.e., in a given dataset), the realized FDP can be far away from the average (FDR) and, hence, from the tolerance level δ (see Genovese and Wasserman 2006 for a more detailed discussion). In other words, specifying the FDR control δ at 5%, say, could still lead to a *realized* FDP *in a particular dataset* being far above 5%, with the incidence of such cases increasing the farther away we move from the independence assumption in the tests. Therefore, FDR control is better suited for cases where a researcher can analyze a large number of data sets, allowing one to make confidence statements about the average realized FDP across such data sets.[5]

[5]We thank Michael Wolf for clarifying this important difference to us.

2. Data and trading strategies

In this section, we construct a set $\mathcal{S}^{\mathcal{E}}$ of randomly generated strategies and apply various MHT methods to this set. Monthly returns and prices are obtained from CRSP. Annual accounting data come from the merged CRSP/COMPUSTAT files. We collect all items included in the balance sheet, the income statement, the cash-flow statement, and other miscellaneous items for the years 1972 to 2015. We choose 1972 as the beginning of our sample as it corresponds to the first year of trading on Nasdaq that dramatically increased the number of stocks in the CRSP dataset. All our results are robust to beginning the sample in 1963, which is the first date on which the COMPUSTAT data are not affected by backfilling bias. Following convention, we set a six month lag between the end of the fiscal year and the availability of accounting information.

We impose several filters on the data to obtain our basic sample. First, we include only common stocks with CRSP share codes of 10 or 11. Second, we require that data for each variable be available for at least 300 firms each month for at least 30 years during the sample period. Third, in FM regressions described later, we require that data be available for all independent variables (including the variable of interest) for at least 300 firms each month for at least 30 years during the sample period.

There are 185 variables that survive the above filters and can be used to develop trading signals. We list these variables in Appendix Table A1. We refer to these variables as *Levels*. We also construct *Growth rates* from one year to the next for these variables. Since it is common in the literature to construct ratios of different variables we also compute all possible combinations of ratios of two levels, denoted *Ratios of two*. Finally, we also compute all possible combinations that can be expressed as a ratio between the difference of two variables to a third variable (i.e., $(x_1 - x_2)/x_3$). We refer to this last group as *Ratios of three*. We obtain a total of 2,393,641 possible signals.

We evaluate trading signals by (i) estimating abnormal performance of the hedge portfolios using a factor model and (ii) by evaluating the ability of the signal to explain the

cross-section of returns.

2.1 Hedge portfolios

We sort firms into value-weighted deciles on June 30 of each year and rebalance these portfolios annually. We discuss results using alternative 2×3 portfolios in robustness Section 5.3. The first portfolio formation date is June 1973 and the last formation date is June 2015. We require a minimum of 30 stocks in each decile (300 stocks in total) in a month to consider that month as having a valid return. The signal is considered to have generated a valid portfolio if there are at least 360 months of valid returns. We consider long-short portfolios only. Thus, we compute a hedge portfolio return that is long in decile ten and short in decile one. Since we do not know ex-ante which of the two extreme portfolios has the largest average return, our hedge portfolios can have either positive or negative average returns. Obviously, it is always possible to obtain a positive average return for a hedge portfolio that has a negative average return by taking the opposite positions. For expositional convenience, we decide not to force average returns to be positive.

We compute abnormal returns for our strategies using the Fama and French (2015) five-factor model augmented with the momentum factor (Carhart 1997). Results using alternative factor models are presented in the robustness Section 5.3. For each trading strategy, we run a time-series regression of the corresponding hedge portfolio returns on the factors and obtain the alpha as well as its heteroskedasticty-adjusted t-statistic, t_α.

2.2 Fama-MacBeth regressions

Given that the alphas of the long-short portfolio effectively consider the efficacy of the strategy in only 20% of the sample, we also evaluate a signal's ability to predict returns in the cross-section of stocks using FM regressions. However, note that these regressions impose linearity in the relation between firm characteristics and returns. We evaluate the ability of the signal to explain stock returns by estimating the following cross-sectional regression

each month:

$$R_{it} - \widehat{\beta}_i F_t = \lambda_{0t} + \lambda_{1t} X_{it-1} + \lambda_{2t} Z_{it-1} + e_{it}, \tag{1}$$

where X is the variable that represents the signal and Z's are control variables. We use the most commonly used control variables, namely size (i.e., the natural logarithm of the firm's market capitalization), natural logarithm of the book-to-market ratio, past one-month and 11-month return (skipping the most recent month), asset growth, and profitability ratio. Book-to-market is calculated following Fama and French (1993) while asset growth and profitability are calculated following Fama and French (2015). We risk-adjust the returns on the left-hand-side of equation (1) following Brennan, Chordia, and Subrahmanyam (1998). We use the same factor model used to calculate hedge portfolio alphas, and calculate full-sample betas, $\widehat{\beta}$s, for each stock. We require at least 60 months of valid returns to estimate the time-series regression beta. All right-hand-side variables are winsorized at the 0.5 and 99.5 perecntiles in these regressions. We discuss results using ranks of right-hand-side variables (instead of their numerical values) in the robustness Section 5.3.

In estimating the cross-sectional regressions, we require a minimum of 300 stocks in a month. Finally, we require a minimum of 360 valid monthly cross-sectional estimates during the sample period to calculate a valid λ_1 coefficient for a signal. Then, we calculate the FM coefficient λ_1 as well its heteroskedasticty-adjusted t-statistic (t_λ). Given that we require a valid beta for each stock and data on additional control variables, the data requirements for the FM regressions are slightly more stringent than those for portfolio formation.

2.3 Raw returns, alphas, and FM coefficients

Table 1 describes the empirical distributions of t-statistics for monthly raw returns, alphas, and FM coefficients. For each variable, we calculate the cross-sectional means, medians, standard deviation, minimum, and maximum across all signals. Also reported are the percentage of t-statistics that, in absolute value, cross thresholds of 1.96 and 2.57 corresponding

to CHT significance levels of 95% and 99% respectively.

Table 1: Significance of trading strategies

The table reports the cross-sectional mean, median, standard deviation, minimum, and maximum of the t-statistics of monthly average return, alpha and FM coefficients of the 2,393,641 trading strategies, whose construction is described in the text. The factor model is the Fama and French (2015) five-factor model augmented with the momentum factor. We run FM regressions as in Brennan, Chordia, and Subramanyan (1998), and we include size, book-to-market, profitability, asset growth, and one- and twelve-month lagged returns as additional controls beside the variable used to construct the trading strategy. All right-hand-side variables are winsorized at the 0.5 and 99.5 percentile in these regressions. We also report the percentage of strategies with t-statistics that are larger than 1.96 and 2.57 in absolute value. The sample period is 1972 to 2015.

	Mean	Median	Std	Min	Max	$\%\|t\| > 1.96$	$\%\|t\| > 2.57$
Return	−0.08	−0.10	1.21	−7.04	6.72	10.6	3.7
Alpha	−0.19	−0.19	1.62	−7.80	9.01	23.2	11.9
FM	0.07	0.03	1.64	−8.63	8.12	20.0	11.2

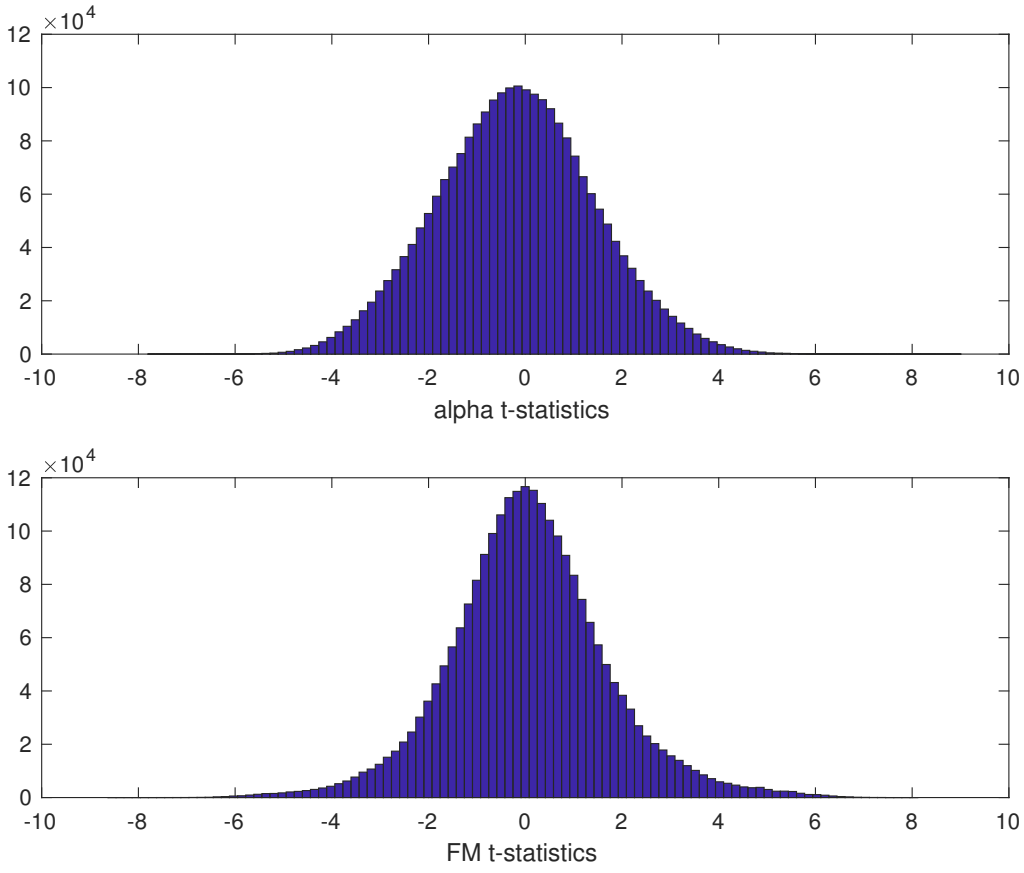
A large number of strategies have average return t-statistics that exceed conventional statistical significance levels: 254,593 (88,898) strategies have average return t-statistics larger than 1.96 (2.57) (in absolute value), corresponding to 10.6% (3.7%) of the total number of strategies. In unreported results, we find that the dispersion in the performance of strategies is largest in the subset of strategies *Ratios of three*. The most profitable and statistically significant returns come from this group. The largest absolute average monthly return is 1.75% and the largest absolute t-statistic is 7.04.

The distribution of alpha and FM coefficient t-statistics reveals even more exceptional performance of strategies than that in raw returns. For example, 23.2% (11.9%) of alpha t-statistics are higher than 1.96 (2.57). The fraction of FM coefficient t-statistics that cross these conventional statistical thresholds is also high. For example, 20.0% (11.2%) of t_λs are higher than 1.96 (2.57), in absolute value.

Figure 1 depicts the histograms for the cross-section of alpha and FM coefficient t-statistics. The distributions are generally centered around zero and appear to be normally distributed. The support for the distributions is consistent with the cross-sectional standard deviations in Table 1. Note that the distributions of t_α and t_λ are fat-tailed, consistent with

the large number of CHT rejections in Table 1.

Figure 1: Empirical distributions of portfolios t-statistics
We construct trading strategies as described in the text. The figure shows cross-sectional histograms for alpha t-statistics and Fama-MacBeth coefficient t-statistics. Alphas are computed relative to the Fama and French (2015) five-factor model augmented with a momentum factor. Please see Table 1 for further details. The sample period is 1972 to 2015.



Despite their impressive performance, most of the strategies in the right-tail do not have an immediate economic motivation and would presumably be never considered by finance researchers. As we explained in the introduction, and as will become clear shortly, we use the set $\mathcal{S}^{\mathcal{E}}$ of randomly generated strategies only as a crutch to help us get closer to the set $\mathcal{S}^{\mathcal{R}}$. Accordingly, we next apply MHT methods to these data and calculate the corresponding rejections rates.

2.4 Applying MHT to the data

As discussed in Section 1, all MHT methods consist of adjustments to the threshold t-statistic associated with a desired level of significance. We use $\alpha = 5\%$ for significance level, $\gamma = 5\%$ for FDP control, and $\delta = 5\%$ for FDR control. The t-statistic thresholds and the fraction of rejections corresponding to this threshold are reported in Table 2. Among FWER control methods, thresholds for both alpha and FM coefficient t-statistics are high with extremely low rejection rates; only 0.04% percent of alpha t-statistics are higher than the threshold of 5.60.[6] Thresholds are also high for BHY and RSW methods but the rejection rates are higher than those for FWER control methods. For instance, the RSW threshold for alpha t-statistic is 3.96 and 1.4% of alphas cross this threshold. As expected, BH method imposes lower thresholds and has more rejections than the BHY method.

Comparing rejection rates from CHT and MHT gives us an estimate of the proportion of CNM rejections. Table 1 shows that 23.2% of strategies have $|t_\alpha| > 1.96$. Table 2 shows that 1.4% of the strategies have $|t_\alpha| > 3.96$, the threshold imposed by the RSW method. This yields a CNM rejection rate of 94.1% ($= 1 - 1.4/23.2$).

We tabulate the proportion of CNM rejections in the last row of each panel in Table 2. Methods that control FWER have such high thresholds that there are very few rejections under these methods. This also implies that these methods classify essentially all CHT rejections as CNM rejections. By construction, lower MHT thresholds lead to more MHT rejections and a lower proportion of CNM rejections. Accordingly, we find that the proportion of CNM rejections is the lowest for the BH method. Since FM coefficient thresholds are lower than the corresponding alpha thresholds, we also find that the FM coefficient CNM rejections are lower than the alpha CNM rejections. Overall, we find that the proportion of CNM rejections ranges from 57.4% to 99.9%.

[6]Even so, 0.04% of more than two million strategies is a sizeable number. However, we reiterate that our goal is not to find outperforming strategies. These strategies in the extreme right-tail lack economic foundation.

Table 2: MHT thresholds and rejection rates

The table shows alpha and FM coefficient statistical thresholds adjusted for MHT as well as the percentage of rejected strategies. We use FWER control (Bonf(erroni) and Holm), FDR control (BH and BHY), and FDP control (RSW) MHT methods. We use $\alpha = 5\%$ for significance level, $\gamma = 5\%$ for FDP control, and $\delta = 5\%$ for FDR control. Fraction of CNM rejections is computed as follows. 23.2% of the strategies have $|t_\alpha| > 1.96$; 1.4% strategies have $|t_\alpha| > 3.96$, the RSW threshold; the estimate of the proportion of CNM rejections is $1 - 1.4/23.2 = 94.1\%$. The sample period is 1972 to 2015.

	Bonf	Holm	BH	BHY	RSW
	Alpha t-statistic				
Threshold	5.60	5.60	2.87	4.18	3.96
Rejections	<0.1	<0.1	8.1	0.9	1.4
CNM Rejections	99.9	99.9	64.9	96.2	94.1
	FM coefficient t-statistic				
Threshold	5.42	5.42	2.86	3.89	3.32
Rejections	0.4	0.4	8.5	3.1	5.5
CNM Rejections	97.2	97.2	57.4	84.5	72.7
	Alpha and FM coefficient t-statistics				
Rejections	<0.1	<0.1	0.8	<0.1	0.1
CNM Rejections	99.9	99.9	84.3	99.1	97.9

2.4.1 Dual hurdles

We also calculate rejection rates for strategies that cross the statistical threshold for both alpha and FM coefficient t-statistic. In other words, we require consistency between results obtained by studying portfolio returns and those derived from FM regressions. There are economic and statistical reasons for doing so.

Economically, as discussed in Section 2.2, there are advantages and disadvantages to both the portfolio sorts and the FM regressions. We would like a trading signal to not only generate a high long-short portfolio alpha but also to explain the broader cross-section of returns in a regression setting. Therefore, many papers in the finance literature report both. It is our general understanding that many researchers follow the procedure of not rejecting the null hypotheses of no predictability if they have statistically significant t_α but insignificant t_λ or vice-versa.

Statistically, imposing dual hurdles can be seen as a way to improve the size of the tests. The probability of a strategy having statistically significant alpha *and* FM coefficient by luck is lower than having either of them (if the alphas and FM coefficients are not perfectly correlated). These size improvements apply to both CHT and MHT testing methods.

Imposing the dual filter drastically decreases the number of rejections even for CHT. For example, Table 1 shows that 23.2% (20.0%) of the strategies have alpha (FM coefficient) t-statistic greater than 1.96. The intersection of these two sets gives us only 5.4% of the total number of strategies (number not reported in the table). Thus, consistency across testing procedures play an important role in restricting the set of statistically significant strategies even with CHT.

Imposing the constraint of consistency between alpha and FM coefficients, the proportion of CNM rejections ranges from 84.3% (for the BH method) to close to 100% (for FWER control methods). Looking a bit ahead, if we use the best MHT method, namely RSW, the proportion of CNM rejections is 97.9%.

3. Monte Carlo simulations

We run Monte Carlo simulations with two main objectives in mind. First, we wish to clarify the properties of the MHT methods discussed in Section 1. This allows us to give preference to some MHT methods over others and, consequently, rely on only those methods for inference. Second, we investigate the advantage of implementing the dual hurdle procedure of Section 2.4.1 on alphas and FM coefficients. This practice allows us to show in a controlled experiment that the proportion of false discoveries is greatly reduced by the dual hurdle.

3.1 Simulation details

3.1.1 Data generating process

Stocks: There are N stocks in the economy. Stock returns are generated from a linear factor model:

$$R_{it} = \alpha_i + \beta_i' F_t + \epsilon_{it}. \tag{2}$$

Alphas are drawn from a normal distribution $\mathcal{N}(0, \sigma_\alpha)$. The random shocks ϵ are distributed as $\mathcal{N}(0, \sigma_\epsilon)$ where $\sigma_\epsilon = 15.1\%$ to match the average monthly standard deviation of residuals in the data. We set N = 2,000 to roughly equal the monthly average number of stocks for which we can generate a signal in the real data (2,383), and $T = 500$ to roughly equal the number of months in the data (516). Cross-correlation in stock returns is mechanically produced by the factor structure. The simulated empirical distribution of factors and factor sensitivities are matched to the corresponding empirical quantities in the data. In particular, we simulate a six-factor model that mimics the statistical properties of the five Fama and French (2015) factors augmented with the momentum factor. Each month, we draw the factor returns, F_t, from a multivariate normal distribution with means and covariance matrix that match those of the actual distribution of factor returns. For each stock i, we draw the 6×1 vector of betas, β_i, from a multivariate normal distribution with means and covariance matrix matched to that of the cross-sectional distribution of betas from the actual data.

Signals: In each period t, the econometrician draws S noisy signals from a set $\mathcal{S}^{\mathcal{E}}$. Some of the signals contain information about returns and some do not. We can think about a signal as a variable constructed from accounting data. Some signals (for example, book to market ratio) are informative in the sense that they carry some cross-sectional predictability about stocks' abnormal returns (i.e., α_i), while others (for example, the ratio of debt due in one year to depreciation) are not. We draw an informative signal with probability π. After

drawing a signal, its realizations corresponding to stock i in time period t are given by:

$$\begin{aligned} s_{it} &= \alpha_i + \eta^s_{it} && \text{if the signal is informative} \\ &= \eta^s_{it} && \text{if the signal is uninformative.} \end{aligned} \tag{3}$$

Even informative signals, however, are not perfect and $\eta^s \sim \mathcal{N}(0, \sigma_\eta)$ is signal-specific noise. The signals' ability to predict returns depends on the signal-to-noise ratio, $\pi \times \sigma^2_\alpha / \sigma^2_\eta$. Since signals in the real world are correlated (many signals come from data sources such as balance sheet and cash flow statements), we allow for a constant correlation ρ amongst signals. Unless otherwise stated, we set $S = 10{,}000$ and $\pi = 5\%$.

3.1.2 Portfolios and cross-sectional regressions

We follow similar procedure to analyze the simulated data as we do with the real data. Every month we sort stocks into deciles based on signal s and hold these portfolios for one month. The hedge portfolio goes long in decile 10 and short in decile 1. We run a time-series regression of the hedge portfolio returns, R_s, based on signal s, and record the t-statistic of the alpha, t_{α_s}.

$$R_{st} = \alpha_s + \beta'_s F_t + u_{st}. \tag{4}$$

We end up with S portfolios and, hence, S t_{α_s}s. Every period t we also run a univariate cross-sectional regression of risk-adjusted returns on the signal s, as in Brennan, Chordia, and Subrahmanyam (1998):

$$R_{it} - \hat{\beta}'_i F_t = \lambda_{0t} + \lambda_{st} s_{it} + \psi_{it}, \tag{5}$$

and compute the usual Fama and MacBeth (1973) t-statistic, t_{λ_s}. As before, we end up with S t_{λ_s}s.

3.2 Properties of different MHT methods

We first highlight the properties of different MHT methods and make suggestions about their use in the field of finance. Table 3 reports thresholds, rejection rates, the average and the standard deviation of false discovery proportion (FDP), and the average false non-discovery proportion (FNDP). The latter is a measure of power introduced by Genovese and Wasserman (2006); a large E[FNDP] corresponds to lower power.

We tabulate results obtained from applying MHT methods to the alpha t-statistics obtained from our simulations. We adopt $\alpha = 5\%$ statistical significance level, $\gamma = 5\%$ for FDP control, and $\delta = 5\%$ for FDR control. Each entry in the table corresponds to the average or standard deviation over $K = 1{,}000$ repetitions. We can think of each simulated economy as an independent dataset that is used to study the informativeness of the S signals.

The baseline parameters for the simulation are $N = 2{,}000$, $T = 500$, $\pi = 5\%$, $\sigma_\eta = 20\%$, $\sigma_\alpha = 2\%$, $\rho = 0$, and $S = 10{,}000$. Different panels of Table 3 vary one parameter while keeping the others fixed. Panel A varies the signal-to-noise ratio by varying σ_α from 1.5% to 2.5%, Panel B varies the correlation of signals ρ from zero to 20%,[7] Panel C varies the proportion of informative signals π from zero to 30%, and Panel D varies the number of signals S from 5,000 to 500,000. The high end of some of these ranges are not meant to be realistic and are presented only for illustrative purposes. To avoid clutter, we do not discuss each panel separately but rather discuss only the following three conclusions that we draw from the ensemble of the results.

First, FWER control methods are very strict and impose very high statistical thresholds, that deliver very few rejections and poor power when the signal to noise ratio is low (see Panel A). These methods are also not adaptive to changing many of the parameters. For example, increasing the fraction π of informative signals does not change the thresholds (see Panel C). Rather, these methods depend primarily on the number of signals S; higher S (see

[7] Green, Hand, and Zhang (2017) show that the absolute correlation amongst strategies, that are promoted as being very profitable, is only 9%.

Panel D) mechanically increases the thresholds, reduces rejections, and decreases power. Changing correlation (Panel B) does not seem to affect performance of these methods in an appreciable way. We conclude that FWER control methods are less suitable for finance applications that exhibit poor signal-to-noise ratio.

Second, amongst procedures that control FDR and FDP, we notice the following. (a) BHY has the lowest rejections while BH has the highest rejections, with RSW somewhere in between (see Panel A). (b) FDR $\equiv$ E[FDP] is close to 5% for BH but only 0.5% for BHY; thus BHY is too strict. Even though RSW is not designed to control FDR, we find that it controls the FDR to well below 5%. This is to be expected because, as noted in Section 1.3, FDP control at $\gamma = 5\%$ implies an FDR of below 5%. (c) As also mentioned in Section 1.3, one drawback of procedures designed to control FDR is that they do not guarantee that the FDP will be below 5% in *a particular data set*. This is reflected in the relatively high standard deviation (across simulations) of FDP for BH; BHY still manages to have low std[FDP] because of its high thresholds and low rejections (again see Panel A). High correlation of signals creates the biggest problems for BH (see Panel B). (d) Power (= 1−E[FNDP]) is directly associated with the quality of the signal. RSW sits in the middle between BH and BHY in terms of power.

Third, thresholds do not increase mechanically with an increase in signals for all MHT methods (see Panel D). Thresholds do increase for FWER control methods; this fact may have contributed to the erroneously held popular belief that MHT thresholds will keep increasing as the profession keeps discovering more strategies. However, thresholds are relatively insensitive for methods that control FDR and FDP. If anything, higher S leads to a slight gain in efficiency, as evidenced by the lower std[FDP], for these methods. We conclude that (a) having a large number of strategies does not impose any bias on methods that control FDR and FDP, and (b) there is slight efficiency advantage to having more strategies than fewer strategies.

In summary, the evidence presented in this section suggests that MHT methods that

Table 3: MHT properties in simulations

Data are generated as described in the text for N stocks and S signals. Stocks alphas are normally distributed with mean zero and variance σ_α^2. A fraction π signals are informative and σ_η is the noise in the signal. Signals are correlated with constant correlation of ρ. The Bonf(erroni) and Holm methods control FWER, the BH and BHY methods control FDR, and RSW controls FDP. We report thresholds, rejection rates, the average and the standard deviation of false discovery proportion (FDP), and the average false non-discovery proportion (FNDP) for alpha t-statistics, where FDP and FDNP are calculated for each simulation based on the knowledge of informative signals. We use $\alpha = 5\%$ statistical significance level, $\gamma = 5\%$ for FDP control, and $\delta = 5\%$ for FDR control. The statistics below are reported for $K = 1{,}000$ repetitions. Unless otherwise noted, $N = 2{,}000$, $T = 500$, $\pi = 5\%$, $\sigma_\eta = 20\%$, $\sigma_\alpha = 2\%$, $\rho = 0$, and $S = 10{,}000$. Panel A varies the signal-to-noise ratio by changing σ_α, Panel B varies the correlation ρ among signals, Panel C varies the proportion π of informative signals, and Panel D varies the number S of signals.

Panel A: Signal-to-noise ratio						Panel B: Correlation between signals					
σ_α	Bonf	Holm	BH	BHY	RSW	ρ	Bonf	Holm	BH	BHY	RSW
	Thresholds						Thresholds				
1.5	4.56	4.56	3.06	3.76	3.54	0	4.56	4.49	3.02	3.82	3.42
2.0	4.56	4.49	3.02	3.82	3.42	10	4.56	4.50	3.02	3.88	3.48
2.5	4.56	4.44	3.02	4.12	3.42	20	4.56	4.50	3.02	3.95	3.56
	Rejections						Rejections				
1.5	1.8	1.8	4.6	3.4	3.7	0	5.0	5.0	5.2	5.0	5.1
2.0	5.0	5.0	5.2	5.0	5.1	10	5.0	5.0	5.3	5.0	5.1
2.5	5.0	5.0	5.2	5.0	5.1	20	5.0	5.0	5.3	5.0	5.0
	E[FDP]						E[FDP]				
1.5	<0.1	<0.1	4.8	0.5	1.0	0	<0.1	<0.1	4.7	0.5	1.2
2.0	<0.1	<0.1	4.7	0.5	1.2	10	<0.1	<0.1	4.8	0.5	1.0
2.5	<0.1	0.2	4.7	0.5	1.2	20	<0.1	<0.1	4.8	0.5	0.8
	std[FDP]						std[FDP]				
1.5	0.1	0.1	1.0	0.4	0.5	0	<0.1	0.1	1.0	0.3	0.5
2.0	<0.1	0.1	1.0	0.3	0.5	10	<0.1	<0.1	2.3	0.5	0.8
2.5	<0.1	0.1	1.0	0.3	0.5	20	0.1	0.1	4.3	0.8	1.2
	E[FNDP]						E[FNDP]				
1.5	63.9	63.7	12.7	33.0	25.8	0	0.3	0.3	<0.1	<0.1	<0.1
2.0	0.3	0.3	<0.1	<0.1	<0.1	10	0.2	0.2	<0.1	<0.1	<0.1
2.5	0	0	0	0	0	20	0.3	0.3	<0.1	<0.1	<0.1

control FWER do not seem appropriate for finance applications as they lack power and are mechanically affected by the number of hypotheses being tested. Among the other MHT methods, there are reasons to prefer the RSW method. First, this method is in between the

... continued Table 3

Panel C: Proportion of informative signals						Panel D: Number of signals					
π	Bonf	Holm	BH	BHY	RSW	S	Bonf	Holm	BH	BHY	RSW
	Thresholds						Thresholds				
0	4.56	4.49	3.02	3.82	3.42	5,000	4.42	4.38	3.03	4.06	3.47
10	4.56	4.49	2.80	3.54	3.15	10,000	4.56	4.49	3.02	3.82	3.42
20	4.56	4.49	2.57	3.32	2.82	100,000	5.03	5.01	3.01	3.73	3.42
30	4.56	4.45	2.42	3.19	2.60	500,000	5.20	5.20	3.01	3.74	3.41
	Rejections						Rejections				
0	5.0	5.0	5.2	5.0	5.1	5,000	5.0	5.0	5.2	5.0	5.0
10	10.0	10.0	10.5	10.0	10.2	10,000	5.0	5.0	5.2	5.0	5.1
20	19.9	20.0	20.8	20.1	20.4	100,000	4.9	4.9	5.2	5.0	5.1
30	29.9	29.9	31.1	30.1	30.6	500,000	4.9	4.9	5.2	5.0	5.1
	E[FDP]						E[FDP]				
0	<0.1	<0.1	4.7	0.5	1.2	5,000	<0.1	<0.1	4.7	0.5	0.9
10	<0.1	<0.1	4.4	0.5	1.5	10,000	<0.1	<0.1	4.7	0.5	1.2
20	<0.1	<0.1	3.9	0.4	1.8	100,000	<0.1	<0.1	4.7	0.4	1.2
30	<0.1	<0.1	3.5	0.4	2.1	500,000	<0.1	<0.1	4.7	0.4	1.2
	std[FDP]						std[FDP]				
0	<0.1	0.1	1.0	0.3	0.5	5,000	0.1	0.1	1.3	0.5	0.7
10	<0.1	<0.1	0.7	0.2	0.5	10,000	<0.1	0.1	1.0	0.3	0.5
20	<0.1	<0.1	0.4	0.1	0.4	100,000	<0.1	<0.1	0.4	0.1	0.3
30	<0.1	<0.1	0.3	0.1	0.4	500,000	<0.1	<0.1	0.3	0.1	0.3
	E[FNDP]						E[FNDP]				
0	0.3	0.3	<0.1	<0.1	<0.1	5,000	0.2	0.2	0	<0.1	<0.1
10	0.3	0.3	<0.1	<0.1	<0.1	10,000	0.3	0.3	0	<0.1	<0.1
20	0.3	0.2	<0.1	<0.1	<0.1	100,000	1.2	1.1	<0.1	<0.1	<0.1
30	0.3	0.2	0	<0.1	0	500,000	1.8	1.9	<0.1	<0.1	<0.1

two FDR control methods in terms of both size and power. Second, as this method is meant to control the tail behavior of FDP, it avoids the concerns that controlling FDR exposes a researcher to the possibility that the realized FDP varies significantly across applications. Thus, our suggestion is to adopt RSW.

3.3 Dual hurdles

Section 2.4.1 discusses the economic merits of imposing dual hurdles of alpha and FM coefficients. In this section, we explore the statistical advantages to such a procedure. In particular, we compare rejection rates and FDR $\equiv$ E[FDP] obtained from applying CHT and RSW to alpha, FM, and both alpha and FM t-statistics. We set N = 2,000, T = 500, S = 10,000, $\rho = 0$, and $\sigma_\eta = 20\%$. We set π at either 0 or 5%, and vary σ_α from 1.5% to 2.5%.

Table 4: Dual hurdles

Data are generated as described in the text for N stocks and S signals. Stocks alphas are normally distributed with mean zero and variance σ_α^2. A fraction π signals are informative and σ_η is the noise in the signal. Signals are correlated with constant correlation of ρ. We set N = 2,000, T = 500, S = 10,000, $\rho = 0$, and $\sigma_\eta = 20\%$. We set π at either 0 or 5%, and vary σ_α from 1.5 to 2.5%. We report results for CHT (with threshold 1.96) and RSW. We use $\gamma = 5\%$ for FDP control. All rejection rates are for a significance level of $\alpha = 5\%$. We report thresholds, rejection rates, and FDR $\equiv$ E[FDP]. We calculate these statistics separately for alpha, FM coefficient, and jointly for alpha and FM coefficient.

		Threshold	Rejections		E[FDP]	
π	σ_α	RSW	RSW	CHT	RSW	CHT
		Portfolio alpha: t_α				
0	2.0	4.02	<0.1	4.9	—	—
5	1.5	3.54	3.7	9.6	1.1	48.7
5	2.0	3.42	5.1	9.7	1.2	48.2
5	2.5	3.42	5.1	9.7	1.2	48.4
		Fama-MacBeth: t_λ				
0	2.0	4.01	<0.1	5.0	—	—
5	1.5	3.19	5.1	9.8	2.7	48.9
5	2.0	3.19	5.1	9.8	2.7	48.9
5	2.5	3.19	5.1	9.8	2.8	48.9
		t_α and t_λ				
0	2.0	—	<0.1	1.5	—	—
5	1.5	—	3.7	6.3	0.2	22.0
5	2.0	—	5.0	6.4	0.2	21.8
5	2.5	—	5.0	6.4	0.2	21.9

Table 4 shows that the rejection rates for CHT for either alpha or FM are close to 5% for $\pi = 0$ and close to 10% for $\pi = 5\%$. The rejection rates increase with signal-to-noise

ratio. These numbers are to be expected as CHT rejects 5% of the strategies that should be rejected (assuming very good power) but also reject 5% of the strategies from the null (false rejections). Imposing a dual hurdle reduces the proportion of rejections to almost half. CHT still rejects 1.5% strategies with $\pi = 0$ and about 6% strategies with $\pi = 5\%$. As with the real data, imposing dual hurdle helps with proportion of rejections. A similar story emerges when looking at rejection rates from RSW. The rejection rates for either alpha or FM are low to begin with (lower than 5%) but a dual hurdle reduces the rejection rates further.

A sharper picture is obtained by looking at FDR≡E[FDP]. Recall that FDR is the proportion of rejections that are false (meaning that the signal is uninformative under the null in the population). Since CHT rejects roughly twice the strategies than π, the FDR under CHT is close to 50% for either alpha or FM. Imposing the dual hurdle reduces the FDR by half. Thus, imposing a dual hurdle helps in curbing the false rejection rate to about 25% for CHT methods. This may still be deemed to be too high in some research applications. If improvement in size is what is desired, this is precisely where the MHT helps. The dual hurdle reduces the FDR to about 0.2% when asking strategies to have t-statistics that cross both alpha and FM thresholds. In other words, MHT is designed to limit false rejections but permits some false rejections (to improve power). Imposition of dual hurdles reduces the chances of false rejections further.

This further reduction of FDR might not be desirable, however. A researcher might want to directly control the FDR (or FDP) of the dual hurdle procedure at a particular level. Unfortunately, we are not aware of any statistical procedure that guarantees an ex-ante FDP control for a dual hurdle test. Nevertheless, given the importance of this issue, we offer some heuristics of application of dual hurdle thresholds in Section 4.2.1.

The results from this section demonstrate that often-followed practice of dual hurdles in finance helps in reducing the probability of false rejections for all testing methods. Coupled with MHT methods, it makes for a very effective tool to guard against false rejections.

4. Researcher versus econometrician

The results so far show that applying the RSW procedure to the set of randomly generated signals, $\mathcal{S}^{\mathcal{E}}$, produces CNM rejections of 97.9%. On the other hand, Harvey, Liu, and Zhu's (2016) report that applying the BHY procedure to the set of 296 published factors, produces a CNM rejection rate of 27% (i.e., 80 out of 296).[8,9]

We propose a indirect method to recover the proportion of false rejections by using both CNM rejection estimates. Our plan, in a nutshell, is to infer the statistical properties of the set $\mathcal{S}^{\mathcal{R}}$ of strategies that the researchers study. With that in hand, we can accomplish both the normative and positive goals of the paper: quantify the proportion of false rejections and propose MHT adjusted thresholds.

We start by specifying the different sampling scheme used by the econometrician and the researchers. The econometrician uses a signal s drawn from the set $\mathcal{S}^{\mathcal{E}}$, where a fraction π of signals has the ability to predict returns. We assume that researchers adopt a sophisticated sampling scheme that gives them an advantage over the econometrician. In particular, they draw informative signals with probability $\Omega \times \pi$, where $\Omega > 1$ quantifies the advantage that the researchers have over the econometricians.

The modeling choice is meant to capture the idea that a researcher's superior ability comes from the fact that they only sample signals for which they can provide a rationale, by means of a theoretical model or a strong narrative. Consequently, the set $\mathcal{S}^{\mathcal{R}}$ will contain more informative signals and fewer uninformative signals than $\mathcal{S}^{\mathcal{E}}$. Nonetheless, we allow for the possibility that, given the nature of research, sometimes a false narrative can also be

[8]Harvey, Liu, and Zhu (2016) do calculate higher CNM rejection rates using methods that control FWER, such as Bonferroni and Holm. However, as we have shown in Section 3.2, these methods are very conservative and suffer from poor power. Harvey, Liu, and Zhu (in their Appendix A) also estimate missed significant factors with t-statistics between 1.96 and 2.57. Using this enhanced sample, they report higher MHT thresholds presumably implying higher CNM rejections. Please see Section 4.2 for a discussion of how our exercise relates to theirs.

[9]On the other hand, using a similar laboratory of over 18,000 randomly generated signals, Yan and Zheng (2017) find that a vast majority of them can be rejected using bootstrap method of Fama and French (2010), but not a formal MHT procedure. Harvey and Liu (2019) show that the bootstrap methods have severe size biases that leads to large over-rejections.

provided for signals that come from the null of no predictability; researchers sample these signals with probability $(1 - \Omega \times \pi)$. Researchers apply a simple rule and circulate a report only for signals for which they can compute a t-statistic that is larger than a threshold, say 1.96. The set of strategies that are published, $\mathcal{S}^{\mathcal{P}} = \{s \in \mathcal{S}^{\mathcal{R}},\ t_s > 1.96\}$, is the simulation counterpart of the set of signals analyzed by Harvey, Liu, and Zhu (2016).

The correct threshold and the proportion of false rejections can be computed using the set $\mathcal{S}^{\mathcal{R}}$, and the knowledge of which signals are informative and which are not (precise details are in the next section). Please note the issue here is not that the *number* of strategies in the sets $\mathcal{S}^{\mathcal{E}}$, $\mathcal{S}^{\mathcal{R}}$, and $\mathcal{S}^{\mathcal{P}}$ is different; our simulations from Section 3.2 show that the number of strategies has a negligible role in determining thresholds and number of rejections. We use the statistical framework characterized by equations (2) through (5) to solve this problem. We fix some parameters of the model that can be directly observed in the data, and calibrate those that are unobservable by asking the model to match some observable quantities that are informative about the said parameters. With all the parameters in hand, we then calculate the MHT adjusted threshold and proportion of false rejections in the set $\mathcal{S}^{\mathcal{R}}$. We describe the calibration details next. We undertake robustness analysis on some of the simplifying assumptions, such as normality, in Section 5.

4.1 Calibration

Calibrated parameters: In order to minimize the uncertainty in the calibration, we only calibrate parameters that are unobservable and set the rest to quantities that are either directly measurable in the data or for which there are close estimates in the literature. We calibrate parameters related to the data generating process for stocks and signals in equations (2) and (3) and the researchers' advantage over econometricians in picking informative signals. Specifically, we calibrate σ_η, π, and Ω.

Fixed parameters: The distributions of factors, betas, and residuals, that define equation (2) are matched to their respective counterparts in the real data sample, as discussed in

Section 3. We also fix the standard deviation of the cross-sectional distribution of alphas, σ_α, to average cross-sectional standard deviation of the realized stock alphas, 1.9%. To compute this quantity in the real data we proceed as follows: for each stock, we run ninety-month rolling window time-series regressions of the stock excess returns on the six-factor model, where we require at least sixty observations. We then compute the cross-sectional standard deviation of the stocks' alphas for each time period, and finally average across time. As there is only limited information about the correlation structure between strategies, we fix ρ to zero (we return to this issue in the robustness checks presented in the Section 5.1).

Target quantities: We choose target quantities that are the most responsible for pinning down the parameters to be calibrated. While all parameters affect all target quantities, the connection between parameters and target quantities simply refers to which target quantity is affected the most by changing one particular parameter.

The fraction of informative signals π essentially defines the set $\mathcal{S}^{\mathcal{E}}$. We calibrate π by asking the simulation to produce the same fraction of CNM discoveries implied by the RSW thresholds that we observe in the data (i.e., CNM rejections in $\mathcal{S}^{\mathcal{E}} = 97.9\%$).

The factor Ω that defines the superior ability of researchers (relative to the econometrician) to draw informative signals is calibrated by asking the simulation to produce the same proportion of CNM rejections in the set $\mathcal{S}^{\mathcal{P}}$ (i.e., CNM rejections in $\mathcal{S}^{\mathcal{P}} = 27.0\%$).

We calibrate σ_η by asking the simulation to match two quantities: the average of cross-sectional standard deviations of residuals from the cross-sectional FM regressions in equation (5) (i.e., StDev(ψ) = 17.2%), and the ratio of (significant) portfolio alphas to (significant) stock alphas. Intuitively, the residuals from the FM regressions are inversely related to the ability of signals to explain the cross-section of returns. For each signal, we first run the FM regression as specified in equation (5) and obtain StDev(ψ) by averaging it across signals.[10]

The noise in signals also directly affects the ability to convert stock alphas into profitable

[10]There is a slight mismatch in the computation of StDev(ψ) for the data and the simulation. This quantity for the data is obtained from equation (1) but obtained from equation (5) for the simulations as we do not have the control variables such as size, book-to-market ratio, past returns, etc., for the simulations.

trading strategies. The more noise there is in the (informative) signals, the harder it is to construct trading strategies with a large alpha. We construct a measure of "conversion" of stock alphas to portfolio alphas as the ratio of significant portfolio alphas to significant stock alphas, $E(|\text{sign. } \alpha_s|)/E(|\text{sign. } \alpha_i|)$. For stocks, we consider only alphas that are significant at conventional levels. Thus, $E(|\text{sign. } \alpha_i|)$ is simply $E(|\alpha_i||t_{\alpha_i}| > 1.96)$. For portfolios, consider only those strategies for which both alpha and FM coefficient are significant. Thus, $E(|\text{sign. } \alpha_s|)$ is $E(|\alpha_s||t_\alpha| > 1.96 \text{ and } |t_\lambda| > 1.96))$. Note here that the requirement of statistical significance is used only as a device to isolate the tails of the distributions (since the average alphas for both stocks and portfolios are close to zero in the simulation and in real data). This quantity is computed to be 0.12 (= 0.47%/3.8%) in the real data.

Procedure: In the simulated data, we follow a similar procedure for target quantities, wherein we further average across the 1,000 simulations. We use a global minimization algorithm (particle swarm of Kennedy and Eberhart 1995) to find the vector of parameters that minimizes the sum of the squared percentage distance of the target quantities.

Outputs: After obtaining the calibrated parameters, we run 1,000 simulations of the set $\mathcal{S}^{\mathcal{R}}$. In each simulation we compute the ratio of uninformative signals for which both the alpha and FM coefficient t-statistics are larger than 1.96 (i.e., false rejections) relative to the total number of signals that are rejected by CHT. We report the proportion of false rejections as the average of this number across the 1,000 simulations. As mentioned above, we can compute false rejections only because we know exactly which signals are informative and which are uninformative. Similarly, for each simulation we compute the RSW thresholds for alpha and FM coefficient t-statistics and report their averages across simulations.

We tabulate results of the calibration in Table 5. Panel A shows the calibrated parameters, Panel B compares the target parameters obtained from the simulation to the respective figures from the data, and Panel C reports the MHT adjusted thresholds and the proportion of false discoveries in the set $\mathcal{S}^{\mathcal{R}}$.

Overall the simulated economy is very close to the actual data. For instance, 96.5% of the

Table 5: False rejection rates and adjusted MHT thresholds

We use a global minimization algorithm (particle swarm) to calibrate the statistical framework presented in Section 4. We find the vector of parameters that minimizes the sum of the squared percentage distance of the target quantities. In Panel A we report the calibrated parameters (σ_η and π are reported in percentages). In Panel B, we compare the target quantities obtained from the simulation to the respective quantities from the data. In Panel C we report the adjusted MHT thresholds and the proportion of CNM and false discoveries in the set $\mathcal{S}^{\mathcal{R}}$.

Panel A: Calibrated parameters		
σ_η	π	Ω
4.8	0.06	29.0

Panel B: Target quantities		
	Data	Simulation
CNM rejections in $\mathcal{S}^{\mathcal{E}}$	97.9	96.5
StDev(ψ) in $\mathcal{S}^{\mathcal{E}}$	17.2	17.5
$E(\lvert\text{sign. } \alpha_s\rvert)/E(\lvert\text{sign. } \alpha_i\rvert)$ in $\mathcal{S}^{\mathcal{E}}$	0.12	0.12
CNM rejections in $\mathcal{S}^{\mathcal{P}}$	27.0	30.9

Panel C: Outcome quantities in the set $\mathcal{S}^{\mathcal{R}}$	
Alpha threshold, $\mathcal{T}_\alpha$	3.84
FM threshold, $\mathcal{T}_\lambda$	3.38
CNM rejections	42.5
False rejections	45.3

signals in the econometrician set $\mathcal{S}^{\mathcal{E}}$ are deemed as CNM rejections according to the RSW threshold relative to 97.9% in the real data. Similarly, the proportion of CNM rejections in the published research set $\mathcal{S}^{\mathcal{P}}$ is 30.9%, which is close to 27.0% in the actual data. The residuals from the FM regressions have very similar distributions (standard deviation of 17.5% compared to 17.2% in the data). Finally the "conversion" ratio of signals from stock to portfolios is the same at 0.12.

Moving to the calibrated parameters, we find that the econometrician set $\mathcal{S}^{\mathcal{E}}$ is endowed with 0.06% of informative signals; this low number corresponds to the intuitive notion that an overwhelming majority of the random strategies that we construct are uninformative. In order for the econometrician (researcher) to estimate a proportion of lucky discoveries of 96.5% (30.9%), we find that the researcher must be 29 times better at drawing informative

signals than the econometrician. Thus, our model sustains the hypothesis that researchers do not simply mine the data but are engaged in the process of discovering informative signals.

Ultimately, the calibration allows us to characterize the researcher set $\mathcal{S}^{\mathcal{R}}$, the proportion of false rejections under CHT, and MHT thresholds based on the RSW procedure. We find a proportion of false rejections of 45.3%, and thresholds of 3.84 and 3.38 for alpha and FM coefficient t-statistics, respectively.[11]

4.2 Discussion

4.2.1 Thresholds

Our threshold estimates quantitatively confirm the intuition of Harvey, Liu, and Zhu (2016) that accounting for not only the right-tail of the strategies in the published set $\mathcal{S}^{\mathcal{P}}$ and also for strategies that do not reach statistical significance in the researcher set $\mathcal{S}^{\mathcal{R}}$ would lead to thresholds higher than three.[12]

We note that we derive our thresholds by independently applying the RSW procedure to the alpha and FM coefficient t-statistics. However, we use (and suggest that researchers also do so) *both* thresholds to establish which signals are significant. Thus, we effectively create a dual hurdle that trading signals have to surpass. While our statistical framework is cast in a Frequentist setting, one can view this dual hurdle in a Bayesian setting (Harvey 2017) as having more skeptical priors than those for a single hurdle. The dual threshold is also very much in the spirit of other procedures that researchers often adopt to reduce the likelihood of data-mining, such as testing the same hypotheses with international data.

To consider the effects of such a dual hurdle, we start by looking at the realized FDR

[11] We also refine the set of stocks used to construct the strategies. It is widely known that anomalies are stronger in small-cap and micro-cap stocks (Fama and French 2008). At portfolio formation, we exclude all stocks with a price less than three dollars and a market capitalization lower than the 20th percentile of the NYSE distribution. This filter also alleviates concerns about transaction costs as well those about generalizability and relevance (Novy-Marx and Velikov 2016). We find that the proportion of false rejections in this set is 41.7%. Details of these results are available upon request.

[12] Harvey, Liu, and Zhu (2016, p.38) report that accounting for missing factors increases the 5% BHY threshold from 2.78 to 3.18.

(average FDP across simulations). As mentioned before, since FDP control is stricter than FDR control, we expect the realized FDR to be lower than the 5% FDP control. We do find that, using the baseline parameters from Table 5, the realized FDR is 0.8% for t_α, 2.2% for t_λ, and 0.2% for the dual hurdles.

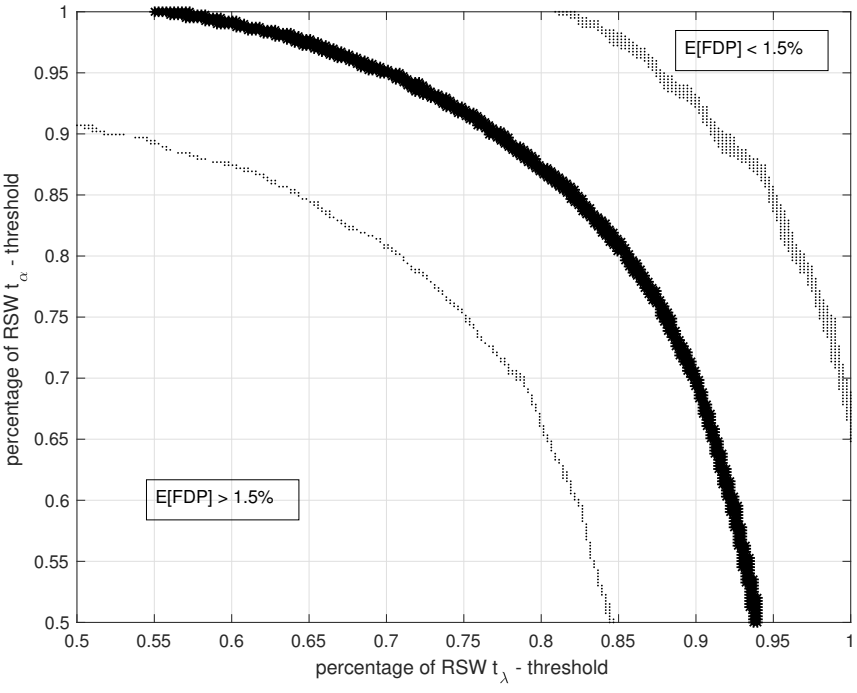
While the extremely low FDR of the RSW method achieves the goal of reducing the fraction of false discoveries, it also implies an FDP control of dual hurdles that is lower than what a researcher might desire (for example, the control that was used to derive the individual thresholds). Ideally, we would prefer an MHT procedure that ensures an ex-ante specified control from the dual hurdles. Absent such a procedure, we offer some heuristics on how to adjust the thresholds to achieve an arbitrary goal. To keep things simple, say that we aim for a realized FDR of 1.5% where 1.5% is the average realized FDR obtained by individually applying the time-series and cross-sectional thresholds.

Figure 2 shows the combination of thresholds for which the realized FDR is 1.5%. The solid black line is for the average (across simulations) and the two light lines show two-standard-deviations bounds around the average. Below the line the realized FDR is more than 1.5%, above the line it is less than 1.5%. The two axes are presented in terms of percentage of the thresholds that we report in Table 5, so that the top right corner represents the realized FDR of 0.2% corresponding to the two thresholds of 3.84 and 3.34. There are many combinations that produce a realized FDR of 1.5%: for example, a dual threshold of 3.07 (80% of 3.84 for t_α) and 2.88 (85% of 3.34 for t_λ). We leave a more explicit analysis of these issues to future research.

Finally, we note that thresholds obtained by applying any MHT procedure only correct the multiplicity problem that naturally arises in the research process. These higher thresholds may not be helpful in aligning researchers' personal incentives with the scientific discovery process. Indeed, it is possible that advocating for higher thresholds than those prescribed by CHT might have the unintended consequence of increased data mining. We refer readers to Harvey (2017) for a detailed discussion of this problem and some possible solutions.

Figure 2: FDR from application of dual RSW thresholds

Using the statistical framework of Section 4 and the baseline parameters from Table 5, we calculate the FDR after application of dual hurdles on alpha and FM coefficient t-statistics. The solid black line reports the RSW thresholds for the two t-statistics that attain an average FDR of 1.5% (across simulations). The two light lines show two-standard-deviations bounds around the average. Below the solid line, the realized FDR is more than 1.5%, above the line it is less than 1.5%. The two axes are presented in terms of percentage of the t-statistic thresholds reported in Table 5; the top-right corner cor



4.2.2 False rejections

Since the 45.3% estimate is obtained by averaging across the 1,000 simulations, we can calculate how the simulation error affects the estimate by considering the standard deviation of false rejection percentage across simulations. We find that our estimate has a pretty narrow distribution with a standard deviation of 2.1%. In other words, we can statistically confidently say that our estimate is different from that obtained from the set $\mathcal{S}^{\mathcal{P}}$.

We have noted in the introduction that, in general, the false rejections in the set $\mathcal{S}^{\mathcal{R}}$ may not be close to CNM rejections because of power problems (short time history of trading strategies and/or low signal-to-noise ratio). These issues hinder inferences drawn from application of MHT to real data (see, for example, Andrikogiannopoulou and Papakonstantinous 2019 in the context of mutual fund data). Our simulation sidesteps this issue by design as

we have a long time-series (500 observations).

Our calibration implies that, without any correction for the multiple testing problem, the expected proportion of false rejections would be 45.3% going forward (assuming that the statistical distribution of individual stock returns does not change; researchers do not improve their ability in discovering trading strategies; and MHT does not induce more data mining). In essence, if our model is an accurate representation of researchers' ability to draw informative signals, we should expect to continue to discover abnormally profitable trading strategies with associated t-statistics larger than 1.96. Some of these strategies would be false. If researchers were to adopt MHT thresholds, the percentage of false rejections would be much smaller; possibly as small as the size of the FDP control specified by the MHT procedure.

4.2.3 Relation to out-of-sample

One alternative approach in the literature to estimate false rejections is out-of-sample studies. If a signal is true under the null (does not produce an alpha), then the signal's predictive ability should be weak to non-existent in an out-of-sample period. This is the viewpoint adopted by Linnainmaa and Roberts (2018), who study the performance of trading strategies in the years before the original study that discovered them.[13] They report that, of the 36 strategies that they study, 20 strategies have insignificant Fama and French (1993) three-factor alphas in the pre-discovery period. This represents a false rejection rate of 55.6%. There are obvious differences in the sample period, number of strategies, and performance evaluation method between the two studies but the two estimates are in the same ballpark. Nevertheless, we return to the issue of different estimates obtained by considering different proportion of lucky rejections in the set $\mathcal{S}^{\mathcal{P}}$ in Section 5.2.

[13]McLean and Pontiff (2016) also study out-of-sample performance of trading strategies. They report a deterioration of performance after the publication of the strategies. Since this decline could be due to an increase in arbitrage activity, their results speak less to our study of false rejections.

4.2.4 Relation to literature

As mentioned several times, one of the main difficulties in the application of MHT is the unobservability of the set $\mathcal{S}^{\mathcal{R}}$. Our approach to uncovering the properties of this set is to combine information from the observed sets $\mathcal{S}^{\mathcal{E}}$ and $\mathcal{S}^{\mathcal{P}}$.

Harevey, Liu, and Zhu (2016) also provide an estimate of the set of strategies in the set $\mathcal{S}^{\mathcal{R}}$ by fitting a truncated exponential distribution to the tails of the published strategies in the set $\mathcal{S}^{\mathcal{P}}$. They allow for estimation error in this fitted exponential distribution as well as for the possibility that they may not have captured all the strategies that have been successfully tried. The main difference between our approach and theirs is that instead of matching t-statistics corresponding to a certain frequency (i.e., a certain quantile), we match the frequencies corresponding to certain t-statistics (i.e., the fraction of t-statistics larger than a threshold). Thus, our approach to fitting CNM rejections is akin to fitting the tails of the t-statistic distribution.

The advantage of our approach is that we do not have to necessarily specify the details of the publication process and we avoid under-sampling insignificant strategies. The limitation of our approach is that the shape of the distribution of t-statistics in our simulation may not match the shape of the observed truncated distribution (we match only the fractions of t-statistics larger than 1.96 and larger than the RSW threshold).

Chen (2019) also estimates the truncated part of the distribution of t-statistics (that is present in the set $\mathcal{S}^{\mathcal{R}}$ but not in the set $\mathcal{S}^{\mathcal{P}}$). He fits a variety of models and reports that t-statistic thresholds vary a lot depending on the model. One counter-intuitive aspect is that Chen sometimes finds that researchers never sample strategies from the null of no alpha. Translated into our framework, this means that Ω (researchers superior ability) is so high that, in the limit, $\Omega \times \pi$ approaches one (Chen's p_0 is close to zero). If the researcher never samples from the null, then, of course, there is no reason for any statistical testing (classical or multiple). In such a situation, the t-statistic threshold would be very low, and false rejections would be close to zero, as Chen (2019) and Chen and Zimmermann (2018)

report. It is useful to note, however, that even Chen shows that, in situations where his estimated p_0 is further away from zero, the adjusted MHT thresholds are higher than 1.96 and that FDR is much higher than 5%.

We believe $p_0 \approx 0$ is not an accurate description of reality. We know from McLean and Pontiff (2016) and Linnainmaa and Roberts (2018) that several strategies do not survive out-of-sample tests. Further, while Hou, Xue, and Zhang (2019) is not a truly out-of-sample study, they too find a large fraction of false rejections. There are also biases in the publication process as highlighted by Harvey (2017). Finally, some of the research in anomalies is not predicated on theory, at least not on ex-ante theory.

Our results imply that researchers draw uninformative signals 98% ($1 - \Omega \times \pi \approx 0.98$) of the time. In other words, p_0 is close to 0.98 in our estimates while Chen (2019) suggests that it is much lower. Our high estimate of p_0 is a direct consequence of the very low π. Which of these two estimates is closer to truth might depend on one's priors. Given that our paper is concerned with abnormally profitable strategies, we think that a reasonable prior is based on what we know about the performance of industry practitioners (which we understand to be low).

5. Robustness checks

Our estimation of false rejections depends critically on the assumptions underlying our statistical framework that allows us to characterize the set of strategies studied by the researcher, and on the two estimates of CNM rejections (27.0% and 97.9%) that we use. We next conduct a variety of robustness tests to gauge the variation in thresholds and false rejection rates to our modeling choices.

5.1 Alternative assumptions about stock returns and signals

We first consider variations to the assumptions about the data generating process described in Section 4. In particular, randomness in the statistical framework is generated from normal distributions. In practice, we know that stock returns exhibit negative skewness and fat tails, and that some of the factors are prone to extremely negative returns (for example, momentum crashes; see Daniel and Moskowitz 2016). We also know that the published strategies show some, albeit weak, cross correlation (for example, Green, Hand, and Zhang 2017 report an average pairwise correlation of 9%). We start with the assumptions about stock returns. Equation (2) describes three sources of randomness in stock returns: alphas, factor returns, and residuals.

We first estimate the empirical distribution of the cross-section of alphas in the data. We follow the same procedure of rolling time-series regression, as in Section 4, and compute the time-series averages of the cross-sectional standard deviation, skewness, and kurtosis of estimated stock's alphas. We find skewness of 1.17 and kurtosis of 18.01. Accordingly, we modify the alpha generating process from a normal distribution to a Pearson system with mean zero, standard deviation 1.9%, skewness of 1.17, and kurtosis of 18.01 (see Johnson, Kotz, and Balakrishnan 1994). Results of the calibration are reported in Table 6 under the column entitled 'Alpha.' While some of the parameters slightly change, we obtain a proportion of false rejections of 45.9%, which is close to the baseline scenario estimate of 45.3%. The alpha and FM coefficient t-statistic thresholds also do not change much.

Second, instead of simulating factor realizations from a multivariate normal distribution, we bootstrap factor returns by resampling them in stationary blocks with replacement (Politis and Romano 1994) directly from the observed six-factor monthly returns. The bootstrap preserves the higher-order moments and factor crashes observed in the data (and is easier to implement than a simulated system of factors that preserves time-series properties of data counterparts). We find a proportion of false rejections of 46.3%, but higher thresholds at 4.41 and 3.50, for alpha and FM coefficient t-statistics, respectively.

Table 6: Alternative assumptions about stock returns and signals

We use a global minimization algorithm (particle swarm) to calibrate the statistical framework presented in Section 4. We find the vector of parameters that minimizes the sum of the squared percentage distance of the target quantities. We consider four modifications of the basic assumptions: alternative distribution for alphas, factors and residuals, and signal cross-correlation of 9% (as opposed to zero). In Panel A we report the calibrated parameters. In Panel B, we compare the target quantities obtained from the simulation to the respective quantities from the data. In Panel C we report the adjusted MHT thresholds and the proportion of false discoveries.

		Alpha	Factors	Residuals	$\rho = 9\%$
		Panel A: Calibrated parameters			
σ_η		4.2	9.7	7.6	6.2
π		0.06	0.06	0.06	0.08
Ω		28.1	22.1	29.1	22.5
		Panel B: Target quantities			
	Data	Simulation			
CNM rejections in $\mathcal{S}^{\mathcal{E}}$	97.9	96.5	95.8	96.3	95.6
StDev(ψ) in $\mathcal{S}^{\mathcal{E}}$	17.2	17.7	17.5	14.7	17.5
$E(\lvert\text{sign. } \alpha_s\rvert)/E(\lvert\text{sign. } \alpha_i\rvert)$ in $\mathcal{S}^{\mathcal{E}}$	0.12	0.10	0.13	0.10	0.12
CNM rejections in $\mathcal{S}^{\mathcal{P}}$	27.0	30.0	30.4	28.7	31.2
		Panel C: Outcome quantities in the set $\mathcal{S}^{\mathcal{R}}$			
Alpha threshold, $\mathcal{T}_\alpha$		3.84	4.41	3.82	3.79
FM threshold, $\mathcal{T}_\lambda$		3.34	3.50	3.40	3.29
CNM rejections		43.3	44.9	41.2	42.6
False rejections		45.9	46.3	42.8	45.5

Next, we modify the properties of residuals in Equation (2). We first compute the cross-sectional average of residuals standard deviations, skewness, and kurtosis (again using rolling regressions for each stock) in the real data. We then modify our simulation so that residuals are drawn from a distribution with standard deviation of 14.7%, skewness of 0.83, and kurtosis of 6.38. Re-calibrating, we obtain a proportion of false rejections of 42.8%, with critical values of 3.82 and 3.40, for alpha and FM coefficient t-statistics, respectively.

Finally, relying on Green, Hand, and Zhang (2017), we fix the cross-sectional correlation, ρ, between strategies at 9% and repeat the calibration. We find a proportion of false rejections of 45.5% and associated thresholds that are close to the baseline specification at 3.79 and

3.29, for alpha and FM coefficient t-statistics, respectively.

None of these modifications affect our estimate of false rejections in a meaningful way for the following reason. Even though our calibration attempts to match four different quantities in Table 5, the two main drivers are the CNM rejections in the sets $\mathcal{S}^{\mathcal{E}}$ and $\mathcal{S}^{\mathcal{P}}$ (i.e., 97.9% and 27.0%, respectively). Therefore, while some calibrated parameters change, the false rejection proportion is more tightly linked to the two CNM rejections, and relatively independent of the exact specification of randomness in the simulation.

The robustness experiments, thus, highlight a limitation of our results: the tight link between the two CNM rejection fractions and the calibration outputs. Lower (higher) CNM rejections in either set will lower (raise) our estimate of false rejections. For instance, if the proportion of CNM rejections in $\mathcal{S}^{\mathcal{P}}$ was much lower (as, for example, argued by Chen and Zimmermann 2018), then, this would intuitively imply that the researcher is very good at screening out null strategies (high Ω in our setup) and, therefore, the false rejection fraction in the set $\mathcal{S}^{\mathcal{R}}$ will also be low. The opposite would be obtained with a high proportion of CNM rejections in $\mathcal{S}^{\mathcal{P}}$. We explore the sensitivity of our results to the two estimates of CNM rejections in the next two subsections.

5.2 Alternative definition of CNM rejections in $\mathcal{S}^{\mathcal{P}}$

One essential element of the calibration discussed in section 4.1 is the estimate of CNM rejection in the set of public strategies. We have used the 27% estimate from Harvey, Liu, and Zhu (2016). While this estimate has the advantage that it is constructed by considering the pristine set of discoveries as they are presented by the original authors, it has the drawback that such discoveries were not based on the same set of procedures.

As a possible alternative, we consider a measure constructed by applying the same performance evaluation procedure to a set of public strategies. In particular, we consider the portfolio returns constructed using 156 signals that are available from Andrew Chen's website (see Chen and Zimmermann 2018). This alternative is not free of its own set of problems.

For example, as shown by Green, Hand, and Zhang (2017) and Hou, Xue, and Zhang (2019), there is no guarantee that in a replication sample, and by changing the evaluation method, the signals would still produce a profitable and statistically significant strategy. Therefore, our replication procedure introduces a bias against discoveries by artificially imposing a threshold for publication that is not completely representative of the true publication process. Moreover, even if a signal was profitable in the earlier sample period, it may become unprofitable later because of investor learning (McLean and Pontiff 2016) and will be deemed unpublishable in our replication universe. The contamination of in-sample (relative to the original study) with out-of-sample data, and the use of a different factor model in the evaluation produces non-trivial consequences for the 156 trading strategies. Only 77 strategies have statistically significant average returns, and of those 57 have statistically significant alphas using the Fama and French (2015) five factors plus the momentum factor. Thus, the set of t-statistics that we can derive from this sample is not as representative of published signals as the one used by Harvey, Liu, and Zhu (2016).

Remaining cognizant of these problems, we compute a CNM rejection rate for those 57 strategies using the same BHY at 5% procedure as that used by Harvey, Liu, and Zhu (2016). We obtain a CNM rejection rate of 50.6%. We repeat the calibration exercise with the new estimate of CNM rejections in $\mathcal{S}^{\mathcal{P}} = 50.6\%$ and report the results in Table 7.

The most immediate difference, relative to the baseline case, is that to allow a higher CNM rejection rate in the set $\mathcal{S}^{\mathcal{P}}$, the researcher must not be as skilled at drawing informative signals. Accordingly, the relative advantage of researcher over the econometrician, Ω, declines from 29.0 in the baseline case to only 13.7 in this experiment. Lower Ω, in turn, implies that there are fewer informative signals in the $\mathcal{S}^{\mathcal{R}}$. Combined with a slight increase in the signal to noise ratio, σ_η, leads to an increase in thresholds (i.e., from 3.83 to 4.09 for t_α, and from 3.31 to 3.52 for t_λ) and, finally, an increase in the proportion of false rejections to 64.9%. Thus, the impact of higher proportion of CNM rejections in the set of public signals on the estimate of the proportion of false rejections is quantitatively important.

Table 7: Alternative definition of CNM rejections in $\mathcal{S}^{\mathcal{P}}$

We use a global minimization algorithm (particle swarm) to calibrate the statistical framework presented in Section 4. We find the vector of parameters that minimizes the sum of the squared percentage distance of the target quantities. We consider an alternative to the proportion of CNM rejections in $\mathcal{S}^{\mathcal{P}}$ provided by Harvey, Liu, and Zhu (2016): we take 156 strategies studied by Chen and Zimmerman (2018) and calculate CNM rejection rate in this set. In Panel A we report the calibrated parameters. In Panel B, we compare the target quantities obtained from the simulation to the respective quantities from the data. In Panel C we report the adjusted MHT thresholds and the proportion of false discoveries.

Panel A: Calibrated parameters		
σ_η	π	Ω
3.8	0.06	13.7

Panel B: Target quantities		
	Data	Simulation
CNM rejections in $\mathcal{S}^{\mathcal{E}}$	97.9	96.8
CNM rejections in $\mathcal{S}^{\mathcal{P}}$	50.6	48.5
StDev(ψ) in $\mathcal{S}^{\mathcal{E}}$	17.2	17.5
$E(\lvert\text{sign. } \alpha_s\rvert)/E(\lvert\text{sign. } \alpha_i\rvert)$ in $\mathcal{S}^{\mathcal{E}}$	0.12	0.13

Panel C: Outcome quantities in the set $\mathcal{S}^{\mathcal{R}}$	
Alpha threshold, $\mathcal{T}_\alpha$	4.08
FM threshold, $\mathcal{T}_\lambda$	3.51
CNM rejections	63.1
False rejections	64.9

5.3 Alternative specification of CNM rejections in $\mathcal{S}^{\mathcal{E}}$

5.3.1 Different factor models

One way to consider the impact of alternative measures of CNM rejections in the set $\mathcal{S}^{\mathcal{E}}$ is to use different factor models. This exercise is also useful as a robustness check on our subjective choice of considering the Fama and French (2015) five-factor model augmented by momentum factor as the baseline factor model.

We consider four substitutes: CAPM, FF3, BS, and HXZ. CAPM is a one factor model with the market factor. FF3 is the Fama and French (1993) three-factor model. BS is the Barillas and Shanken (2018) six-factor model. HXZ is the Hou, Xue and Zhang (2015) q-

model. For each trading strategy, we run a time-series regression of the corresponding hedge portfolio returns on the factors and obtain the alpha as well as its heteroskedasticty-adjusted t-statistic, t_α. The additional control variables in FM regressions correspond to the factor model used. Thus, we do not include any other controls when risk adjusting stock returns by the excess market return. We include size and book-to-market when adjusting stock returns using the FF3 model. In all the other cases, we include size, book-to-market, profitability, asset growth, and one- and 11-month lagged returns. Summary statistics for alpha and FM t-statistics for these factor models are shown in Panels A1 and A2 of Appendix Table A2.

Table 8 shows the results of applying CHT and MHT for each factor model. We find quite some variation in rejection rates across factor models in the randomly generated sample of over two million signals. For example, HXZ model rejects the fewest (22.3%), while the BS model rejects the most (35.7%) of the null of zero alphas at conventional statistical levels. Rejection rates for FM coefficient are more similar across models. Using FM regressions, the FF3 model has the most rejections (8.3%) and the HXZ model has the least (4.9%). After applying the RSW thresholds, rejections are very low and all below 1%.[14]

Detailed results of the calibration of our statistical model for the four factor model alternatives are reported in Appendix Table A3. In Table 8, we report only the RSW thresholds and the proportion of false rejections obtained from the researcher set $\mathcal{S}^{\mathcal{R}}$. Thresholds are all very close to the figures reported in Table 5 across the four possible models.

False rejection percentages vary between 40.1% and 46.5%, with CAPM producing the lowest estimate and BS the highest. Although there are other reasons that contribute to the estimate of false rejections, everything else being equal, a higher CNM rejection rate in the set of randomly generated signals, $\mathcal{S}^{\mathcal{E}}$, pushes the magnitude of the estimate of false rejections in the set $\mathcal{S}^{\mathcal{R}}$ higher.

[14]Taken literally, a fraction close to 0% means that, in our sample of randomly generated strategies, any MHT rejection is likely false when evaluated by the HXZ model. To put it differently, the HXZ model remarkably explains returns of almost all randomly generated returns.

Table 8: Alternative factor models

The table reports CHT rejections, RSW rejection rates and proportion of CNM rejections for the full set of 2,393,641 signals and all stocks, as well as RSW thresholds and proportion of false rejections obtained from the calibration (see Appendix Table A3). The CAPM uses the market factor. FF3 is the Fama and French (1993) three-factor model. BS is the Barillas and Shanken (2015) six-factor model. HXZ is the Hou, Xue and Zhang (2015) q-model augmented with the momentum factor. In FM regressions, we do not include any other control when risk adjusting stock returns CAPM. We include size and book-to-market when adjusting stock returns using the FF3 model. In all the other cases, we include size, book-to-market, profitability, asset growth, and one- and twelve-month lagged returns. The significance level is 5% for all tests and we use only the RSW method for MHT control. All rejections rates are reported in percent. The sample period is 1972 to 2015.

	Alpha	FM	Alpha and FM
		CAPM	
CHT Rejections in $\mathcal{S}^{\mathcal{E}}$	23.4	24.8	7.7
RSW Rejections in $\mathcal{S}^{\mathcal{E}}$	3.4	8.2	0.6
RSW CNM rejections in $\mathcal{S}^{\mathcal{E}}$	—	—	92.0
Thresholds in $\mathcal{S}^{\mathcal{R}}$	3.76	3.36	—
False rejections in $\mathcal{S}^{\mathcal{R}}$	—	—	40.1
		FF3	
CHT Rejections in $\mathcal{S}^{\mathcal{E}}$	27.3	24.8	7.8
RSW Rejections in $\mathcal{S}^{\mathcal{E}}$	4.9	8.3	0.4
RSW CNM rejections in $\mathcal{S}^{\mathcal{E}}$	—	—	94.3
Thresholds in $\mathcal{S}^{\mathcal{R}}$	3.81	3.42	—
False rejections in $\mathcal{S}^{\mathcal{R}}$	—	—	44.2
		BS	
CHT Rejections in $\mathcal{S}^{\mathcal{E}}$	35.7	20.6	7.9
RSW Rejections in $\mathcal{S}^{\mathcal{E}}$	10.6	5.6	0.7
RSW CNM rejections in $\mathcal{S}^{\mathcal{E}}$	—	—	91.1
Thresholds in $\mathcal{S}^{\mathcal{R}}$	3.81	3.38	—
False rejections in $\mathcal{S}^{\mathcal{R}}$	—	—	46.5
		HXZ	
CHT Rejections in $\mathcal{S}^{\mathcal{E}}$	22.3	20.6	4.4
RSW Rejections in $\mathcal{S}^{\mathcal{E}}$	0.4	4.9	0.1
RSW CNM rejections in $\mathcal{S}^{\mathcal{E}}$	—	—	99.2
Thresholds in $\mathcal{S}^{\mathcal{R}}$	3.84	3.40	—
False rejections in $\mathcal{S}^{\mathcal{R}}$	—	—	45.6

5.3.2 Alternative definition and construction of the set of strategies

We conduct three more robustness checks by changing the definition of the set of randomly generated strategies. First, we repeat our analysis after eliminating signals labeled as *Ratios of three* from the sample set $\mathcal{S}^{\mathcal{E}}$. This smaller set contains 13,748 strategies. Summary statistics for this sample of strategies is reported in Panel B of Appendix Table A2. Second, instead of decile portfolios, we consider 2×3 portfolios as follows. We sort stocks into two groups based on size with NYSE median determining the cutoff. We independently sort stocks into three groups using as cutoffs the 30th and 70th percentile of the variable of interest. The long-short portfolio is then defined to be long (short) in the average of two size groups for tercile three (one).[15] Correspondingly, in the simulation, we sort stocks into three groups using as cutoffs the 30th and 70th percentile of the variable of interest. Third, in FM regressions, we replace raw variables with their ranks that run from zero to one. Such regressions are sometimes used in the literature to remove the effect of outliers (we do winsorize the variables at the 0.5 and 99.5 percentiles in the main analysis). Correspondingly, in the simulation, we run FM regressions with ranks.

Detailed results from the calibration exercises are reported in Appendix Table A4. Table 9 documents the CHT and MHT rejections, the RSW thresholds, and the proportion of false rejections on these alternate set of strategies. We find that the proportion of false rejections is around 40%, while thresholds remain relatively stable.

6. Conclusion

The finance profession has discovered a number of profitable trading strategies. Given that the CRSP and Compustat datasets have been studied for over four decades, it is likely

[15]The 2×3 approach to constructing portfolios is common when constructing factors (see, for example, Fama and French 1993, 2015). Since our paper tries to address the false rejections problem in the context of anomalies, the question of which 2×3 'factors' are false discoveries is a different question from the one that we are trying to answer. Nevertheless, it is of interest to examine our results if we do follow the alternative portfolio sorting approach.

Table 9: Alternative definitions and construction of strategies

The table reports CHT rejections, RSW rejection rates, the proportion of CNM rejections, as well as RSW thresholds and proportion of false rejections obtained from the calibration (see Appendix Table A4). We consider one alternative to the definition of trading strategies by excluding *Ratios of three*. We also consider two alternatives ways of constructing trading strategies: sorting stocks into 2×3 groups (instead of deciles), and using ranks of variables (instead of raw variables) in FM regressions. The significance level is 5% for all tests and we use only the RSW method for MHT control. All rejections rates are reported in percent. The sample period is 1972 to 2015.

	Alpha	FM	Alpha and FM
	Small set of strategies		
CHT Rejections in $\mathcal{S}^{\mathcal{E}}$	27.6	19.0	5.2
RSW Rejections in $\mathcal{S}^{\mathcal{E}}$	3.7	4.6	0.3
RSW CNM rejections in $\mathcal{S}^{\mathcal{E}}$	86.6	75.6	93.8
Thresholds in $\mathcal{S}^{\mathcal{R}}$	3.76	3.21	—
False rejections in $\mathcal{S}^{\mathcal{R}}$	—	—	40.1
	2×3 portfolios		
CHT Rejections in $\mathcal{S}^{\mathcal{E}}$	35.5	25.5	11.4
RSW Rejections in $\mathcal{S}^{\mathcal{E}}$	16.8	7.4	3.1
RSW CNM rejections in $\mathcal{S}^{\mathcal{E}}$	52.2	71.1	72.7
Thresholds in $\mathcal{S}^{\mathcal{R}}$	3.67	3.26	—
False rejections in $\mathcal{S}^{\mathcal{R}}$	—	—	40.8
	FM regressions with ranks		
CHT Rejections in $\mathcal{S}^{\mathcal{E}}$	27.5	25.6	10.3
RSW Rejections in $\mathcal{S}^{\mathcal{E}}$	8.5	7.5	1.8
RSW CNM rejections in $\mathcal{S}^{\mathcal{E}}$	69.3	70.7	82.1
Thresholds in $\mathcal{S}^{\mathcal{R}}$	3.85	3.38	—
False rejections in $\mathcal{S}^{\mathcal{R}}$	—	—	41.4

that the profitability of some of the discovered strategies appears exceptional due to luck. Further, out of the many strategies that are evaluated in isolation only the profitable ones are reported.

We account for the fact that only a small fraction of the studied strategies are reported and apply MHT to obtain a estimate of the proportion of false discoveries. Even after allowing for the possibility that researchers are much more likely, compared to the econometricians, to over-sample strategies that are false under the null, the proportion of false rejections using CHT is about 45.3%.

Appendices

A. MHT methods

We are interested in testing the performance of trading strategies by analyzing the abnormal returns generated by S signals. The test statistic is either a t-statistic or equivalently a p-value. The null hypothesis corresponding to each strategy is labeled as H_i. For ease of notation, we will relabel the strategies and order them from the best (highest t-statistic) to the worst (lowest t-statistic). In other words, it is assumed that $t_1 \geq t_2 \geq \ldots \geq t_S$, or equivalently $p_1 \leq p_2 \leq \ldots \leq p_S$. Some of the methods used in this section use a bootstrap procedure which is described later in in this section.

A.1 FWER

The strictest idea in MHT is to try to avoid any false rejections. This translates to controlling the FWER, which is defined as the probability of rejecting even one of the true null hypotheses:

$$\text{FWER} = \text{Prob}\ (F_1 > 1).$$

Thus, FWER measures the probability of even one false discovery (i.e., rejecting even one true null hypothesis). A testing method is said to control the FWER at a significance level α if FWER $\leq \alpha$. There are many approaches to controlling FWER.

A.1.1 Bonferroni method

The Bonferroni method, at level α, rejects H_i if $p_i \leq \alpha/S$. The Bonferroni method is a single-step procedure because all p-values are compared to a single critical value. This critical p-value is equal to α/S. For a very large number of strategies, this leads to an extremely small (large) critical p-value (t-statistic). While widely used for its simplicity, the biggest disadvantage of the Bonferroni method is that it is very conservative and leads to a loss of power. One of the main reasons for the lack of power is that the Bonferroni method ignores the cross-correlations that are bound to be present in most financial applications.

A.1.2 Holm method

This is a step-down method based on Holm (1979) and works as follows. The procedure starts by checking the most significant hypothesis and works its way down to less significant hypotheses. The null hypothesis H_i is rejected at level α if $p_i \leq \alpha/(S-i+1)$ for $i = 1, \ldots, S$. The intuition for the method is, if one is at stage of the hypothesis i, then hopefully the hypotheses $H_1, H_2, \ldots, H_{i-1}$ have been correctly rejected. One can then apply the Bonferroni correction to hypothesis H_i using the number of remaining hypotheses, $S - i + 1$.

In comparison with the Bonferroni method, the criterion for the smallest p-value is equally strict at α/S but it becomes less and less strict for larger p-values. Thus, the Holm method

will typically reject more hypotheses and is more powerful than the Bonferroni method. However, because it also does not take into account the dependence structure of the individual p-values, the Holm method is also very conservative.

A.1.3 Bootstrap reality check

Bootstrap reality check (BRC) is based on White (2000). The idea is to estimate the sampling distribution of the largest test statistic taking into account the dependence structure of the individual test statistics, thereby asymptotically controlling FWER.

The implementation of the method proceeds as follows. Bootstrap the data using procedure described in Section A.5. For each bootstrapped iteration b, calculate the highest (absolute) t-statistic across all strategies and call it $t_{\max}^{(b)}$, where the superscript b is used to clarify that these t-statistics come from the bootstrap. The critical value is computed as the $(1-\alpha)$ empirical percentile of B bootstrap iterations values $t_{\max}^{(1)}, t_{\max}^{(2)}, \ldots, t_{\max}^{(B)}$.

Statistically speaking, BRC can be viewed as a method that improves upon Bonferroni by using the bootstrap to get a less conservative critical value. From an economic point of view, BRC addresses the question of whether the strategy that appears the best in the observed data really beats the benchmark. However, BRC method does not attempt to identify as many outperforming strategies as possible.

A.1.4 StepM method

This method, based on Romano and Wolf (2005) addresses the problem of detecting as many out-performing strategies as possible. The stepwise StepM method is an improvement over the single-step BRC method in very much the same way as the stepwise Holm method improves upon the single-step Bonferroni method. The implementation of this procedure proceeds as follows:

1. Consider the set of all S strategies. For each cross-sectional bootstrap iteration, compute the maximum t-statistic, thus obtaining the set $t_{\max}^{(1)}, t_{\max}^{(2)}, \ldots, t_{\max}^{(B)}$. Then compute the critical value c_1 as the $(1-\alpha)$ empirical percentile of the set of maximal t-statistics, as in BRC method. Apply now the c_1 threshold to the set of original t-statistics and determine the number of strategies for which the null can be rejected. Say that there are R_1 strategies, for which $t_i \geq c_1$. We have now $S - R_1$ strategies remaining with t-statistics ordered as $t_{R_1+1}, t_{R_1+2}, \ldots, t_S$.

2. Consider the set of remaining $S - R_1$ strategies. For each bootstrapped iteration b, calculate the highest (absolute) t-statistic across all remaining strategies. To avoid cluttering up the notation, we will use the same symbols as before and call the maximal t-statistics of the b bootstrap iteration across the $S - R_1$ remaining strategies as $t_{\max}^{(b)}$. The critical value c_2 is computed as the $(1-\alpha)$ empirical percentile of B bootstrap iterations values $t_{\max}^{(1)}, t_{\max}^{(2)}, \ldots, t_{\max}^{(B)}$. Say that there are R_2 strategies, for which $t_i \geq c_2$, and are, therefore, rejected in this step. After this step, $S - R_1 - R_2$ strategies remain with t-statistics ordered as $t_{R_1+R_2+1}, t_{R_1+R_2+2}, \ldots, t_S$.

3. Repeat the procedure until there are no further strategies that are rejected. The StepM critical value for the entire procedure is equal to the critical value of the last step and the number of strategies that are rejected is equal to the sum of the number of strategies that are rejected in each step.

Like the Holm method, the StepM method is a step-down method that starts by examining the most significant strategies. The main advantage of the method is that, because it relies on bootstrap, it is valid under arbitrary dependence structure of the test statistics. As mentioned before, this method will detect many more out-performing strategies than the Bonferroni method or the BRC approach.

It is easy to see that the BRC approach amounts to only step one of the above procedure, namely computing only the critical value c_1. By continuing the method after the first step, more false null hypotheses can be rejected. Moreover, since typically $c_1 > c_2 > \ldots$, the critical value in StepM method is less conservative than that in BRC approach. Nevertheless, the StepM procedure still asymptotically controls FWER at significance level α.

A.2 k-FWER

By relaxing the strict FWER criterion, one can reject more false hypotheses. For instance, k-FWER is defined as the probability of rejecting at least k of the true null hypotheses:

$$k\text{-FWER} = \text{Prob}\{\text{Reject at least } k \text{ of the true null hypothesis}\}.$$

A testing method is said to control for k-FWER at a significance level α if k-FWER $\leq \alpha$. Testing methods such as Bonferroni and Holm, discussed earlier, can be generalized for k-FWER testing. Please refer to Romano, Shaikh, and Wolf (2008) for further details. Here we discuss only the extension of the StepM method which is known as the k-StepM method.

A.2.1 k-StepM method

The implementation of this procedure proceeds as follows:

1. Consider the set of all S strategies. For each bootstrapped iteration b, calculate the k-highest (absolute) t-statistic across all strategies and call it $t^{(b)}_{k\text{-max}}$, where the superscript b is used to clarify that these t-statistics come from the bootstrap. Compute the critical value c_1 as the $(1-\alpha)$ empirical percentile of B bootstrap iterations values $t^{(1)}_{k\text{-max}}, t^{(2)}_{k\text{-max}}, \ldots, t^{(B)}_{k\text{-max}}$. Say that there are R_1 strategies, for which $t_i \geq c_1$, and are, therefore, rejected in this step. After this step, $S - R_1$ strategies remain with t-statistics ordered as $t_{R_1+1}, t_{R_1+2}, \ldots, t_S$. Apart from the use of k-max instead of max, this step is identical to the first step of StepM procedure.

2. Consider the set of remaining $S - R_1$ strategies. Call this set `Remain`. Also consider a number $k-1$ of strategies from the set of already rejected strategies. Call this set `Reject`. Now consider the union of these two sets, `Consider` = `Remain` $\cup$ `Reject`. For each bootstrapped iteration b, calculate the k-highest (absolute) t-statistic across all strategies in the set `Consider` and call it $t^{(b)}_{k\text{-max}}$. Compute the $(1-\alpha)$ empirical

percentile of B bootstrap iterations values $t^{(1)}_{k\text{-max}}, t^{(2)}_{k\text{-max}}, \ldots, t^{(B)}_{k\text{-max}}$. This empirical percentile will depend on which $k-1$ strategies were included in the set `Reject`. Given that there are $\binom{R_1}{k-1}$ possible ways of choosing $k-1$ strategies from a set of R_1 strategies, the critical value c_2 is computed as the maximum across all these permutations. Say that there are R_2 strategies, for which $t_i \geq c_2$, and are, therefore, rejected in this step. After this step, $S - R_1 - R_2$ strategies remain with t-statistics ordered as $t_{R_1+R_2+1}, t_{R_1+R_2+2}, \ldots, t_S$.

3. Repeat the procedure until there are no further strategies that are rejected. The critical value of the procedure is equal to the critical value of the last step and the number of strategies that are rejected is equal to the sum of the number of strategies that are rejected in each step.

The key innovation in the k-StepM procedure is in the inclusion of (some of the) rejected strategies while calculating subsequent critical values (c_2 and thereafter). The intuition is as follows. Remember that ideally we want to calculate the empirical critical value from the set of strategies that are true under the null hypothesis. This set is unknown in practice. However, we can use the results of the first step to arrive at this set. The set `Remain` of remaining strategies that have not (yet) been rejected is an obvious candidate for strategies that are true under the null. If we are in the second step of the procedure, it stands to reason that the first step was not able to control k-FWER. In other words, less than k true null hypotheses were rejected in the first step. Let's say that number is in fact $k-1$. Obviously, we do not know with precision which $k-1$ true nulls have been rejected among the many strategies rejected in the first step. Therefore, to be cautious, Romano, Shaikh, and Wolf (2008) suggest looking at all possible combinations of $k-1$ rejected hypotheses from the set `Reject`.

A.3 False discovery rate (FDR)

A multiple testing method is said to control FDR at level δ if FDR $\equiv$ E(FDP) $\leq \delta$. Since FDR is already an expectation, controlling for FDR does not need additional specification of probabilistic significance level. One of the earliest methods to controlling FDR is by Benjamini and Hochberg (1995) and proceeds in a stepwise fashion as follows. Assuming as before that the individual p-values are ordered from the smallest to largest, and defining:

$$j^* = \max\left\{ j : p_j \leq \left(\frac{j}{S}\right)\delta \right\},$$

one rejects all hypotheses $H_1, H_2, \ldots, H_{j^*}$ (i.e., j^* is the index of the largest p-value among all hypotheses that are rejected). This is a step-up method that starts with examining the least significant hypothesis and moves up to more significant test statistics.

Benjamini and Hochberg (1995) show that their method controls FDR if the p-values are mutually independent. Benjamini and Yekutieli (2001) show that a more general control of FDR under a more general dependence structure of p-values can be achieved by replacing

the definition of j^* with:

$$j^* = \max\left\{j : p_j \leq \left(\frac{j}{S \times C_S}\right)\delta\right\},$$

where the constant $C_S = \sum_{i=1}^{S} 1/i \approx \log(S)$. However, the Benjamini and Yekutieli method is less powerful than that of Benjamini and Hochberg.

A.4 False discovery proportion (FDP)

A multiple testing method is said to control FDP at proportion γ and level α if Prob(FDP $> \gamma) \leq \alpha$. Lehmann and Romano (2005) and Romano and Shaikh (2006) develop extensions of the Holm method for FDP control. Here we discuss only the extension of the StepM procedure developed by Romano and Wolf (2007).

A.4.1 FDP-StepM method

The StepM procedure for control of FDP is as follows:

1. Let $k = 1$.
2. Apply the k-StepM method and denote by R_k the number of hypotheses rejected.
3. If $R_k < k/\gamma - 1$, then stop. Else, let $k = k + 1$ and return to step 2.

The FDP-StepM method is, thus, a sequence of k-StepM procedures. The intuition of applying an increasing series of k's is as follows. Consider controlling FDP at proportion $\gamma = 10\%$. We start by applying the 1-StepM method. Denote by R_1 the number of strategies rejected at this stage. Since the basic 1-StepM procedure controls for FWER, we can be confident that no false rejections have occurred so far, which in turn also implies that FDP has also been controlled. Consider now the issue of rejecting the strategy H_{R_1+1}, the next most significant strategy (recall that StepM is a step-down procedure).

Rejection of H_{R_1+1}, if the null of this strategy is true, renders the false discovery proportion to be equal to $1/(R_1 + 1)$. Since we are willing to tolerate 10% of false rejections, we would be willing to tolerate rejecting this strategy if $1/(R_1 + 1) < 0.1$ which is true if $R_1 > 9$. Thus if $R_1 < 9$ the procedure would stop at the first step. Alternatively, if $R_1 > 9$, the procedure would continue with the 2-StepM method, which by design should not reject more than one true hypothesis.

Besides the fact that the FDP-StepM method allows the researcher to directly control FDP, one other big advantage of this method for us is that it accounts for arbitrary dependence structure in the data and, therefore, in the individual p-values.

A.5 Bootstrap method

Some of the methods describe above rely on a bootstrap. We describe the details of the bootstrap in this section. This approach, inspired by Kosowski, Timmermann, Wermers, and White (2006) and Fama and French (2010) and used recently by Yan and Zheng (2017),

relies on bootstrapping the cross-section of fund returns through time thereby preserving the cross-sectional dependence structure in strategy returns and ultimately their alpha estimates.

To bootstrap under the null, say of zero alpha, we first subtract the factor-model alpha from the monthly portfolio returns. Each bootstrap run is a random sample (with replacement) of the alpha-adjusted returns and the factors over the sample period. To preserve the cross-sectional correlation we apply the same bootstrap draw to all portfolios and to the factors. To preserve possible autocorrelation in the return structure, we construct the stationary bootstrap of Politis and Romano (1994) by drawing random blocks with an average length of six months. Due to the computational constraints imposed by the large scale of our exercise we limit the exercise to 1,000 bootstrap samples. For each bootstrap run we obtain the portfolio alphas and their t-statistics under the null of zero alpha. This bootstrapped distribution of t-statistics is used in the MHT methods that need it.

We conduct a similar experiment for FM coefficients. In particular, for each signal variable we start by subtracting the average from the time-series of the FM coefficient thus obtaining a time-series of adjusted coefficients under the null of no explanatory power. We then bootstrap 1,000 times the time-series of pseudo coefficients and calculate the means and t-statistics for each bootstrap iteration. This generates the bootstrapped distribution of t-statistics of FM coefficients under null.

Table A1: Basic variables used to construct trading strategies

#	Short	Long	#	Short	Long
1	aco	Current Assets Other Total	61	ebit	Earnings Before Interest and Taxes
2	acox	Current Assets Other Sundry	62	ebitda	Earnings Before Interest
3	act	Current Assets Total	63	emp	Employees
4	am	Amortization of Intangibles	64	epsfi	Earnings Per Share (Diluted) Including Extraordinary Items
5	ao	Assets Other	65	epsfx	Earnings Per Share (Diluted) Excluding Extraordinary Items
6	aox	Assets Other Sundry	66	epspi	Earnings Per Share (Basic) Including Extraordinary Items
7	ap	Accounts Payable Trade	67	epspx	Earnings Per Share (Basic) Excluding Extraordinary Items
8	aqc	Acquisitions	68	esub	Equity in Earnings Unconsolidated Subsidiaries
9	aqi	Acquisitions Income Contribution	69	esubc	Equity in Net Loss Earnings
10	aqs	Acquisitions Sales Contribution	70	fca	Foreign Exchange Income (Loss)
11	at	Assets Total	71	fopo	Funds from Operations Other
12	bkvlps	Book Value Per Share	72	gp	Gross Profit (Loss)
13	caps	Capital Surplus-Share Premium Reserve	73	ib	Income Before Extraordinary Items
14	capx	Capital Expenditures	74	ibadj	Income Before Extraordinary Items Adjusted for Common Stock Equivalents
15	capxv	Capital Expend Property, Plant and Equipment Schd V	75	ibc	Income Before Extraordinary Items (Cash Flow)
16	ceq	Common-Ordinary Equity Total	76	ibcom	Income Before Extraordinary Items Available for Common
17	ceql	Common Equity Liquidation Value	77	icapt	Invested Capital Total
18	ceqt	Common Equity Tangible	78	idit	Interest and Related Income Total
19	ch	Cash	79	intan	Intangible Assets Total
20	che	Cash and Short-Term Investments	80	intc	Interest Capitalized
21	chech	Cash and Cash Equivalents Increase-(Decrease)	81	invfg	Inventories Finished Goods
22	cogs	Cost of Goods Sold	82	invrm	Inventories Raw Materials
23	cshfd	Common Shares Used to Calc Earnings Per Share Fully Diluted	83	invt	Inventories Total
24	csho	Common Shares Outstanding	84	invwip	Inventories Work In Process
25	cshpri	Common Shares Used to Calculate Earnings Per Share Basic	85	itcb	Investment Tax Credit (Balance Sheet)
26	cshr	Common-Ordinary Shareholders	86	itci	Investment Tax Credit (Income Account)
27	cstk	Common-Ordinary Stock (Capital)	87	ivaeq	Investment and Advances Equity
28	cstkcv	Common Stock-Carrying Value	88	ivao	Investment and Advances Other
29	cstke	Common Stock Equivalents Dollar Savings	89	ivch	Increase in Investments
30	dc	Deferred Charges	90	ivst	Short-Term Investments Total
31	dclo	Debt Capitalized Lease Obligations	91	lco	Current Liabilities Other Total
32	dcpstk	Convertible Debt and Preferred Stock	92	lcox	Current Liabilities Other Sundry
33	dcvsr	Debt Senior Convertible	93	lct	Current Liabilities Total
34	dcvsub	Debt Subordinated Convertible	94	lifr	LIFO Reserve
35	dcvt	Debt Convertible	95	lifrp	LIFO Reserve Prior
36	dd	Debt Debentures	96	lo	Liabilities Other Total
37	dd1	Long-Term Debt Due in One Year	97	lse	Liabilities and Stockholders Equity Total
38	dd2	Debt Due in 2nd Year	98	lt	Liabilities Total
39	dd3	Debt Due in 3rd Year	99	mib	Noncontrolling Interest (Balance Sheet)
40	dd4	Debt Due in 4th Year	100	mibt	Noncontrolling Interests Total Balance Sheet
41	dd5	Debt Due in 5th Year	101	mii	Noncontrolling Interest (Income Account)
42	dlc	Debt in Current Liabilities Total	102	mrc1	Rental Commitments Minimum 1st Year
43	dltis	Long-Term Debt Issuance	103	mrc2	Rental Commitments Minimum 2nd Year
44	dlto	Other Long-term Debt	104	mrc3	Rental Commitments Minimum 3rd Year
45	dltp	Long-Term Debt Tied to Prime	105	mrc4	Rental Commitments Minimum 4th Year
46	dltr	Long-Term Debt Reduction	106	mrc5	Rental Commitments Minimum 5th Year
47	dltt	Long-Term Debt Total	107	mrct	Rental Commitments Minimum 5 Year Total
48	dm	Debt Mortgages Other Secured	108	msa	Marketable Securities Adjustment
49	dn	Debt Notes	109	ni	Net Income (Loss)
50	do	Discontinued Operations	110	niadj	Net Income Adjusted for Common-Ordinary Stock (Capital) Equivalents
51	dp	Depreciation and Amortization	111	nopi	Nonoperating Income (Expense)
52	dpact	Depreciation, Depletion and Amortization (Accumulated)	112	nopio	Nonoperating Income (Expense) Other
53	dpc	Depreciation and Amortization (Cash Flow)	113	np	Notes Payable Short-Term Borrowings
54	dpvieb	Depreciation (Accumulated) Ending Balance (Schedule VI)	114	ob	Order Backlog
55	ds	Debt-Subordinated	115	oiadp	Operating Income After Depreciation
56	dv	Cash Dividends (Cash Flow)	116	oibdp	Operating Income Before Depreciation
57	dvc	Dividends Common-Ordinary	117	pi	Pretax Income
58	dvp	Dividends Preferred-Preference	118	ppegt	Property, Plant and Equipment Total (Gross)
59	dvpa	Preferred Dividends in Arrears	119	ppent	Property, Plant and Equipment Total (Net)
60	dvt	Dividends Total	120	ppeveb	Property, Plant, and Equipment Ending Balance (Schedule V)

#	Short	Long	#	Short	Long
121	prstkc	Purchase of Common and Preferred Stock	154	txfo	Income Taxes Foreign
122	pstk	Preferred-Preference Stock (Capital) Total	155	txp	Income Taxes Payable
123	pstkc	Preferred Stock Convertible	156	txr	Income Tax Refund
124	pstkl	Preferred Stock Liquidating Value	157	txs	Income Taxes State
125	pstkn	Preferred-Preference Stock Nonredeemable	158	txt	Income Taxes Total
126	pstkr	Preferred-Preference Stock Redeemable	159	txw	Excise Taxes
127	pstkrv	Preferred Stock Redemption Value	160	wcap	Working Capital (Balance Sheet)
128	re	Retained Earnings	161	xacc	Accrued Expenses
129	rea	Retained Earnings Restatement	162	xad	Advertising Expense
130	reajo	Retained Earnings Other Adjustments	163	xi	Extraordinary Items
131	recco	Receivables Current Other	164	xido	Extraordinary Items and Discontinued Operations
132	recd	Receivables Estimated Doubtful	165	xidoc	Extraordinary Items and Discontinued Operations (Cash Flow)
133	rect	Receivables Total	166	xint	Interest and Related Expense Total
134	recta	Retained Earnings Cumulative Translation Adjustment	167	xlr	Staff Expense Total
135	rectr	Receivables Trade	168	xopr	Operating Expenses Total
136	reuna	Retained Earnings Unadjusted	169	xpp	Prepaid Expenses
137	revt	Revenue Total	170	xpr	Pension and Retirement Expense
138	sale	Sales-Turnover (Net)	171	xrd	Research and Development Expense
139	seq	Stockholders Equity Parent	172	xrdp	Research Development Prior
140	siv	Sale of Investments	173	xrent	Rental Expense
141	spi	Special Items	174	xsga	Selling, General and Administrative Expense
142	sppe	Sale of Property	175	ret1	1m Past Return
143	sstk	Sale of Common and Preferred Stock	176	ret3	3m Past Return
144	tlcf	Tax Loss Carry Forward	177	ret6	6m Past Return
145	tstk	Treasury Stock Total (All Capital)	178	ret9	9m Past Return
146	tstkc	Treasury Stock Common	179	ret12	1y Past Return
147	tstkn	Treasury Stock Number of Common Shares	180	size	Market Capitalization
148	txc	Income Taxes Current	181	price	Price
149	txdb	Deferred Taxes (Balance Sheet)	182	volume	Traded Volume
150	txdc	Deferred Taxes (Cash Flow)	183	dvolume	Dollar Traded Volume
151	txdi	Income Taxes Deferred	184	turn	Turnover
152	txditc	Deferred Taxes and Investment Tax Credit	185	vol	1y Return Volatility

Table A2: Descriptive statistics of portfolio raw returns on trading strategies – SubSamples and different factor models

This table reports the cross-sectional mean, median, standard deviation, minimum, and maximum of the t-statistics of monthly average return, alpha and FM coefficients as in Table 1 but for subsamples and different factor models. Panel A uses all stocks and all strategies but uses different factor models described in Table 8. Panel B is for the subsample of the main set of strategies that does not contain portfolio returns constructed using *Ratio of three* signals. The subsample is thus composed by 13,748 strategies and uses the Fama and French (2015) five-factor model augmented with the momentum factor.The row entitled 'Alpha 2×3' sorts stocks into 2×3 groups (instead of deciles), and the rwo entitled 'FM rank' uses ranks of variables (instead of raw variables) in FM regressions. The sample period is 1972 to 2015.

	Mean	Median	Std	Min	Max	$\%\|t\| > 1.96$	$\%\|t\| > 2.57$
	Panel A1: Alpha t-statistics for different factor models						
CAPM	−0.09	−0.12	1.64	−7.53	7.88	23.4	12.2
FF3	−0.24	−0.26	1.76	−8.15	8.69	27.4	14.8
BS	−0.29	−0.29	2.09	−8.23	7.96	35.7	23.1
HXZ	−0.19	−0.18	1.58	−7.30	7.78	22.3	11.1
	Panel A2: FM t-statistics for different factor models						
CAPM	0.14	0.03	1.97	−11.55	11.55	24.8	14.4
FF3	−0.02	−0.07	1.84	−9.27	8.58	24.8	14.7
BS	0.07	0.03	1.66	−8.58	8.33	20.6	11.6
HXZ	0.05	0.01	1.64	−8.47	7.92	20.6	11.4
	Panel B: Small set of strategies						
Return	−0.08	−0.10	1.21	−7.04	6.72	10.6	3.7
Alpha	−0.19	−0.19	1.62	−7.80	9.01	23.2	11.9
Alpha 2×3	0.14	0.10	2.25	−7.38	7.97	35.1	24.1
FM	0.07	0.03	1.64	−8.63	8.12	20.0	11.2
FM rank	−0.38	−0.31	1.73	−5.46	5.95	28.0	16.1

Table A3: Alternative factor models: Calibration

We use a global minimization algorithm (i.e., particle swarm) to calibrate the statistical framework presented in Section 3 and 4. We find the vector of parameters that minimizes the sum of the squared percentage distance of the target quantities. We consider four alternative factor models: CAPM, FF3, BS, and HXZ. In Panel A we report the calibrated parameters. In Panel B, we compare the target quantities obtained from the simulation to the respective quantities from the data. In Panel C we report the adjusted MHT thresholds and the proportion of false discoveries.

	CAPM		FF3		BS		HXZ	
	Panel A: Calibrated parameters							
σ_η	3.0		3.1		7.4		3.2	
π	0.13		0.09		0.10		0.03	
Ω	16.1		21.1		17.1		56.9	
	Panel B: Target quantities							
	Data	Sim	Data	Sim	Data	Sim	Data	Sim
CNM rejections in $\mathcal{S}^{\mathcal{E}}$	92.0	91.9	94.3	94.2	91.1	94.1	99.2	98.6
StDev(ψ) in $\mathcal{S}^{\mathcal{E}}$	17.1	17.5	17.1	17.5	17.2	17.6	17.2	17.7
$E(\lvert\text{sign. } \alpha_s\rvert)/E(\lvert\text{sign. } \alpha_i\rvert)$ in $\mathcal{S}^{\mathcal{E}}$	0.20	0.20	0.14	0.16	0.12	0.12	0.15	0.11
CNM rejections in $\mathcal{S}^{\mathcal{P}}$	27.0	26.7	27.0	29.9	27.0	32.0	27.0	32.5
	Panel C: Outcome quantities in the set $\mathcal{S}^{\mathcal{R}}$							
Alpha threshold, $\mathcal{T}_\alpha$	3.76		3.81		3.81		3.84	
FM threshold, $\mathcal{T}_\lambda$	3.36		3.42		3.38		3.40	
CNM rejections	38.0		42.0		44.1		42.2	
False rejections	40.1		44.2		46.5		45.6	

Table A4: Alternative definitions and construction of strategies: Calibration

We use a global minimization algorithm (particle swarm) to calibrate the statistical framework presented in Section 4. We find the vector of parameters that minimizes the sum of the squared percentage distance of the target quantities. We consider one alternative to the definition of trading strategies by excluding ratios of three. We also consider two alternatives ways of constructing trading strategies: sorting stocks into 2×3 groups (instead of deciles), and using ranks of variables (instead of raw variables) in FM regressions. In Panel A we report the calibrated parameters. In Panel B, we compare the target quantities obtained from the simulation to the respective quantities from the data. In Panel C we report the adjusted MHT thresholds and the proportion of false discoveries.

	Small set of strategies		2×3 portfolios		FM regressions with ranks	
	Panel A: Calibrated parameters					
σ_η	5.6		19.3		13.9	
π	0.10		0.75		0.31	
Ω	21.1		3.4		5.1	
	Panel B: Target quantities					
	Data	Sim	Data	Sim	Data	Sim
CNM rejections in $\mathcal{S}^{\mathcal{E}}$	93.8	93.7	72.7	71.0	82.2	77.9
StDev(ψ) in $\mathcal{S}^{\mathcal{E}}$	14.2	15.6	14.2	15.6	14.3	15.6
$E(\lvert\text{sign. } \alpha_s\rvert)/E(\lvert\text{sign. } \alpha_i\rvert)$ in $\mathcal{S}^{\mathcal{E}}$	0.11	0.12	0.08	0.08	0.13	0.13
CNM rejections in $\mathcal{S}^{\mathcal{P}}$	27.0	28.3	27.0	28.2	27.0	28.0
	Panel C: Outcome quantities in the set $\mathcal{S}^{\mathcal{R}}$					
Alpha threshold, $\mathcal{T}_\alpha$	3.76		3.67		3.85	
FM threshold, $\mathcal{T}_\lambda$	3.21		3.26		3.38	
CNM rejections	38.2		36.7		38.9	
False rejections	40.9		40.8		41.4	

References

Andrikogiannopoulou, Angie, and Filippos Papakonstantinous, 2019, Reassessing False Discoveries in Mutual Fund Performance: Skill, Luck, or Lack of Power? forthcoming *Journal of Finance*.

Barillas, Francisco and Jay Shanken, 2018, Comparing Asset Pricing Models, *Journal of Finance* 73, 715–754.

Barras, Laurent, Olivier Scaillet, and Russ Wermers, 2010, False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas, *Journal of Finance* 65, 179–216.

Benjamini, Yoav, and Yosef Hochberg, 1995, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 289–300.

Benjamini, Yoav, and Daniel Yekutieli, 2001, The Control of the False Discovery Rate in Multiple Testing under Dependency, *Annals of Statistics* 29, 1165–1188.

Bonferroni, Carlo Emilio, 1936, *Teoria Statistica Delle Classi e Calcolo Delle Probabilità* (Libreria Internazionale Seeber).

Brennan, Michael, Tarun Chordia, and Avanidhar Subrahmanyam, 1998, Alternative Factor Specifications, Security Characteristics, and the Cross-Section of Expected Stock Returns, *Journal of Financial Economics* 49, 345–373.

Carhart, Mark M., 1997, On Persistence in Mutual Fund Performance, *The Journal of Finance* 52, 57–82.

Chang, Andrew C., and Phillip Li, 2018, Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Often Not," *Critical Finance Review* 7.

Chen, Andrew Y., and Tom Zimmermann, 2018, Publication Bias and the Cross-Section of Stock Returns, Working paper.

Chen, Andrew Y., 2019, Do T-Stat Hurdles Need to be Raised? Identification of Publication Bias in the Cross-Section of Stock Returns, Working paper.

Chordia, Tarun, Avanidhar Subrahmanyam, and Qing Tong, 2014, Have Capital Market Anomalies Attenuated in the Recent Era of High Liquidity and Trading Activity?, *Journal of Accounting and Economics* 58, 41–58.

Conrad, Jennifer, Michael Cooper, and Gautam Kaul, 2003, Value versus Glamour, *Journal of Finance* 58, 1969–1995.

Daniel, Kent, and Tobias J. Moskowitz, 2016, Momentum Crashes, *Journal of Financial Economics* 122, 221–247.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson, 1986, Replication in Empirical Economics: The Journal of Money, Credit, and Banking Project, *American Economic Review* 76, 587–630.

Fama, Eugene F., and Kenneth R. French, 1993, Common Risk Factors in the Returns on Stocks and Bonds, *Journal of Financial Economics* 33, 3–56.

Fama, Eugene F., and Kenneth R. French, 2008, Dissecting Anomalies, *Journal of Finance* 63, 1653–1678.

Fama, Eugene F., and Kenneth R. French, 2010, Luck Versus Skill in the Cross-Section of Mutual Fund Returns, *Journal of Finance* 65, 1915–1947.

Fama, Eugene F., and Kenneth R. French, 2015, A Five-Factor Asset Pricing Model, *Journal of Financial Economics* 116, 1–22.

Fama, Eugene F., and James D. MacBeth, 1973, Risk, Return and Equilibrium: Empirical Tests, *Journal of Political Economy* 81, 607–636.

Foster, F. Douglas, Tom Smith, and Robert E. Whaley, 1997, Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R2, *Journal of Finance* 52, 591–607.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, 2018, Dissecting Characteristics Nonparametrically, Working paper.

Green, Jeremiah, John R. M. Hand, and X. Frank Zhang, 2017, The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns, *Review of Financial Studies* 30, 4389–4436.

Genovese, Christopher R., and Larry Wasserman, 2006, Exceedance Control of the False Discovery Proportion, *Journal of the American Statistical Association* 101, 1408–1417.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2018, Empirical Asset Pricing via Machine Learning, Working paper.

Harvey, Campbell R., 2017, The Scientific Outlook in Financial Economics, *Journal of Finance* 72, 1399–1440.

Harvey, Campbell R., and Yan Liu, 2019, False (and Missed) Discoveries in Financial Economics, Working paper.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, ... and the Cross-Section of Expected Returns, *Review of Financial Studies* 29, 5–68.

Holm, Sture, 1979, A Simple Sequentially Rejective Multiple Test Procedure, *Scandinavian Journal of Statistics* 6, 65–70.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015, Digesting Anomalies: An Investment Approach, *Review of Financial Studies* 28, 650–705.

Hou, Kewei, Chen Xue, and Lu Zhang, 2019, Replicating Anomalies, forthcoming *Review of Financial Studies*.

Ioannidis, John P. A., 2005, Why Most Published Research Findings Are False, *PLoS Medicine* 2, 696–701.

Johnson, Norman L., Samuel Kotz, and N. Balakrishnan, 1994, *Continuous Univariate Distributions, Volume 1* (Wiley-Interscience).

Karolyi, G. Andrew, and Bong-Chan Kho, 2004, Momentum Strategies: Some Bootstrap Tests, *Journal of Empirical Finance* 11, 509–536.

Kennedy, James, and Russell Eberhart, 1995, Particle Swarm Optimization, *Proceedings of the IEEE International Conference on Neural Networks*, 1942–1948.

Kosowski, Robert, Allan Timmermann, Russ Wermers, and Hal White, 2006, Can Mutual Fund "Stars" Really Pick Stocks? New Evidence from a Bootstrap Analysis, *Journal of Finance* 61, 2551–2595.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2018, Shrinking the Cross Section, forthcoming *Journal of Financial Economics*.

Lehmann, Eric L., and Joseph P. Romano, 2005, Generalizations of the Familywise Error Rate, *Annals of Statistics* 33, 1138–1154.

Leamer, Edward E., 1978, *Specification Searches* (Wiley, New York).

Leamer, Edward E., 1983, Let's Take the Con Out of Econometrics, *American Economic Review* 73, 31–43.

Linnainmaa, Juhani T., and Michael Roberts, 2018, The History of the Cross-Section of Stock Returns, *Review of Financial Studies* 31, 2606–2649.

Lo, Andrew W., and A. Craig MacKinlay, 1990, Data-Snooping Biases in Tests of Financial Asset Pricing Models, *Review of Financial Studies* 3, 431–467.

McCullough, B. D., and H. D. Vinod, 2003, Verifying the Solution from a Nonlinear Solver: A Case Study, *American Economic Review* 93, 873–892.

McLean, R. David, and Jeffrey Pontiff, 2016, Does Academic Research Destroy Stock Return Predictability? *Journal of Finance* 71, 5–32.

Novy-Marx, Robert, and Mihail Velikov, 2016, A Taxonomy of Anomalies and Their Trading Costs, *Review of Financial Studies* 29, 104–147.

Politis, Dimitris N., and Joseph P. Romano, 1994, The Stationary Bootstrap, *Journal of the American Statistical Association* 89, 1303–1313.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf, 2008, Formalized Data Snooping Based On Generalized Error Rates, *Econometric Theory* 24, 404–447.

Romano, Joseph P., and Azeem M. Shaikh, 2006, Stepup Procedures for Control of Generalizations of the Familywise Error Rate, *Annals of Statistics* 34, 1850–1873.

Romano, Joseph P., and Michael Wolf, 2005, Stepwise Multiple Testing as Formalized Data Snooping, *Econometrica* 73, 1237–1282.

Romano, Joseph P., and Michael Wolf, 2007, Control of Generalized Error Rates in Multiple Testing, *Annals of Statistics* 35, 1378–1408.

Schwert, G. William, 2003, Anomalies and Market Efficiency, in: George M. Constantinides, Milton Harris, and René M. Stulz (ed.), *Handbook of the Economics of Finance*, edition 1, volume 1, chapter 15, 939–974 Elsevier.

Sullivan, Ryan, Allan Timmermann, and Halbert White, 1999, Data-Snooping, Technical Trading Rule Performance, and the Bootstrap, *Journal of Finance* 54, 1647–1691.

Yan, Xuemin (Sterling), and Lingling Zheng, 2017, Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach, *Review of Financial Studies* 30, 1382–1423.

White, Halbert, 2000, A Reality Check for Data Snooping, *Econometrica* 68, 1097–1126.

Swiss Finance Institute

Swiss Finance Institute (SFI) is the national center for fundamental research, doctoral training, knowledge exchange, and continuing education in the fields of banking and finance. SFI's mission is to grow knowledge capital for the Swiss financial marketplace. Created in 2006 as a public–private partnership, SFI is a common initiative of the Swiss finance industry, leading Swiss universities, and the Swiss Confederation.

swiss:finance:institute

c/o University of Geneva, Bd. Du Pont d'Arve 42, CH-1211 Geneva 4
T +41 22 379 84 71, rps@sfi.ch, www.sfi.ch