Error Variance, Fairness, and the Curse on Minorities

Beauxis-Aussalet, Emma

published in
Machine Learning and Principles and Practice of Knowledge Discovery in Databases
2021

DOI (link to publisher)
10.1007/978-3-030-93733-1_25

document version
Publisher's PDF, also known as Version of record

Error Variance, Fairness, and the Curse on Minorities

Emma Beauxis-Aussalet(✉)

Vrije Universiteit Amsterdam, De Boelelaan 1111, Amsterdam, The Netherlands
e.m.a.l.beauxisaussalet@vu.nl

Abstract. Machine learning systems can make more errors for certain populations and not others, and thus create discriminations. To assess such fairness issue, errors are typically compared across populations. We argue that we also need to account for the variability of errors in practice, as the errors measured in test data may not be exactly the same in real-life data (called *target* data). We first introduce statistical methods for estimating random error variance in machine learning problems. The methods estimate how often errors would exceed certain magnitudes, and how often the errors of a population would exceed that of another (e.g., by more than a certain range). The methods are based on well-established sampling theory, and the recently introduced Sample-to-Sample estimation. The latter shows that small target samples yield high error variance, even if the test sample is very large. We demonstrate that, in practice, minorities are bound to bear higher variance, thus amplified error and bias. This can occur even if the test and training sets are accurate, representative, and extremely large. We call this statistical phenomenon the curse on minorities, and we show examples of its impact with basic classification and regression problems. Finally, we outline potential approaches to protect minorities from such curse, and to develop variance-aware fairness assessments.

Keywords: Fairness · Transparency · Error variance · Discrimination

1 Introduction

Machine learning systems have been shown to make systematically more errors for certain populations, thus discriminating them [2,12]. Such fairness issues are part of larger societal issues, and cannot be fully addressed by technical solutions [17]. Yet technical solutions for measuring machine learning errors enable a variety of fairness assessments [9,16] but not without concerns about their appropriate use [6]. For instance, the choice of error metrics and sampling approach must be adapted to the domain and context.

In this position paper, we focus on bias defined as systematic error differences among populations. We discuss the statistical methods for estimating the range of error and bias to expect in practice, among random samples (e.g., to audit a model before deployment). We only discuss random error variance, i.e., the error

M. Kamp et al. (Eds.): ECML PKDD 2021 Workshops, CCIS 1525, pp. 352–365, 2021.
https://doi.org/10.1007/978-3-030-93733-1_25

variations that are solely due to the effect of randomly sampling the test data and the real-life data samples.

However narrow or well-know random variance may be, we show its implications for high-level fairness and transparency issues. We argue that i) error variance should be considered as another dimension of fairness, ii) algorithm audits should use appropriate statistics to make error variance transparent, and iii) regulations should adopt variance-aware approaches when setting limitations on error and bias.

We first review prior work on error and bias metrics, and their variance estimation (Sect. 2) before outlining their application to machine learning problems (Sect. 3). We then argue that error variance is an important dimension of fairness issues (Sect. 4).

We demonstrate that unequal variance amplifies bias, and increases the frequency of large error discrepancies. Minorities are by definition smaller populations, and we demonstrate the higher variance and bias that ensue. Such statistical phenomenon occurs regardless of the quality of test or training data, and can be considered as a curse on minorities. We finally outline potential approaches for mitigating such curse, and for informing stakeholders, users, or policy makers.

2 Statistical Theory

We review essential work on error and bias metrics (Sect. 2.1) and their random sample variance (Sects. 2.2) for classification and regression problems.

2.1 Error and Bias Metrics

The errors measured using *test data* are used to estimate of the errors to expect in the *target data* processed in real life.[1] Such error estimation relies on the assumption that the test data is representative of the target data, i.e., test and target data are randomly sampled from the same population. We distinguish two assumptions: simple random sampling, and stratified random sampling. The strata can be protected populations (e.g., potentially discriminated by higher errors) and/or any class of interest (e.g., in classification problems).

In practice, it is often the case that the target data should be considered as drawn from stratified randomly sampling. For instance, certain classes or protected populations can be larger at certain time periods or locations (e.g., more binational citizens at certain zip codes, or more unemployed females in times of crisis). In this case, we must assume that the target data is drawn from a stratified sampling, with strata for each class or sub-population.

In the sections below, we outline which error metrics are not applicable if the target data follows a stratified sampling, and not a simple random sampling.

[1] In this paper, the *target data* is the real-life data processed in practice, not the target variables to predict. The *training data* is not discussed as it is irrelevant for estimating the errors in the target data. Only the *test data* is relevant, and the model may also be semi- or self-supervised.

Classification Problems. Different metrics of classification errors have been established, and address different needs and use cases [8,13,14]. We distinguish three kinds of basic error rates, depending on their denominator (Table 1).

- **Actual class size** $n_{x.}$ as denominator: θ_{xy} (1), e.g., True Positive rates θ_{11}
- **Predicted class size** $n_{.y}$ as denominator: e_{xy} (2), e.g., Precision e_{11}
- **Total sample size** $n_{..}$ as denominator: Accuracy A (3), with the correct classifications n_{xx} as numerator

The latter two (e.g., Precision and Accuracy) are biased estimators if the target data can have highly varying class proportions (e.g., for stratified target samples, with strata of data points from the same true class, with highly varying class sizes). In such case, the error rates θ_{xy} should be used [1].

Table 1. Notation and basic variables of error metrics.

n_{xy}	Number of data points actually belonging to class x and classified as class y (errors if $x \neq y$). In binary problems, n_{11} are True Positives, n_{00} are True Negatives, n_{01} are False Positives, and n_{10} are False Negatives
n_{xx}	Number of correct classifications for class x
$n_{x.}$	Total number of data points actually belonging to class x (i.e., $\sum_y n_{xy} = n_{x.}$)
$n_{.y}$	Total number of data points classified as class y (i.e., $\sum_x n_{xy} = n_{.y}$)
$n_{..}$	Total number of data points (i.e., $\sum_x n_{x.} = \sum_y n_{.y} = n_{..}$)
θ_{xy}	Rate of error n_{xy} on actual class size $n_{x.}$ (1) (e.g., True Positive Rate if $x = y$)
e_{xy}	Rate of error n_{xy} on predicted class size $n_{.y}$ (2) (e.g., Precision if $x = y$)
A	Accuracy (3), i.e., rate of correct classification on total sample size

$$\theta_{xy} = \frac{n_{xy}}{n_{x.}} \quad (1) \qquad e_{xy} = \frac{n_{xy}}{n_{.y}} \quad (2) \qquad A = \frac{\sum_x n_{xx}}{n_{..}} \quad (3)$$

These three kinds of error rates at the core of a variety of bias metrics [16]. Metrics θ_{xy} (1) with actual class size as denominator apply to:

- Predictive Equality or False Positive Rate Balance (with $x = 0$ and $y = 1$).
- Equal Opportunity or False Negative Rate Balance (with $x = 1$ and $y = 0$).
- Equalized Odds, Conditional Procedure Accuracy Equality, or Disparate Mistreatment (with $y = 1$).

Metrics e_{xy} (2) with predicted class size as denominator apply to:

- Predictive Parity or Outcome Test (with $x = 1$ and $y = 1$).
- Condition Use Accuracy Equality (with $x \neq y$).

Finally, accuracy A (3), with total sample size as denominator, applies to Overall Accuracy Equality.

Regression Problems. Metrics of regression errors basically rely on residuals, i.e., the difference between the true and predicted values for a data point i ($r_i = y_i - \widehat{y_i}$). Residuals are typically averaged over data points, and using the absolute or squared values ($|r_i|$ or r_i^2) to avoid signed values. The latter's mean typically tend to zero in regression problems, as an effect of how regressions are fitted by machine learning systems (or other statistical methods). On the contrary, the mean absolute error (MAE $= {}^1\!/_{n_{..}} \sum_i |r_i|$), the mean squared error (MSE $= {}^1\!/_{n_{..}} \sum_i r_i^2$), or the root mean squared error (RMSE $= \sqrt{{}^1\!/_{n_{..}} \sum_i r_i^2}$) have non-null values that represent the range of residuals to expect in a data sample (i.e., error variance MSE and standard deviation RMSE).

Comparing MAE, MSE, or RMSE across protected populations is already a method for considering fairness issues arising from error variance. In this paper, we will introduce additional methods to account for the size and composition of the target samples.

2.2 Estimating Variance

In this section, we first discuss random variance in classification problems, before introducing the Sample-to-Sample method applied to both regression and classification problems.

Classification Problems. Sampling theory defines the random variance of a rate [5] and is applicable to error rates in classification problems. In essence, the smaller the sample, the larger the variance. Applied to the three kinds of errors rates θ_{xy}, e_{xy}, A (1)–(3), simple formulas can estimate the variance of random samples. With specific sampling assumptions, the formulas below can be applied.

$$V\left(\boldsymbol{\theta}_{\boldsymbol{xy}}^{(x.)}\right) = \frac{\theta_{xy}^*(1-\theta_{xy}^*)}{n_{x.}} \tag{4}$$

$$V\left(\boldsymbol{e}_{\boldsymbol{xy}}^{(.y)}\right) = \frac{e_{xy}^*(1-e_{xy}^*)}{n_{.y}} \tag{5}$$

$$V\left(\boldsymbol{A}^{(..)}\right) = \frac{A^*(1-A^*)}{n_{..}} \tag{6}$$

These formulas concern infinite populations or infinite stratum population ($n_{x.}^* \to \infty$, $n_{.y}^* \to \infty$, $n_{..}^* \to \infty$) and do not apply finite population correction.

For **errors rates $\boldsymbol{\theta_{xy}}$** (4), the formula concerns random samples of size $n_{x.}$ drawn from the stratum of data points that actually belong to class x, and denoted with upperscript $^{(x.)}$. The stratum's error rate is θ_{xy}^* (e.g., when $n_{x.} \to \infty$).

For **errors rates $\boldsymbol{e_{xy}}$** (5), the formula concerns random samples of size $n_{.y}$ drawn from the stratum of data points that are classified as class y, denoted with upperscript $^{(.y)}$. The stratum's error rate is e_{xy}^* (e.g., when $n_{.y} \to \infty$).

For **accuracy $\boldsymbol{A}$** (6), the formula concerns simple random samples of size $n_{..}$ drawn from a population of all classes, denoted with upperscript $^{(..)}$. The population accuracy is A^* (e.g., when $n_{..} \to \infty$).

Sample-to-Sample Estimation. When the errors in a test sample are used to estimate the errors in a target sample, one sample's error rate is used as an estimator of another sample's error rate. Such estimators, called *Sample-to-Sample* [1], are impacted by the random variance in both test and target samples. We detail this estimation method using **prime symbols** (e.g., n'_{xy}) when referring to the target sample, and **no prime** when referring to the test sample.

The error differences between the test and target samples vary as the difference of two random variables does (e.g., $\boldsymbol{\theta_{xy}} - \boldsymbol{\theta'_{xy}}$). We assume that the test and target samples are independent (e.g., no overlap) with null covariance, e.g., $\mathrm{V}(\boldsymbol{\theta_{xy}} - \boldsymbol{\theta'_{xy}}) = V(\boldsymbol{\theta_{xy}}) + V(\boldsymbol{\theta'_{xy}})$.

Hence when estimating the error rates in a target sample from the error rates measured in a test sample, the variance of such Sample-to-Sample estimates is the sum of both test and target sample variance (7)–(9). Thus error estimates for small target sets can have large variance, even if the test sample is very large.

$$\widehat{V(\boldsymbol{\theta'^{(x.)}_{xy}})} = V\big(\boldsymbol{\theta^{(x.)}_{xy}}\big) + V\big(\boldsymbol{\theta'^{(x.)}_{xy}}\big) = \frac{\theta^*_{xy}(1-\theta^*_{xy})}{n_{x.}} + \frac{\theta^*_{xy}(1-\theta^*_{xy})}{n'_{x.}} \tag{7}$$

$$\widehat{V(\boldsymbol{e'^{(.y)}_{xy}})} = V\big(\boldsymbol{e^{(.y)}_{xy}}\big) + V\big(\boldsymbol{e'^{(.y)}_{xy}}\big) = \frac{e^*_{xy}(1-e^*_{xy})}{n_{.y}} + \frac{e^*_{xy}(1-e^*_{xy})}{n'_{.y}} \tag{8}$$

$$\widehat{V(\boldsymbol{A'^{(..)}})} = V\big(\boldsymbol{A^{(..)}}\big) + V\big(\boldsymbol{A'^{(..)}}\big) = \frac{A^*(1-A^*)}{n_{..}} + \frac{A^*(1-A^*)}{n'_{..}} \tag{9}$$

It is important to mention, as in Sect. 2.1, that estimates for $\widehat{\boldsymbol{e'^{(.y)}_{xy}}}$ and $\widehat{\boldsymbol{A'^{(..)}}}$ are not applicable if class proportions largely vary among target samples.

Sample-to-Sample in Regression Problems. Sample-to-Sample originally addressed classification problems. Prior work mentions the issue of error variance for regression fairness, but without investigating random sample variance in target samples [7,15]. We thus briefly investigate how residuals in target samples may randomly vary from those measured in a test sample (since the latter is used to estimate the former, i.e., $\widehat{MSE'} = MSE$). The variance of mean squared errors (MSE) is the variance of a mean. According to the central limit theorem, we assume that MSE are normally distributed with $V(\boldsymbol{MSE}) = MSE^{*2}(1/n_{..})$ and $V(\boldsymbol{MSE'}) = MSE^{*2}(1/n'_{..})$.

The error differences between the test and target samples vary as the difference of two random variables (e.g., $\boldsymbol{MSE'} - \boldsymbol{MSE}$). We assume that test and target samples are independent with null covariance, thus $V(\boldsymbol{MSE} - \boldsymbol{MSE'}) = V(\boldsymbol{MSE}) + V(\boldsymbol{MSE'}) = MSE^{*2}(1/n_{..} + 1/n'_{..})$.

In practice, the population MSE^* can be estimated from the test set, i.e., $\widehat{MSE^*} = MSE$. Thus we may apply Sample-to-Sample estimates as in (10), and we evaluate this approach in Sect. 4. Meanwhile, in Sect. 3 we focus on classification problems.

$$\widehat{V}(\widehat{\boldsymbol{MSE'}}) = MSE^2(1/n'_{..} + 1/n_{..}) \tag{10}$$

3 Estimating Variance in Practice

The theory for variance estimation is based on the overall population error rates $(\theta^*_{xy}, e^*_{xy}, A^*)$ which are unknown. In practice, the error rates from the test data can be used to estimate the population error rates [1].

$$\widehat{\theta^*_{xy}} = \theta_{xy} \qquad \widehat{e^*_{xy}} = e_{xy} \qquad \widehat{A^*} = A \tag{11}$$

We can then compute the variance estimates of error rates (Sect. 3.1) and error differences among sub-populations (Sect. 3.2). Using these methods, we demonstrate the practical impacts of variance on minorities, i.e., larger ranges of error and bias occur for smaller sub-populations (Sect. 3.3). Finally, we outline methods for drawing confidence intervals, to visualise the range of error and bias to expect in practice (Sect. 3.4).

3.1 Estimating Error Variance

We can estimate error variance in target samples using equations (7–9) and (11).

$$\widehat{V}(\widehat{\boldsymbol{\theta'}^{(x.)}_{\boldsymbol{xy}}}) = \theta_{xy}(1-\theta_{xy})\left(\frac{1}{n_{x.}} + \frac{1}{\widehat{n'_{x.}}}\right) \tag{12}$$

$$\widehat{V}(\widehat{\boldsymbol{e'}^{(.y)}_{\boldsymbol{xy}}}) = e_{xy}(1-e_{xy})\left(\frac{1}{n_{.y}} + \frac{1}{n'_{.y}}\right) \tag{13}$$

$$\widehat{V}(\widehat{\boldsymbol{A'}^{(..)}}) = A(1-A)\left(\frac{1}{n_{..}} + \frac{1}{n'_{..}}\right) \tag{14}$$

For error rates θ'_{xy} (12) the true class sizes $n'_{x.}$ are unknown. They can be derived as $\widehat{n'_{x.}} = \sum_y e_{xy} n'_{.y}$ but only if class proportions are stable (i.e., test and target data are random samples from the same class distribution). Otherwise $n'_{x.}$ can be estimated using a system of linear equations, which solution is given by inverting the matrix of error rates θ_{xy} and multiplying it with the vector of predicted class sizes $n'_{.y}$ (15). This is called the matrix inversion method [1,3]. Applying it adds another component of variance, which is worth investigating in future research (e.g., many error rates θ_{xy} are involved).

$$\begin{pmatrix} \widehat{n'_{1.}} \\ \widehat{n'_{2.}} \\ \vdots \\ \widehat{n'_{x.}} \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{21} & \dots & \theta_{x1} \\ \theta_{12} & \theta_{22} & \dots & \theta_{x2} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1x} & \theta_{2x} & \dots & \theta_{xx} \end{pmatrix}^{-1} \begin{pmatrix} n'_{.1} \\ n'_{.2} \\ \vdots \\ n'_{.x} \end{pmatrix} \tag{15}$$

3.2 Estimating Bias Variance

So far, we focused on estimating the variance of a single error rate. Now we estimate the variance of bias arising from error discrepancies among protected populations, i.e., the range of error rate differences to expect in practice.

Our notation uses subscripts α and β to represent the populations to compare (e.g., $n'_{..|\alpha}$ is the target sample size for population α and $\theta_{xy|\beta}$ is the error rate for population β). We assume that the error rates of each population are independent (e.g., populations α and β are distinct with no overlap) thus with null covariance, e.g., $V(\boldsymbol{\theta_{xy|\alpha}} - \boldsymbol{\theta_{xy|\beta}}) = V(\boldsymbol{\theta_{xy|\alpha}}) + V(\boldsymbol{\theta_{xy|\beta}})$. Thus the target sample sizes of both populations α and β impact the variance of error differences (16).

$$\widehat{V}\left(\widehat{\boldsymbol{\theta'}^{(x.|\alpha)}_{xy|\alpha}} - \widehat{\boldsymbol{\theta'}^{(x.|\beta)}_{xy|\beta}}\right) = (\theta_{xy|\alpha} - \theta^2_{xy|\alpha})\left(\frac{1}{n_{x.|\alpha}} + \frac{1}{\widehat{n'_{x.|\alpha}}}\right) + (\theta_{xy|\beta} - \theta^2_{xy|\beta})\left(\frac{1}{n_{x.|\beta}} + \frac{1}{\widehat{n'_{x.|\beta}}}\right)$$

$$\widehat{V}\left(\widehat{\boldsymbol{e'}^{(.y|\alpha)}_{xy|\alpha}} - \widehat{\boldsymbol{e'}^{(.y|\beta)}_{xy|\beta}}\right) = (e_{xy|\alpha} - e^2_{xy|\alpha})\left(\frac{1}{n_{.y|\alpha}} + \frac{1}{n'_{.y|\alpha}}\right) + (e_{xy|\beta} - e^2_{xy|\beta})\left(\frac{1}{n_{.y|\beta}} + \frac{1}{n'_{.y|\beta}}\right)$$

$$\widehat{V}\left(\widehat{\boldsymbol{A'}^{(..)}_{\alpha}} - \widehat{\boldsymbol{A'}^{(..)}_{\alpha}}\right) = (A_\alpha - A^2_\alpha)\left(\frac{1}{n_{..|\alpha}} + \frac{1}{n'_{..|\alpha}}\right) + (A_\beta - A^2_\beta)\left(\frac{1}{n_{..|\beta}} + \frac{1}{n'_{..|\beta}}\right) \quad (16)$$

3.3 The Case of Minorities

The formulas above show that the variance of bias (i.e., of error differences) depends on the sample sizes of both populations α and β. If a population α is a minority, its size is smaller than a population β. Thus the bias to expect may have larger variance than if α and β had similar sizes.

For example, we can easily show a case where the bias variance is larger if one population is a minority, compared to populations of equal sizes (unless circumstances are unrealistic, e.g., extremely different or low accuracy). Let's consider accuracy in a target set of $n'_{..} = 1000$ data points with 2 sub-populations. The variance due to the test set is considered null (e.g., $n_{..} \to \infty$) and we use $a = A'_\alpha - A'^2_\alpha$ and $b = A'_\beta - A'^2_\beta$.

- For a minority/majority split into $n'_{..|\alpha} = 100$ and $n'_{..|\beta} = 900$, the random variance from the target sample is approximately $0.01a + 0.0011b$.
- For an equal split into $n'_{..|\alpha} = n'_{..|\beta} = 500$, the variance is $0.002a + 0.002b$.
- If $a = b$, there is more variance in the case of a minority ($0.011a > 0.004a$).
- In general, there is more variance in case of a minority unless $b > 98a$. This is unrealistic since $b < 1$ and $a < 1$, so accuracies would be extremely low. And accuracies would be so different that a fairness issue is already blatant.

This demonstration may not hold for all error rates and population sizes. Future work is needed to formally specify the conditions under which the presence of a minority implies higher bias than more balanced populations. However, we can already observe that the presence of a minority can have significant impacts on the range of bias to expect in practice.

3.4 Confidence Intervals

To communicate how frequently certain magnitudes of error can occur, we can use confidence intervals (CI). Such intervals represent a range of values to expect for a specific proportion of the data samples (having the same sample sizes). For examples, with a confidence level of 50%, the interval represents the range of errors that occur in about 50% of the random samples. Errors are lower than the lower bound in 25% of data samples, and higher than the lower bound in the other 75%.

For example, in Fig. 1 we draw confidence intervals with various confidence levels. They show how frequently certain ranges of error occur among random samples. In this example, True Positive rates lower than 0.89 occur in 25% of female samples (lower bound of 50% CI) but only in 5% of male samples (lower bound of 90% CI). This can be considered as unfair.

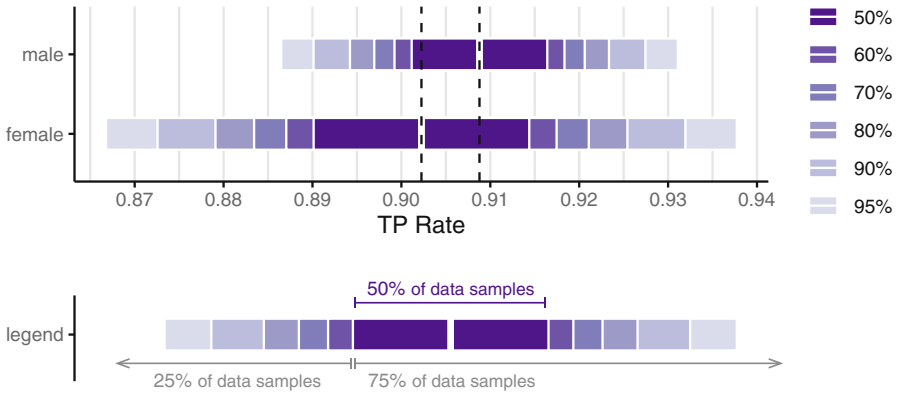

Fig. 1. Example of True Positive (TP) rates θ_{11} compared for female and male populations, with confidence intervals showing the range of error to expect in practice (e.g., 50% intervals show ranges expected for 50% of data samples). We show 50% to 95% CI drawn with the normal approximation and $z \in \{0.67, 0.84, 1.04, 1.28, 1.64, 1.96\}$. *(We acknowledge that this figure omits other genders, and shows only two for simplicity.)*

Specific statistical methods have been established to draw confidence intervals for error rates [11] or differences between error rates [10]. The normal approximation is the simplest method: $\boldsymbol{\theta'_{xy}} \pm z\sqrt{V(\boldsymbol{\theta'_{xy}})}$ e.g., with $z = 1$ for 68% CI, and $z = 1.96$ for 95% CI.

The normal approximation may be the most understandable, e.g., to non-experts. But it is unreliable for small samples (e.g., $n'_{x.} < 30$) and extreme error rates (e.g., $\theta_{xy} \to 0$). In such cases, more complex methods are required such as Wilson scores or Clopper-Pearson estimates [4,18].

Note that in the case of target samples with highly varying class proportions (i.e., stratified sampling), it is complex to account for the variance of estimated class sizes $\widehat{n'_{x.}}$ when the matrix inversion method is required (15). Binary problems can be solved algebraically and Fieller's theorem is applicable [1]. But formulas are complex to derive for multiclass problems, and bootstrapping methods are recommended [3].

4 Error Variance and Fairness

We argue that fairness assessments must consider random error variance, or they may fail to assess the practical risks of bias. Without information on variance, small error differences may seem negligible. However, random error variance makes it possible for large error differences to occur frequently (e.g., depending on sample sizes). In other words, even if error differences remain small on average, large bias and discrimination may occur on occasion. We argue that it is important to estimate how often such occasional circumstances would occur.

To support our argument, we first outline the practical impacts of random error variance, and how human expectations or regulations may overlook such issue (Sect. 4.1). We then discuss a practical case showing how bias is amplified by random variance, in a simple classification problem (Sect. 4.2). Finally, we discuss variance and fairness in regression problems, and test the application of Sample-to-Sample variance estimation (Sect. 4.3).

4.1 Problem Statement

Practical Impacts. With large error variance, a system may largely misclassify an entire population, e.g., for a given batch of data processed at a specific month and at a specific location. The societal impacts can be high: even though such circumstances occur at random, their occurrence can have significant repercussions at the scale of a society. For instance, the medical tests of a hospital could massively fail in a given month. Should an epidemic also strike at the same time, the impacts would be critical.

Minorities, by definition, have smaller populations sizes, and thus larger error variance than a majority group would. In practice, with minorities we run the risk of impacting entire communities, e.g., a local community in a given neighbourhood, at a given period of time. For example, frauds could be massively over-estimated in a local community, which may already suffer from other discriminations for being a minority. Large error discrepancies may be random, but they can have significant repercussions at the scale of a small community, and may not be as rare as expected.

In essence, higher error variance implies higher risks of discrimination, and minorities naturally have higher error variance than majorities. Thus estimating the random variance of errors is needed to inform impact assessments, to prevent discriminations, and to make systems more transparent.

Variance-Aware Human Expectations. It is important that the humans who use, manage, or audit AI systems are aware of the range of errors to expect, especially for small data samples. Stakeholders and decision-makers may not be aware that errors can be much larger than measured on a test sample. Humans may not expect such phenomenon, or may believe it is more rare than it actually is, especially if the test sample is very large.

The users and controllers of an AI system should be aware that random error variance can be significant, that it also depends on the size of target samples, and that therefore minorities have larger variance. Furthermore, they should be aware of the precise range of errors to expect, and at which frequency. This information can be estimated from equations (12)–(14) and (16), and visualized using confidence intervals (Sect. 3.4). Provided with these, humans may understand the range of error and bias to expect in practice.

This is a first step towards informing human expectations, but it is not sufficient. Confidence intervals may be relatively intuitive, but they are not self-explanatory: the information, the visualization, or the terminology may be misunderstood. Moreover, many other aspects must be considered to inform fairness assessments (e.g., test set quality, actual demographics, human recourse and intervention). Future work is required to design visualizations and tutorials, and to explore how fairness criteria and regulations can be adapted.

Variance-Aware Regulations. We focus on a form of regulation that uses thresholds for limiting the error to tolerate (e.g., no less than 90% TP rate), or for limiting the error differences between populations (e.g., bias of maximum -1% TP rate). We argue that a single threshold is not enough to limit error and bias in practice, and that error variance should be considered.

A threshold may be met when measuring errors in a single test sample. But in real life, error magnitudes randomly vary from sample to sample. Hence, the threshold may be passed occasionally, for some samples.

We argue that it is important to know at which frequency a limit would be violated, and to regulate such frequency too. For instance, regulations can specify the frequency at which a threshold may be passed, as in the examples below.

- To limit error: ***TP Rate should be higher than 0.9*** *and* ***no more than 5% of data samples*** *(of size* $n'_{x.}$*)* ***should exceed this limit***.
- To limit bias: ***TP Rate differences should not exceed*** $\pm$***0.1*** *and* ***no more than 5% of data samples*** *(of size* $n'_{x.|\alpha} + n'_{x.|\beta}$*)* ***should exceed this limit***.
- To limit bias: ***The extreme range of error occurring in 5% of male samples*** *(of size* $n'_{x.|\alpha}$*)* ***should not occur in more than 20% of female samples*** *(of size* $n'_{x.|\beta}$*)*.
- To limit bias: ***The 60% CI for females*** *(for sample size* $n'_{x.|\beta}$*)* ***should be contained within the 90% CI for males*** *(for sample size* $n'_{x.|\alpha}$*)*.

To be applicable, such regulations need to specify the sample sizes of interest (e.g., the target sample size $n'_{x.|\alpha}$ for TP rates). Choosing the sample sizes may be arbitrary, but it is necessary, as one could tweak the sample sizes to escape the regulation (e.g., if a limit is not met for a small sample, it may be met for a larger sample). Future work is required for establishing realistic and meaningful sample sizes, and these decisions highly depend on the domain and context.

Among potential approaches, we could use the typical demographics of the impacted populations. For instance, the typical populations in geographic areas of interest, e.g., at the scale of neighbourhoods, cities, or regions. The sample sizes can also correspond to batches of data that are typically processed within a time period, e.g., monthly, weekly, or daily. In a domain-agnostic approach, we could use sample sizes of 100 to reflect the intuition behind expressing error rates in percentages. We explore the latter in the example below (Sect. 4.2).

4.2 Variance and Fairness in Classification Problems

A Simple Example. Let's consider a classifier that detects high-risk patients in hospitals, and that has been tested on thousands of patients. To prevent high-risk patients from remaining undetected (False Negatives, FN), doctors may decide to re-examine the uncertain cases, e.g., within a daily arrival of patients. To understand how many high-risk patients can remain undetected, doctors should be aware of the error rate, and the frequency at which higher error rates may occur.

The True Positive (TP) rate of 90% may be reliable, with limited random variance due to the test set size (e.g., ±1 standard deviation). But it does not represent the range of error to expect in practice. For example, it does not mean that about 10 ±1 patients remain undetected (FN) in a daily batch of 100 high-risk patients. It does not mean either that 10 ±2 FN occur in 95% of the daily batches of 100 high-risk patients.

Variance in Practice. For daily data samples of 100 high-risk patients,[2] we can expect 7 to 13 undetected patients out of 100 for 68% of the days (about 5 days a week), and **13 to 16 undetected patients out of 100 in about 15% of the days** (about once a week). There would be more than 16 undetected patients in 2.5% of the days (about once a month).

For a minority group of 10 high-risk patients,[3] we can expect up to 2 undetected patients out of 10 (80% TP rates) in about 68% of the days[4], and **2 to 3 undetected patients out of 10 (70–80% TP rates) in about 15% of the days** (about once a week).

[2] Sample-to-Sample estimation: $(0.1 \times 0.9)/100 = 0.0009$ and $\sqrt{0.0009 + 0.01^2} \approx 0.03$. So the 68% CI is about 10 ± 3 FN patients out of 100, and the 95% CI is 10 ± 6.

[3] $(0.1 \times 0.9)/10 = 0.009$ and $\sqrt{0.009 + 0.01^2} \approx 0.1$. So the 68% CI is about 1 ± 1 FN patients out of 10, and the 95% CI is 1 ± 2.

[4] The normal approximation of CI is not exact with so few items, especially for the lower bound (e.g., reaching negative values). This higher bound is still informative.

This is **much more than stated in the test results** where the standard deviation for the test sample alone is ±1%. The standard deviations for samples of 100 and 10 high-risk patients are ±3% and ±10% respectively.

Humans may not expect such magnitudes of error to occur so frequently, as it exceeds what is stated in the test results. Furthermore, it illustrates a fairness issue that is inherent to random variance: larger error rates occur more often for a minority than for a majority, and this too is unfair.

4.3 Variance and Fairness in Regression Problems

We evaluated Sample-to-Sample applied to regression problems with the approach in equation (10), Sect. 2.

We simulated a simple regression problem with a single feature $x \in [0, 10]$ with $y_i = 2x + 1 + \epsilon$ and $\widehat{y_i} = 2x + 1$. The noise ϵ is normally distributed $\epsilon \sim N(0, \sigma)$ with $\sigma \in \{1, 1.5, 2\}$. It generates population residuals with mean squared error $MSE^* \in \{1, 2.25, 4\}$.

We used test samples with $n_{..} \in \{1\,000, 10\,000\}$ and target samples with $n'_{..} \in \{100, 1\,000\}$. To compare the residuals in test and target samples, we drew 10 000 pairs of test and target samples of both sizes (40 000 pairs in total).

We assumed that larger range of MSE and differences $MSE - MSE'$ occur with smaller target samples. Our results are consistent with this assumption, but not with the formulas we proposed (10).

The actual residuals did not follow $V(\boldsymbol{MSE'}) = MSE^{*2}(1/n'_{..})$ and instead we observed $V(\boldsymbol{MSE'}) \approx 2\,MSE^{*2}\,(1/n'_{..})$. The estimated residuals $\widehat{\boldsymbol{MSE'}}$ did not follow $\widehat{V}(\widehat{\boldsymbol{MSE'}}) = MSE^{\,2}(1/n_{..} + 1/n'_{..})$ and instead we observed $V(\widehat{\boldsymbol{MSE'}}) \approx 2\,MSE^{\,2}\,(1/n_{..} + 1/n'_{..})$.

Future work is needed to establish the theory that governs these observations. However, our empirical observations show the same variance issues as for classification problems: the variance estimation for target samples depends on the size of both test and target samples. The variance of residuals is larger for smaller target samples, and thus for minorities.

5 Conclusion

We demonstrated that error variance depends on the size of the data processed in real life (called the target sample). Variance increases as a target sample size decreases, and this particularly impacts minorities, since their sample size is consistently lower than majorities. This statistical phenomenon happens even if the test data is extremely large, or has balanced sample size for minorities and majorities.

For minorities, the error variance is naturally higher in practice. This can be considered as a curse on minorities: it is beyond what humans can control, and it impacts communities that are likely to be discriminated otherwise. Therefore, we argue that error variance is an important aspect of fairness assessment, as it creates and amplifies error discrepancies, and thus potential discriminations.

Thus it is important to estimate error variance and confidence intervals, and to account for the actual composition of the real-life populations (e.g., with target sample sizes that are representative of minorities and majorities). However, future work is required to specify the sample and population sizes that are typically encountered in practice, e.g., depending on the application domain.

We also discussed how fairness criteria could take into account error variance, beyond using thresholds to limit error and bias. We argue that fairness criteria should also consider the frequency at which a given threshold may be violated, due to random error variations. Fairness criteria can also rely on drawing confidence intervals with different confidence levels (Fig. 1), to compare the bounds the frequency of certain ranges of error. However, future work is required to design visualizations that are usable and understandable, and to develop guidelines for applying meaningful criteria.

Finally, we recommend that error variance be made transparent and understandable to all users and controllers of AI systems. Indeed, fair system or not, it is important that all stakeholders understand the actual range of errors to expect in practice.

References

1. Beauxis-Aussalet, E., Hardman, L.: Extended methods to handle classification biases. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA) (2017)
2. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, pp. 77–91 (2018)
3. Buonaccorsi, J.P.: Measurement Error: Models. Methods and Applications. CRC Press,Taylor and Francis, Boca Raton (2010)
4. Clopper, C.J., Pearson, E.S.: The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika **26**(4), 404–413 (1934)
5. Cochran, W.G.: Sampling Techniques. John Wiley & Sons, Hoboken (2007)
6. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: a critical review of fair machine learning (2018). arXiv preprint arXiv:1808.00023
7. Foulds, J.R., Islam, R., Keya, K.N., Pan, S.: Bayesian modeling of intersectional fairness: the variance of bias*. In: Proceedings of the 2020 SIAM International Conference on Data Mining, pp. 424–432. SIAM (2020)
8. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. Int. J. Data Mining Knowl. Manage. Process. IJDKP **5**(2), 1–10 (2015)
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning (2019). arXiv preprint arXiv:1908.09635
10. Newcombe, R.G.: Interval estimation for the difference between independent proportions: comparison of eleven methods. Statist. Med. **17**(8), 873–890 (1998)
11. Newcombe, R.G.: Two-sided confidence intervals for the single proportion: comparison of seven methods. Statist. Med. **17**(8), 857–872 (1998)
12. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)

13. Sebastiani, F.: An axiomatically derived measure for the evaluation of classification algorithms. In: International Conference on The Theory of Information Retrieval (2015)
14. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. **45**(4), 1–14 (2009)
15. Steinberg, D., Reid, A., O'Callaghan, S., Lattimore, F., McCalman, L., Caetano, T.: Fast fair regression via efficient approximations of mutual information (2020). arXiv preprint arXiv:2002.06200
16. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pp. 1–7. IEEE (2018)
17. Wieringa, M.: What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 1–18 (2020)
18. Wilson, E.B.: Probable inference, the law of succession, and statistical inference. J. Am. Statist. Assoc. **22**(158), 209–212 (1927)