STATISTICS

The reusable holdout: Preserving validity in adaptive data analysis

Cynthia Dwork,[1]* Vitaly Feldman,[2]* Moritz Hardt,[3]* Toniann Pitassi,[4]* Omer Reingold,[5]* Aaron Roth[6]*

Misapplication of statistical data analysis is a common cause of spurious discoveries in scientific research. Existing approaches to ensuring the validity of inferences drawn from data assume a fixed procedure to be performed, selected before the data are examined. In common practice, however, data analysis is an intrinsically adaptive process, with new analyses generated on the basis of data exploration, as well as the results of previous analyses on the same data. We demonstrate a new approach for addressing the challenges of adaptivity based on insights from privacy-preserving data analysis. As an application, we show how to safely reuse a holdout data set many times to validate the results of adaptively chosen analyses.

Throughout the scientific community there is a growing recognition that claims of statistical significance in published research are frequently invalid. There has been a great deal of effort to understand and propose mitigations for this problem, largely focusing on statistical methods for controlling the false discovery rate in multiple hypothesis testing (*1*). However, the statistical inference theory surrounding this body of work assumes that a fixed procedure is performed, selected before the data are gathered. In contrast, the practice of data analysis in scientific research is, by nature, an adaptive process in which new analyses are chosen on the basis of data exploration and previous analyses of the same data.

It is now well understood that adapting the analysis to data results in an implicit multiple comparisons problem that is not captured in the reported significance levels of standard statistical procedures or by existing techniques for controlling the false discovery rate. This problem, in some contexts referred to as "p-hacking" or "researcher degrees of freedom," is one of the primary explanations as to why research findings are frequently false (*2*–*4*).

The traditional perspective on adaptivity makes it necessary to explicitly account for all of the possible ways to perform the analysis to provide validity guarantees for the adaptive analysis. Although this approach might be possible in simpler studies, it is technically challenging and often impractical in more complicated analyses (*4*). Numerous techniques have been developed by statisticians to address common special cases of adaptive data analysis. Most of these methods focus on a single round of adaptivity—such as variable selection followed by regression on selected variables or model selection followed by testing—and are optimized for specific inference procedures [the literature is too vast to adequately cover here, but see chapter 7 in (*5*) for a starting point]. There are also procedures for controlling false discovery in a sequential setting where tests arrive one-by-one (*6*–*8*). However, these results crucially depend on all tests maintaining their statistical properties despite being sequentially chosen—an assumption that is often difficult to justify in a complex adaptive analysis.

One proposed approach for avoiding the issue of adaptivity is preregistration; that is, defining the entire data analysis protocol ahead of time, thus forcing the analysis to be nonadaptive. A recent open letter (*9*) with more than 80 signatories calls for preregistration in science. Although safe, this proposal can be burdensome on the researcher and may limit the kind of analysis he or she can perform (*4*). As a result, this method has had difficulty gaining momentum in practice. A more popular approach for avoiding problems of this type is to validate data-dependent hypotheses or statistics on a holdout set. The data analyst starts by partitioning data samples randomly into training data and holdout data. The analyst interacts with the training set to obtain a data statistic of interest: for example, correlation between certain traits or the accuracy of a predictive model. The statistic is then validated by computing its value on the holdout set. Because the holdout was drawn from the same data distribution independently of the statistic, standard statistical inference procedures can safely be used.

A major drawback of this basic approach is that the holdout set, in general, is not reusable. If the analyst uses the outcome of the validation to select an additional data statistic, that statistic is no longer independent of the holdout data, and further use of the holdout set for validation can lead to incorrect statistical inference. To preserve statistical validity, the only known safe approach is to collect new data for a fresh holdout set. This conservative approach is very costly and thus is frequently abused, resulting in overfitting to the holdout set (*10*–*12*).

In this work we describe a general method, together with a specific instantiation for reusing a holdout set while maintaining the statistical guarantees of fresh data. The analyst is given unfettered access to the training data set but can only access the holdout set via an algorithm (equivalently, a mechanism) that allows the analyst to validate statistics on the holdout set. Armed with such a mechanism, the analyst is free to explore the (training) data ad libitum, generating and computing statistics, validating them on the holdout, and repeating this procedure, as well as sharing outcomes with other analysts who may also use the same holdout set.

The crucial idea behind our reusable holdout method comes from differential privacy—a notion of privacy preservation in data analysis introduced in computer science (*13*). Roughly speaking, differential privacy ensures that the probability of observing any outcome from an analysis is essentially unchanged by modifying any single data set element. Such a condition is often called a stability guarantee. An important line of work establishes connections between the stability of a learning algorithm and its ability to generalize (*14*–*16*). It is known that certain stability notions are necessary and sufficient for generalization. Unfortunately, the stability notions considered in these prior works do not compose in the sense that running multiple stable algorithms sequentially and adaptively may result in a procedure that is not stable. Differential privacy is stronger than these previously studied notions of stability and, in particular, possesses strong adaptive composition guarantees.

In a nutshell, the reusable holdout mechanism is simply this: access the holdout set only via a differentially private mechanism. The intuition is that if we can learn about the data set in aggregate while provably learning very little about any individual data element, then we can control the information leaked and thus prevent overfitting. More specifically, we introduce a new notion of maximum information that controls overfitting and can be bounded using differential privacy [for an overview, see section 1 of (*17*)]. We present an implementation of the reusable holdout, called Thresholdout, and show that it provably validates a large number of adaptively chosen statistics. We then use a simple classification algorithm on synthetic data to illustrate the properties of Thresholdout. The classifier produced by the algorithm overfits the data when the holdout set is reused naively but does not overfit if used with our reusable holdout.

We operate in a standard setting: an analyst is given a data set $S = (x_1, \ldots, x_n)$ of n samples drawn randomly and independently from some unknown distribution P over a discrete universe X of possible data points. Although our approach can be applied more generally, we focus here on validating statistics that can be expressed as the mean of some arbitrary function $\phi : X \to [0, 1]$ on the data set $E_S[\phi] = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$ [for additional details, see section 1.1 of (*17*)]. Such statistics are

[1]Microsoft Research, Mountain View, CA 94043, USA. [2]IBM Almaden Research Center, San Jose, CA 95120, USA. [3]Google Research, Mountain View, CA 94043, USA. [4]Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3G4, Canada. [5]Samsung Research America, Mountain View, CA 94043, USA. [6]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA.
*Corresponding author. E-mail: dwork@microsoft.com (C.D.); vitaly@post.harvard.edu (V.F.); m@mrtz.org (M.H.); toni@cs.toronto.edu (T.P.); omer.reingold@gmail.com (O.R.); aaroth@cis.upenn.edu (A.R.)

used to estimate the expected value of ϕ on a sample drawn randomly from the distribution P or $P[\phi] = \mathbf{E}_{x\sim P}[\phi(x)]$. A variety of quantities of interest in data analysis can be expressed as the expectation $\mathbf{E}_{x\sim P}[\phi(x)]$ of some function ϕ on P. Examples include true means and moments of individual attributes, correlations between attributes and the generalization error of a predictive model. Moreover, sufficiently precise estimates of these expectations suffice for model selection and assessment.

The data set S is randomly partitioned into training and holdout sets (S_t and S_h, respectively), and the data analyst is allowed to explore the training set without restrictions and generate functions ϕ to estimate the expectation on P. The analyst may access S_h only through Thresholdout. Thresholdout takes the holdout and training sets as input and, for all functions given by the analyst, provides statistically valid estimates of each function's expectation on P. Specifically, for a sufficiently large holdout set, Thresholdout guarantees that for every function $\phi : X \to [0,1]$ generated by the analyst, it will return a value v_ϕ such that $|v_\phi - P[\phi]| \leq \tau$, with probability at least $1 - \beta$, for analyst's choice of error τ and confidence β. The probability space is over the random choice of the data elements in S_h and S_t and the randomness introduced by the mechanism. We emphasize that the estimates are guaranteed to be accurate with respect to the true distribution, even when the functions are generated sequentially and adaptively by the analyst, up to a large number of functions. Our algorithm can equivalently be viewed as producing conservative confidence intervals on adaptively chosen sequences of linear functionals [for the formal connection to confidence intervals, see section 4 of (*17*)].

Given a function ϕ, the algorithm first checks if the difference between the average value of ϕ on the training set S_t (denoted by $E_{S_t}[\phi]$) and the average value of ϕ on the holdout set S_h (denoted by $E_{S_h}[\phi]$) is below a certain threshold $T + \eta$. Here, T is a fixed number such as 0.01, and η is a Laplace noise variable of standard deviation smaller than T by a small factor such as 4. [The Laplace distribution is a symmetric exponential distribution. Adding Laplace noise is one of the most basic operations in differential privacy (*13*).] If the difference is below the threshold, then the algorithm returns $E_{S_t}[\phi]$; that is, the value of ϕ on the training set. If the difference is above the threshold, then the algorithm returns the average value of the function on the holdout after adding Laplacian noise; that is, $E_{S_h}[\phi] + \xi$ (where ξ is a random variable distributed according to the Laplace distribution).

Though it is very simple, Thresholdout gives a surprisingly strong guarantee. Informally, the guarantee is that for any fixed accuracy parameter τ, Thresholdout can continue validating the estimates on the training sets until either the total number of functions asked becomes exponentially large in the size of S_h or the number of functions that fail the validation (meaning average values on S_h and S_t differ by more than the noisy threshold) becomes quadratically large in the size of S_h. Our guarantee can therefore be interpreted as saying that Thresholdout detects a quadratic number of functions that overfit to the training set (or false discoveries) and arise due to adaptivity; additionally, Thresholdout provides a correct estimate of the expectation for those functions. We describe further details of the implementation of Thresholdout and the formal guarantees it satisfies in section 2 of (*17*).

We describe a simple experiment on synthetic data that illustrates the danger of reusing a standard holdout set and how this issue can be resolved by our reusable holdout. The design of this experiment is inspired by Freedman's classical experiment, which demonstrated the dangers of performing variable selection and regression on the same data (*18*). The experiment is commonly referred to as "Freedman's paradox" due to the surprisingly strong effect on the validity of the results.

In our experiment, the analyst wants to build a classifier via the following common strategy. First, the analyst finds a set of single attributes that are correlated with the class label. Then he or she aggregates the correlated variables into a single model of higher accuracy (for example, using boosting or bagging methods). More formally, the analyst is given a d-dimensional labeled data set S of size $2n$ and splits it randomly into a training set S_t and a holdout set S_h of equal size. We denote an element of S by a tuple (x,y), where x is a d-dimensional vector and $y \in \{-1, 1\}$ is the corresponding class label. The analyst wishes to select variables to be included in the classifier. For various values k of the number of variables to select, the analyst picks k variables with the

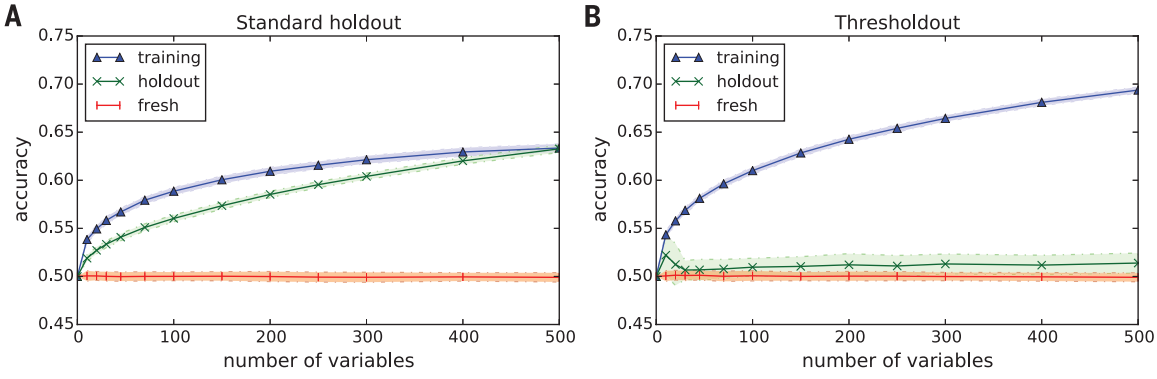

Fig. 1. Learning uncorrelated label. (**A**) Using the standard holdout. (**B**) Using Thresholdout. Vertical axes indicates average classification accuracy over 100 executions (margins are SD) of the classifier on training, holdout, and fresh sets. Horizontal axes show the number of variables selected for the classifier.

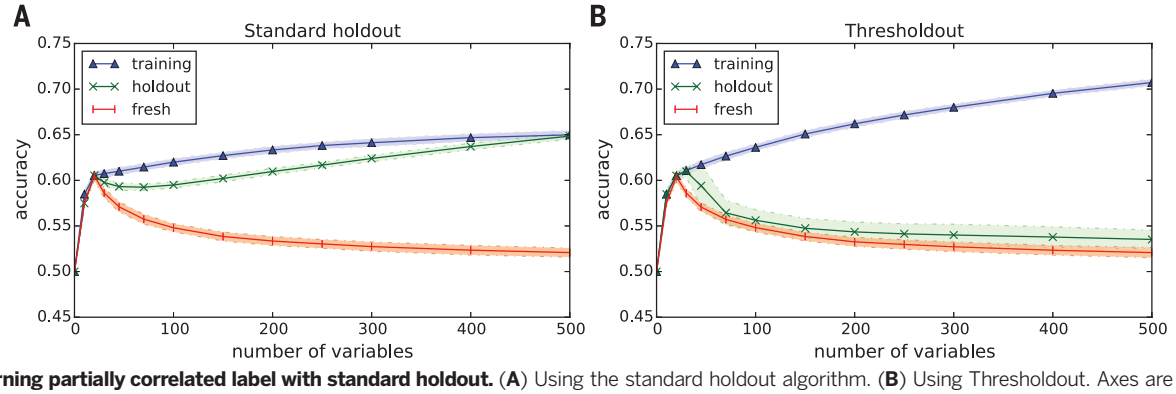

Fig. 2. Learning partially correlated label with standard holdout. (**A**) Using the standard holdout algorithm. (**B**) Using Thresholdout. Axes are as in Fig. 1.

largest absolute correlations with the label. However, he or she verifies the correlations (with the label) on the holdout set and uses only those variables whose correlation agrees in sign with the correlation on the training set and for which both correlations are larger than some threshold in absolute value. The analyst then creates a simple linear threshold classifier on the selected variables using only the signs of the correlations of the selected variables. A final test evaluates the classification accuracy of the classifier on the holdout set. Full details of the analyst's algorithm can be found in section 3 of (*17*).

In our first experiment, each attribute is drawn independently from the normal distribution $N(0,1)$, and we choose the class label $y \in \{-1, 1\}$ uniformly at random so that there is no correlation between the data point and its label. We chose $n = 10{,}000$ and $d = 10{,}000$ and varied the number of selected variables k. In this scenario no classifier can achieve true accuracy better than 50%. Nevertheless, reusing a standard holdout results in reported accuracy of $>63 \pm 0.4\%$ for $k = 500$ on both the training set and the holdout set. The average and standard deviation of results obtained from 100 independent executions of the experiment are plotted in Fig. 1A, which also includes the accuracy of the classifier on another fresh data set of size n drawn from the same distribution. We then executed the same algorithm with our reusable holdout. The algorithm Thresholdout was invoked with $T = 0.04$ and $\tau = 0.01$, which explains why the accuracy of the classifier reported by Thresholdout is off by up to 0.04 whenever the accuracy on the holdout set is within 0.04 of the accuracy on the training set. Thresholdout prevents the algorithm from overfitting to the holdout set and gives a valid estimate of classifier accuracy. In Fig. 1B, we plot the accuracy of the classifier as reported by Thresholdout. In addition, in fig. S2 we include a plot of the actual accuracy of the produced classifier on the holdout set.

In our second experiment, the class labels are correlated with some of the variables. As before, the label is randomly chosen from {–1,1} and each of the attributes is drawn from $N(0,1)$, aside from 20 attributes drawn from $N(y \cdot 0.06, 1)$, where y is the class label. We execute the same algorithm on this data with both the standard holdout and Thresholdout and plot the results in Fig. 2. Our experiment shows that when using the reusable holdout, the algorithm still finds a good classifier while preventing overfitting.

Overfitting to the standard holdout set arises in our experiment because the analyst reuses the holdout after using it to measure the correlation of single attributes. We first note that neither cross-validation nor bootstrap resolve this issue. If we used either of these methods to validate the correlations, overfitting would still arise as a result of using the same data for training and validation (of the final classifier). It is tempting to recommend other solutions to the specific problem on which we based our experiment. Indeed, a substantial number of methods in the statistics literature deal with inference for fixed two-step procedures in which the first step is variable selection [see (*5*) for examples]. Our experiment demonstrates that even in such simple and standard settings, our method avoids false discovery without the need to use a specialized procedure and, of course, extends more broadly. More importantly, the reusable holdout gives the analyst a general and principled method to perform multiple validation steps where previously the only known safe approach was to collect a fresh holdout set each time a function depends on the outcomes of previous validations.

REFERENCES AND NOTES

1. Y. Benjamini, Y. Hochberg, *J. R. Stat. Soc. B* **57**, 289–300 (1995).
2. J. P. A. Ioannidis, *PLOS Med.* **2**, e124 (2005).
3. J. P. Simmons, L. D. Nelson, U. Simonsohn, *Psychol. Sci.* **22**, 1359–1366 (2011).
4. A. Gelman, E. Loken, *Am. Stat.* **102**, 460 (2014).
5. T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Series in Statistics, Springer, New York, ed. 2, 2009).
6. D. Foster, R. Stine, *J. R. Stat. Soc. B* **70**, 429–444 (2008).
7. E. Aharoni, H. Neuvirth, S. Rosset, *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**, 1431–1437 (2011).
8. A. Javanmard, A. Montanari, On online control of false discovery rate. http://arxiv.org/abs/1502.06197 (2015).
9. C. Chambers, M. Munafo, "Trust in science would be improved by study pre-registration," *Guardian US*, 5 June 2013; www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration.
10. J. Reunanen, *J. Mach. Learn. Res.* **3**, 1371–1382 (2003).
11. R. B. Rao, G. Fung, in *Proceedings of the SIAM International Conference on Data Mining 2008* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008), pp. 588–596.
12. G. C. Cawley, N. L. C. Talbot, *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
13. C. Dwork, F. McSherry, K. Nissim, A. Smith, in *Theory of Cryptography* (Lecture Notes in Computer Science Series, Springer, Berlin, 2006), pp. 265–284.
14. O. Bousquet, A. Elisseeff, *J. Mach. Learn. Res.* **2**, 499–526 (2002).
15. T. Poggio, R. Rifkin, S. Mukherjee, P. Niyogi, *Nature* **428**, 419–422 (2004).
16. S. Shalev-Shwartz, O. Shamir, N. Srebro, K. Sridharan, *J. Mach. Learn. Res.* **11**, 2635–2670 (2010).
17. Supplementary materials are available on *Science* Online.
18. D. A. Freedman, *Am. Stat.* **37**, 152–155 (1983).

ACKNOWLEDGMENTS

Authors are listed in alphabetical order. A.R. was supported in part by an NSF CAREER grant (CNS 1253345), NSF grant CCF 1101389, and the Alfred P. Sloan Foundation. T.P. was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. We thank S. Arora, M. F. Balcan, A. Blum, D. Foster, M. Kearns, J. Kleinberg, A. Rakhlin, P. Rigollet, W. Su, and J. Ullman for enlightening discussions about this work. We also thank the Simons Institute for the Theory of Computing at the University of California Berkeley, where part of this research was done.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/349/6248/636/suppl/DC1
Supplementary Text
Figs. S1 and S2
References (*19*–*25*)
Data S1

17 February 2015; accepted 15 June 2015
10.1126/science.aaa9375

ENVIRONMENTAL SCIENCE

Profiling risk and sustainability in coastal deltas of the world

Z. D. Tessler,[1]* C. J. Vörösmarty,[1,2] M. Grossberg,[3] I. Gladkova,[3] H. Aizenman,[3] J. P. M. Syvitski,[4] E. Foufoula-Georgiou[5]

Deltas are highly sensitive to increasing risks arising from local human activities, land subsidence, regional water management, global sea-level rise, and climate extremes. We quantified changing flood risk due to extreme events using an integrated set of global environmental, geophysical, and social indicators. Although risks are distributed across all levels of economic development, wealthy countries effectively limit their present-day threat by gross domestic product–enabled infrastructure and coastal defense investments. In an energy-constrained future, such protections will probably prove to be unsustainable, raising relative risks by four to eight times in the Mississippi and Rhine deltas and by one-and-a-half to four times in the Chao Phraya and Yangtze deltas. The current emphasis on short-term solutions for the world's deltas will greatly constrain options for designing sustainable solutions in the long term.

Deltas present a quintessential challenge for humans amid global environmental change. Home to some of the world's largest urban areas, deltas are also highly dynamic landforms shaped by fluvial and coastal flooding (*1*–*3*). Human activities at the local and regional scales can perturb the water and sedimentary dynamics necessary to maintain a delta's integrity, increasing the rate of relative sea-level rise (RSLR, the combination of land subsidence and offshore sea-level rise) and increasing flood risk (*4*, *5*).

Delta sediments naturally compact over time, requiring new sediment fluxes from the upstream river network and deposition on the delta surface

[1]Environmental CrossRoads Initiative, City University of New York, NY 10031, USA. [2]Department of Civil Engineering, City College of New York, NY 10031, USA. [3]Department of Computer Science, City College of New York, NY 10031, USA. [4]Department of Geological Sciences, University of Colorado–Boulder, Boulder, CO 80309, USA. [5]Department of Civil, Environmental, and Geo- Engineering, University of Minnesota, Minneapolis, MN 55455, USA.
*Corresponding author. E-mail: ztessler@ccny.cuny.edu

The reusable holdout: Preserving validity in adaptive data analysis

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold and Aaron Roth

Science **349** (6248), 636-638.
DOI: 10.1126/science.aaa9375

Testing hypotheses privately

Large data sets offer a vast scope for testing already-formulated ideas and exploring new ones. Unfortunately, researchers who attempt to do both on the same data set run the risk of making false discoveries, even when testing and exploration are carried out on distinct subsets of data. Based on ideas drawn from differential privacy, Dwork *et al.* now provide a theoretical solution. Ideas are tested against aggregate information, whereas individual data set components remain confidential. Preserving that privacy also preserves statistical inference validity.

Science, this issue p. 636