Analyzing Data with Systematic Bias

MANOLIS ZAMPETAKIS
University of California, Berkeley

In many data analysis problems, we only have access to biased data due to some *systematic bias* of the data collection procedure. In this letter, we present a general formulation of systematic bias in data as well as our recent results on how to handle two very fundamental types of systematic bias that arise frequently in econometric studies: *truncation bias* and *self-selection bias*.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics; G.3 [**Probability and Statistics**]: Multivariate Statistics

General Terms: Algorithms, Economics, Theory

Additional Key Words and Phrases: bias, truncation, censoring, self-selection

1. INTRODUCTION

Many problems in data analysis involve the estimation of a property of an unknown probability distribution $\mathcal{P}$, given a set of finite samples from $\mathcal{P}$. In some settings, the goal is to find a full description of the cumulative distribution function or the probability density function of $\mathcal{P}$ (*density estimation*). In other settings, each sample has the form (x, y) and the goal is to estimate the distribution of y given x (*regression or classification*).

A key assumption underlying many widely used methods is that we have access to samples that are *independently and identically distributed* (i.i.d.)—throughout the sampling process, each sample is drawn under the same conditions, it does not affect the rest of the samples, and it is guaranteed to be drawn from the distribution of interest, $\mathcal{P}$. However, this assumption ignores many challenges in the data collection procedure that lead to *biased* or *corrupted* datasets. The presence of bias or corruption in the data can lead to statistical conclusions that are *fallacious* or *unfair* [Mehrabi et al. 2021]. As a result, identifying the sources of bias or corruption, and most importantly developing ways to perform statistical analysis even in the presence of bias is a fundamental problem with many applications in a wide range of scientific areas, including econometrics [Maddala 1986].

Our goal in this letter is to formulate and understand the effects of *systematic bias* in data analysis. In particular, our main focus is on two fundamental types of systematic bias: *truncation bias*, and *self-selection bias* which we introduce in Section 1.1 and Section 1.2, respectively. In Section 2 we provide a general framework that captures systematic bias and we show how truncation and self-selection bias can be realized in this framework. In Section 3, we present some of our recent results on addressing these types of bias for density estimation problems. Finally, we provide open directions in Section 4.

Author's address: `mzampet@berkeley.edu`

1.1 Truncation Bias

Truncation bias occurs when samples falling outside a subset of the support of the population are not observed. The problem of analyzing data with truncation bias has myriad manifestations in economics, social sciences, and all areas of physical sciences, and dates back to famous statisticians like Pearson, Lee, and Fisher [Pearson 1902; Lee 1914; Fisher 1931]. Since then, it has been a central focus of many studies in econometrics [Maddala 1986], epidemiology see [Klein and Moeschberger 2003], and many other scientific fields. Some real world instantiations of truncation bias are the following:

Negative Income Tax Experiment [Hausman and Wise 1977]. In this example, we have a dataset that consists of information about households with low wage rates. The goal of this study is to understand the effect of education level in the annual income of the households. In the dataset of [Hausman and Wise 1977], there was an artificial truncation in the data collection process. In particular, no data points were collected if the dependent variable (annual income) was below 1.5 times the poverty rate. The traditional method for this data analysis tasks is to use the *ordinary least squares (OLS)* regression. It is not hard to see though that in the presence of truncation the OLS outputs solutions that are biased and can lead to fallacious conclusions about the relationship between annual income and education level as was shown in [Hausman and Wise 1977].

Schooling and Earnings of Low Achievers [Hansen et al. 1970]. In this example, we have samples of people that have been rejected from military, i.e. they scored low to the Armed Forces Qualification Test (AFQT), and we wish to estimate the following equation

$$y = f(\text{education}, \text{age}, \dots)$$

where y corresponds to the AFQT score. Again, the data available are truncated due to the fact that the collected data correspond only to people that received a low AFQT score. This truncation makes again the OLS estimator biased, and leads to fallacious conclusions.

Hubble's Law in Astronomy [Woodroofe 1985]. In this example, we have access to astronomical data and our goal is to estimate the relationship between the absolute luminosity (M) of a star and its observed luminosity (m) which also depends on another parameter called the redshift. One main issue with estimating this relationship between m and M is that we only have access to truncated data and in particular we may observe one star only if $m \geq t$ where t is some threshold that depends only on the measurement device.

Efficacy of COVID-19 Vaccine [Dagan et al. 2021]. In many biomedical and epidemiological studies truncation or censoring is a classical type of bias. In this particular study of the efficacy of a COVID-19 vaccine, truncation occurs in the control group of non-vaccinated people because after a while they receive the vaccine.

1.2 Self-selection Bias

Following the example of [Roy 1951], consider a village with two possible occupations: hunting and fishing. Everyone in the village chooses the occupation that maximizes their earning based solely on their own capabilities. Consider now a statistician collecting observations of the earnings and occupations from this village with the goal of estimating a model that predicts the earnings that a particular person would make as a fisherman or as a hunter. It is not hard to see that applying naive statistical analysis techniques, which ignore the self-selection bias, would produce wrong results. Self-selection bias arises in different forms in many fundamental settings:[1]

Imitation Learning. Consider the problem of learning an optimal policy in some contextual bandit setting wherein we observe the arms (e.g. treatments) pulled by an expert (e.g. doctor) in different contexts (e.g. patients). Modeling the reward (e.g. efficacy) from each arm j as an unknown function $f_{w_j}(x, \varepsilon_j)$ of the context x and additional randomness ε_j that the expert might observe but we do not, we assume that the expert selects the arm j with the highest reward $\max_j\{f_{w_j}(x, \varepsilon_j)\}$. Our goal is to learn the underlying models $w_1, \ldots, w_k$ of the arms by observing the expert make decisions in different contexts. This scenario is an instantiation of a statistical analysis task with self-selection bias since we do not get to observe the reward of all the arms $f_{w_1}(x, \varepsilon_1), \ldots, f_{w_k}(x, \varepsilon_k)$, but only the reward of the arm that was selected from the expert, i.e., $\max_j\{f_{w_j}(x, \varepsilon_j)\}$.

Learning from Strategically Reported Data. A widely studied setting featuring self-selected data is one wherein agents strategically chose which data to report. This is a standard challenge in econometrics, which has recently received increased attention in machine learning literature due to the impact of learning-mediated decisions in various contexts; see, e.g., [Hardt et al. 2016; Krishnaswamy et al. 2020; Liu and Garg 2021] and their references. A common example is the reporting of standardized test scores in college admissions, where applicants have a variety of standardized tests available to them, and are only required to report a chosen subset of them.

Learning from Market Data. Following [Fair and Jaffee 1972], consider a linear model of the a market, wherein there is a supply function $S(x, \varepsilon_S) = w_S^{\mathrm{T}} x + \varepsilon_S$ and a demand function $D(x, \varepsilon_D) = w_D^{\mathrm{T}} x + \varepsilon_D$, where x corresponds to a feature vector of the market, w_S and w_D are the coefficients that determine linear functions S and D, and ε_S, ε_D correspond some random noise. If the market is in disequilibrium then supply does not equal demand, i.e., $S(x, \varepsilon_S) \neq D(x, \varepsilon_D)$. So the quantity transacted is $Q(x) = \min\{S(x, \varepsilon_S), D(x, \varepsilon_D)\}$. If we want to estimate w_S, w_D from data of the form $\{(x^{(i)}, Q(x^{(i)}))\}_i$ then this problem can be expressed as a problem of linear regression with self-selection bias [Cherapanamjeri et al. 2022b].

Learning from Auction Data. [Athey and Haile 2002] and a large body of literature in Econometrics consider the problem of learning bid (and valuation) distributions from auction data with partial observability, wherein only the winner of each auction and the price they paid are observed. Consider such observations in repeated first-price auctions. We can cast this problem as an instance of self-selection problem since we only get to see the maximum of the bids at ever iteration. A body of work in the literature has provided estimation and identification results in this setting [Athey and Haile 2007], including recent work of [Cherapanamjeri et al. 2022a] which demonstrates algorithms for estimating the bid distributions non-parametrically (see also Informal Theorem 3.2).

[1] These applications of self-selection bias are from the work of [Cherapanamjeri et al. 2022b]

2. SYSTEMATIC BIAS IN DATA

For simplicity of exposition we present our general framework for the density estimation problem only, but similar formulation can be obtained for regression and classification problems as well, as we discuss in Section 3.1.

Assume that there is an unknown distribution $\mathcal{P}$ with support $S \subseteq \mathbb{R}^d$. The traditional density estimation problem can be formulated as follows.

Definition 2.1 *(Density Estimation).* Let $\varepsilon > 0$, $\mathcal{P}$ a probability distribution that belongs to a family of probability distributions $\mathcal{D}$ and has support $S \subseteq \mathbb{R}^d$. Given as input n i.i.d. samples $x_1, \ldots, x_n$ drawn from $\mathcal{P}$, our goal is to compute a distribution $\mathcal{Q}$ such that $\mathrm{dist}(\mathcal{P}, \mathcal{Q}) \le \varepsilon$ with probability of failure at most 1%, where $\mathrm{dist}(\cdot, \cdot)$ is a distance metric or a divergence between probability distributions, e.g., the total variation distance, or the Kolmogorov distance, or the KL-divergence. If we specify the family of distributions $\mathcal{D}$ and the distance metric dist and there exist a smallest number $f(d, \varepsilon)$ such that for every $n > f(d, \varepsilon)$ the above problem is solvable, then we call $f(d, \varepsilon)$ the sample complexity of this density estimation problem. The running time of the fastest algorithm that takes as input $x_1, \ldots, x_n$ and outputs $\mathcal{Q}$ is the time complexity of this problem.

Remark. A simplification that we make in the formulation above is that we ignore the dependence on the probability of failure, which for the purposes of this letter we assume is a constant, e.g., 1%. In virtually all the settings that we discuss, the probability of failure can be decreased to δ if we pay an additional $\log(1/\delta)$ factor in sample and time complexity.

Example (DKW Inequality). Assume that $d = 1$, $\mathrm{dist}(A, B)$ is the Kolmogov distance, i.e., the maximum difference of the cumulative distribution functions of A and B, and $\mathcal{P}$ is the set of all probability distributions over $\mathbb{R}$. In this setting the celebrated DKW inequality [Dvoretzky et al. 1956] provides a simple algorithm for solving this density estimation problem, with sample and time complexity $O(1/\varepsilon^2)$ which is known to be tight [Massart 1990].

Next, we define the density estimation problem with adversarial corruptions which is a very well studied problem in statistics and machine learning [Huber 2011; Diakonikolas and Kane 2019]. This problem provides some intuition for our formulation of systematic bias.

Definition 2.2 *(Density Estimation with Corruptions).* Let $\alpha > 0$, $\mathcal{P}$ a probability distribution that belongs to a family of probability distributions $\mathcal{D}$ and has support $S \subseteq \mathbb{R}^d$. We are given as input n corrupted samples $z_1, \ldots, z_n$ such that $z_i = h_i(x)$ where $x_1, \ldots, x_n$ are i.i.d. samples drawn from $\mathcal{P}$ and $h_i : S \to S$ are arbitrary unknown functions.[2] For any meaningful estimation to be possible we require that at least $(1 - \alpha) \cdot n$ of the h_i's equal to the identity, i.e, $h_i(x) = x$ for all $x \in S$. Our goal is to compute a distribution $\mathcal{Q}$ such that $\mathrm{dist}(\mathcal{P}, \mathcal{Q}) \le g(\alpha)$ with probability of failure at most 1%. If we specify the family of distributions $\mathcal{D}$, the distance metric dist, and g and there exist a smallest number $f(d, \alpha)$ such that

[2]Observe that since h_i is arbitrary and unknown, the choice of h_i may depend on the rest of the samples $x_1, \ldots, x_{i-1}, x_{i+1}, x_n$ as well.

for every $n > f(d, \alpha)$ the above problem is solvable then we call $f(d, \alpha)$ the sample complexity of this density estimation problem. The running time of the fastest algorithm that takes as input $x_1, \ldots, x_n$ and outputs $\mathcal{Q}$ is the time complexity of this problem.

Example (DKW Inequality Continued). It is easy to see that the the empirical cumulative distribution function that is used to prove the DKW inequality is robust to α fraction of adversarial corruptions. This means that we can solve the density estimation problem with corruptions with $g(\alpha) = \alpha$ and sample and time complexity $O(1/\alpha^2)$.

There are a few things to observe about Definition 2.2.

(1) The problem formulation allows for a very general class of corruptions since we have no restrictions or knowledge about a small fraction of the h_i's.
(2) Most importantly the fraction of corrupted data, i.e., data z_i for which h_i is not equal to the identity, determines the accuracy that we can achieve, i.e., we cannot hope to estimate $\mathcal{P}$ in distance ε unless $\varepsilon \geq g(\alpha)$.

The second point above is what makes the adversarial corruptions framework not applicable in many settings where the data collection procedure introduces bias to all the data and not just a small fraction of them, as the examples from Sections 1.1 and 1.2 illustrate. This leads us to the formulation of density estimation with systematically biased data.

Definition 2.3 *(Density Estimation with Systematic Bias).* Let $\varepsilon > 0$, $\mathcal{P}$ a probability distribution that belongs to a family of probability distributions $\mathcal{D}$ and has support $S \subseteq \mathbb{R}^d$. We are given as input n systematically biased samples $z_1, \ldots, z_n$ such that $z_i = h(x_i)$ where $x_1, \ldots, x_n$ are i.i.d. samples drawn from $\mathcal{P}$ and $h : S \to T \cup \{\perp\}$ is a function that is known to belong to a known family of functions $\mathcal{H}$. Depending on the model we might observe or not observe any z_i with $z_i = \perp$. In that case n is the total number of z_i's observed. Our goal is to compute a distribution $\mathcal{Q}$ such that $\mathrm{dist}(\mathcal{P}, \mathcal{Q}) \leq \varepsilon$ with probability of failure at most 1%. If we specify the family of distributions $\mathcal{D}$, a family of functions $\mathcal{H}$, the distance metric dist and there exist a smallest number $f(d, \varepsilon)$ such that for every $n > f(d, \varepsilon)$ the above problem is solvable then we call $f(d, \varepsilon)$ the sample complexity of this density estimation problem with systematic bias $\mathcal{H}$. The running time of the fastest algorithm that takes as input $x_1, \ldots, x_n$ and outputs $\mathcal{Q}$ is the time complexity of this problem.

The main differences of Definition 2.2 and Definition 2.3 are the following.

(1) In the corruption framework we do not have knowledge about h_i other than that the fraction of corruption is limited, whereas in the systematic bias framework we know that the bias is the same for all the samples, this is why we call it systematic, and also we know the set $\mathcal{H}$ that h belongs to.
(2) On the other hand the important feature of the systematic bias is that in many settings, it allows for estimation up to arbitrarily small error ε, assuming that we have enough samples, even though the bias function h applies to all the data.

We are now ready to show how truncation bias, censoring, and self-selection bias can be expressed in the framework of systematic bias.

Truncation Bias – Censoring. If we know the truncation/censoring then the set of functions $\mathcal{H}$ contains only one function h where $h(x)$ is equal to x if x is inside the survival set $K \subseteq S$ and is equal to $\perp$ otherwise. The difference between truncation and censoring is that for the former we do not observe the $z_i = \perp$ points whereas for the latter we observe them as well.

Self-selection Bias. The density estimation instantiation of self-selection bias appears when x_i is a d-dimensional vector $x_i = (x_{i1}, \ldots, x_{id})$ and

$$h(x_i) = (\max_{j \in [d]} x_{ij}, \arg\max_{j \in [d]} x_{ij}).$$

This corresponds to observing the highest bid and the identity of the highest bidder in a repeated first-price auction and trying to estimate the distribution of each individual agent.

Both of the above problems are impossible if we allow the family of distributions $\mathcal{D}$ unrestricted. As we will see in the next section, to get algorithms with small sample and time complexities we need to assume that $\mathcal{D}$ contains *smooth* distributions for the case of truncation bias or product measures for the case of self-selection bias.

3. RESULTS

Both truncation bias and self-selection bias introduce difficult estimation questions even for the fundamental case where $\mathcal{D}$ is the set of Gaussian distributions, but for simplicity and for consistency with the examples that we presented for classical density estimation and density estimation with corruption we will present our result in the case where $\mathcal{D}$ cannot be parameterized from a small set of parameters. For our results for the Gaussian case we refer to [Daskalakis et al. 2018].

First we need the definition of smooth probability distributions.

Definition 3.1. Let $d = 1$, $S = [0, 1]$, then we say that the family of distributions $\mathcal{D}$ with support S is a *smooth* family of distributions if for every $\mathcal{P} \in \mathcal{D}$ it holds that

(1) $\mathcal{P}$ has a density,

(2) the logarithm of the density of $\mathcal{P}$ is an infinitely differentiable function,

(3) the i-th derivative of the log-density of $\mathcal{P}$ is upper bounded by M^i, for some constant M.

In order to get some intuition of the family of smooth distributions we observe that it contains all the distributions with log-density of the form $f_1(x) + \cdots f_k(x)$ for a finite number k, where f_i can be any of the following:

—a polynomial $\mathrm{poly}(x)$ of constant degree,

—$f_i(x) = \exp(\mathrm{poly}(x))$,

—$f_i(x) = \sin(\mathrm{poly}(x))$.

An example of a class of distributions that are not smooth according to the above definition is the class of distributions with log-density equal to $a \cdot \log(x)$.

We are now ready to present our results for density estimation from truncated data.

INFORMAL THEOREM 3.1. *Assume that we observe n truncated samples $z_1, \ldots, z_n$ from $\mathcal{P} \in \mathcal{D}$, where $\mathcal{D}$ is a smooth family of distributions with support $[0, 1]$ and the truncation is with respect to a known survival set $K \subseteq [0, 1]$. If the measure of K with respect to $\mathcal{P}$ is at least 1% then there exists an algorithm to estimate $\mathcal{P}$ from truncated samples with error ε in total variation distance and with sample and time complexity* $\mathrm{poly}(1/\varepsilon)$.

The surprising conclusion of this theorem is that if we know that the unknown distribution $\mathcal{P}$ is smooth, as per Definition 3.1, then we can identify $\mathcal{P}$ in its whole support $[0, 1]$, even though we observe samples only from a subset K of $[0, 1]$. In other words we can extrapolate and estimate $\mathcal{P}$ even in a region that is completely hidden to us due to the truncation. An intuition for why this is possible comes from Taylor's theorem in calculus. Taylor's theorem suggests that if a function f is sufficiently smooth then the knowledge of the values of all the derivatives of f at a point x_0 is sufficient to determine the values of f in the whole interval $[0, 1]$. The technical contribution of our work is to provide a statistical version of Taylor's theorem where instead of having access to the values of the derivatives, we have access only to truncated samples of the unknown distribution. For more details about Informal Theorem 3.1 and its multi-dimensional generalizations, we refer to [Daskalakis et al. 2021].

Next, we present our results for density estimation from data with self-selection bias.

INFORMAL THEOREM 3.2. *Assume that we observe n samples $z_1, \ldots, z_n$ with the self-selection bias described in the previous section from $\mathcal{P} \in \mathcal{D}$, where $\mathcal{D}$ is the family of product measures over $[0, 1]^d$. Then there exists an algorithm to estimate $\mathcal{P}$ the samples with self-selection bias with error ε in Levy distance[3] and has sample and time complexity $O((1/\varepsilon)^d)$.*

As we already explained the above theorem can be applied to learning from auction data which is a very fundamental problem in econometrics [Athey and Haile 2002]. For a detailed presentation of Informal Theorem 3.2 we refer to [Cherapanamjeri et al. 2022a]. As we show in [Cherapanamjeri et al. 2022a] the exponential dependence on d is necessary for this problem and can only be avoided if we restrict our attention to estimating the probability distribution only on a subset of its support. This can be formulated by changing the distance metric from the Levy distance to one that only measures the difference in a large subset of the support. In that case our sample and time complexity become $\mathrm{poly}(1/\varepsilon)$.

[3]Levy distance is very similar to the Kolmogorov distance but allows for an ε error in the x-axis of the cumularive distribution functions

3.1 Beyond Density Estimation

The main focus of this letter is on density estimation, but similar formulations can be provided in other data analysis tasks like regression and classification. In linear regression, for example, we have access to samples of the form (x, y) where $y = w^T x + \varepsilon$ and our goal is to estimate w. Truncation or self-selection in this setting applies to the dependent set of variables y and we observe $(x, h(y))$ instead of (x, y). In this setting the assumption on the family of probability distributions $\mathcal{D}$ applies to the distribution of the random noise ε. Similar formulations can be done for classification problems as well. For a precise formulation and results on regression and classification problems with systematic bias we refer to [Daskalakis et al. 2019; Ilyas et al. 2020; Daskalakis et al. 2020; Daskalakis et al. 2021; Cherapanamjeri et al. 2022b].

4. OPEN PROBLEMS

As we mentioned in the previous section there are a lot of results for density estimation, regression, and classification problems under truncation or self-selection bias. An interesting direction that is still not well-explored is how truncation or self-selection bias affects the learning algorithms in online or dynamic environments. A first step in this direction has been taken in [Plevrakis 2021], but there are many interesting problems in this area that are still open: (1) Are there online learning or bandit algorithms that are robust to truncation or self-selection bias? For self-selection bias this is closely related with the problem of imitation learning; (2) Can we effectively control dynamical systems for which we cannot observe their state at every time step?

Beyond truncation and self-selection the natural question that arises is the following: Can we provide a characterization of the distribution families $\mathcal{D}$ and the functions $\mathcal{H}$ for which we can solve the density estimation problem with systematic bias?

REFERENCES

Athey, S. and Haile, P. A. 2002. Identification of standard auction models. *Econometrica 70,* 6, 2107–2140.

Athey, S. and Haile, P. A. 2007. Nonparametric approaches to auctions. *Handbook of econometrics 6*, 3847–3965.

Cherapanamjeri, Y., Daskalakis, C., Ilyas, A., and Zampetakis, M. 2022a. Estimation of standard auction models. *arXiv preprint arXiv:2205.02060*.

Cherapanamjeri, Y., Daskalakis, C., Ilyas, A., and Zampetakis, M. 2022b. What makes a good fisherman? linear regression under self-selection bias. *arXiv preprint arXiv:2205.03246*.

Dagan, N., Barda, N., Kepten, E., Miron, O., Perchik, S., Katz, M. A., Hernán, M. A., Lipsitch, M., Reis, B., and Balicer, R. D. 2021. Bnt162b2 mrna covid-19 vaccine in a nationwide mass vaccination setting. *New England Journal of Medicine*.

Daskalakis, C., Gouleakis, T., Tzamos, C., and Zampetakis, M. 2018. Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 639–649.

Daskalakis, C., Gouleakis, T., Tzamos, C., and Zampetakis, M. 2019. Computationally and statistically efficient truncated regression. In *Conference on Learning Theory*. PMLR, 955–960.

Daskalakis, C., Kontonis, V., Tzamos, C., and Zampetakis, E. 2021. A statistical taylor theorem and extrapolation of truncated densities. In *Conference on Learning Theory*. PMLR, 1395–1398.

Daskalakis, C., Rohatgi, D., and Zampetakis, E. 2020. Truncated linear regression in high dimensions. *Advances in Neural Information Processing Systems 33*, 10338–10347.

Daskalakis, C., Stefanou, P., Yao, R., and Zampetakis, E. 2021. Efficient truncated linear regression with unknown noise variance. *Advances in Neural Information Processing Systems 34*, 1952–1963.

Diakonikolas, I. and Kane, D. M. 2019. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 642–669.

Fair, R. C. and Jaffee, D. M. 1972. Methods of estimation for markets in disequilibrium. *Econometrica: Journal of the Econometric Society*, 497–514.

Fisher, R. 1931. Properties and Applications of Hh Functions. *Mathematical tables 1*, 815–852.

Hansen, W. L., Weisbrod, B. A., and Scanlon, W. J. 1970. Schooling and earnings of low achievers. *The American Economic Review*, 409–418.

Hardt, M., Megiddo, N., Papadimitriou, C. H., and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 14-16, 2016*, M. Sudan, Ed. ACM, 111–122.

Hausman, J. A. and Wise, D. A. 1977. Social experimentation, truncated distributions, and efficient estimation. *Econometrica: Journal of the Econometric Society*, 919–938.

Huber, P. J. 2011. Robust statistics. In *International encyclopedia of statistical science*. Springer, 1248–1251.

Ilyas, A., Zampetakis, E., and Daskalakis, C. 2020. A theoretical and practical framework for regression and classification from truncated samples. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4463–4473.

Klein, J. P. and Moeschberger, M. L. 2003. *Survival analysis: techniques for censored and truncated data*. Vol. 1230. Springer.

Krishnaswamy, A. K., Li, H., Rein, D., Zhang, H., and Conitzer, V. 2020. Classification with strategically withheld data. *CoRR abs/2012.10203*.

Lee, A. 1914. Table of the gaussian "tail" functions; when the "tail" is larger than the body. *Biometrika*.

Liu, Z. and Garg, N. 2021. Test-optional policies: Overcoming strategic behavior and informational gaps. In *EAAMO 2021: ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Virtual Event, USA, October 5 - 9, 2021*. ACM, 11:1–11:13.

Maddala, G. S. 1986. *Limited-dependent and qualitative variables in econometrics*. Number 3. Cambridge university press.

Massart, P. 1990. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, 1269–1283.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR) 54,* 6, 1–35.

Pearson, K. 1902. On the Systematic Fitting of Frequency Curves. *Biometrika 2*, 2–7.

Plevrakis, O. 2021. Learning from censored and dependent data: The case of linear dynamics. In *Conference on Learning Theory*. PMLR, 3771–3787.

Roy, A. D. 1951. Some thoughts on the distribution of earnings. *Oxford economic papers 3,* 2, 135–146.

Woodroofe, M. 1985. Estimating a distribution function with truncated data. *The Annals of Statistics 13,* 1, 163–177.