

Statistical Methods for Assessing Differences in False Non-Match Rates Across Demographic Groups

Michael Schuckers

Mathematics, Computer Science and Statistics
St. Lawrence University
Canton, NY, USA

Email: schuckers@stlawu.edu

Sandip Purnapatra, Kaniz Fatima, Daqing Hou, Stephanie Schuckers

Computer and Electrical Engineering,
Clarkson University
Potsdam, NY, USA

Email: {purnaps, fatimak, dhou, sschucke}@clarkson.edu

Abstract—Biometric recognition is used across a variety of applications from cyber security to border security. Recent research has focused on ensuring biometric performance (false negatives and false positives) is fair across demographic groups. While there has been significant progress on the development of metrics, the evaluation of the performance across groups, and the mitigation of any problems, there has been little work incorporating statistical variation. This is important because differences among groups can be found by chance when no difference is present. In statistics this is called a Type I error. Differences among groups may be due to sampling variation or they may be due to actual difference in system performance. Discriminating between these two sources of error is essential for good decision making about fairness and equity. This paper presents two novel statistical approaches for assessing fairness across demographic groups. The first methodology is a bootstrapped-based hypothesis test, while the second is simpler test methodology focused upon non-statistical audience. For the latter we present the results of a simulation study about the relationship between the margin of error and factors such as number of subjects, number of attempts, correlation between attempts, underlying false non-match rates(FNMR's), and number of groups.

I. INTRODUCTION

Biometric recognition is a technology that has broad application for border security, e-commerce, financial transactions, health care, and benefit distribution. With its explosion in use, there are concerns about the fairness of solutions across the broad spectrum of individuals, based on factors such as age, race, ethnicity, gender, education, socioeconomic status, etc. In particular, since biometric recognition has a possibility of error, both false negatives (false rejection) and false positives (false acceptance), the expectation is that solutions have performance which are “fair” across demographic groups. Buolamwini, et. al. found that gender classification based on a single face image had a higher error rate for darker-skinned females with a high 34.7% error rate, compared to other groups (intersections of skin types and genders) [1], [2]. While focused on gender classification rather than face recognition, these papers brought considerable attention to this issue. Others found demographic differences in face recognition for some algorithms and systems [3], [4].

To quantify the equitability of the various face recognition algorithms, multiple metrics have been proposed to evaluate fairness. Proposed Fairness Discrepancy Rate (FDR) weights the two types of errors seen in biometric recognition (false accept and false reject rates), either equally or otherwise, and balances FDR across groups [5]. The U.S. National Institute of Standards and Technology (NIST) introduced the Inequity Rate (IR) metrics for face recognition algorithm performance testing and [6] proposed two interpretability criterion for biometric systems, i.e. Functional Fairness Measure Criteria (FFMC) and Gini Aggregation Rate for Biometric Equitability (GARBE). In other artificial intelligence (AI) research, evaluation metrics include demographic parity, equalized odds, and equal opportunity [7] [8] [9] [10].

However, with all of these active research and analyses, there has been limited contribution towards recommending appropriate statistical methods for determining when two or more groups are “equal” or not. This is essential, as any metric when measured in a sample, will have uncertainty which is a function of variability, correlation, number of groups, and other factors. This uncertainty can be measured through statistical methods, e.g. confidence intervals, to determine the likelihood that differences are found by chance or are a true difference. Given that exact “equality” is unlikely, if not impossible, for a set of groups, these methods allow for appropriate conclusions to be drawn from results.

This paper focuses on statistical methods for fairness solely for false negatives. Biometric solutions used widely by the public are typically based on “verification” or one-to-one matching. A false negative error is when the correct individual is falsely rejected, e.g., does not match their enrollment on a mobile device, passport, bank, or government benefits provider. This “error” may block an individual from accessing benefits which they are entitled to. The goal of this paper is to consider the statistical methods to address differences in false non-match rates based on number of subjects, number of attempts, correlation in attempts, and number of possible demographic groups in the test. The number of subjects and number of attempts can decrease the variability as the number

of subjects and attempts increase; whereas, incorporating the correlation between number of attempts may increase sample variability. Most importantly, the number of demographic groups being compared impacts the variation as an increased number of groups increases the chances that a difference between groups may be found “by chance”, and thus adjustments need to be made in the test due to this effect, often called multiplicity [11].

This paper develops two approaches to detecting differences in FNMR’s between demographic groups. Additionally, we explore the trade-off among variation parameters based on simulations of a hypothetical equity study. In addition to giving guidance on expected outcomes of such a study, this paper will provide suggestions for “practical” thresholds that could be used for when to say that a group is different that would minimize the possibility that that difference was based upon chance alone. In the next section we discuss related statistical work that has been done to assess differences in FNMR’s between groups. Section III introduces the basic statistical structures needed to estimate variation in FNMR estimation. A bootstrap hypothesis test for the equality of FNMR across G groups is presented in Section IV, as well as a simplified alternative that yields a margin of error for detecting differences among groups. That section also includes results of a simulation study. We summarize and discuss this work and possible alternatives in Section V.

II. RELATED WORK

In this section, we discuss other work on statistical methods for comparison of bioauthentication across demographic groups. The NIST Information Technology Laboratory (ITL) quantifies the accuracy of face recognition algorithms for the demographic groups of sex, age, and race [3]. A component of the evaluation focuses on FNMR for one-to-one verification algorithms on four large datasets of photographs collected in U.S. governmental applications (domestic mugshots, immigration application photos, border crossing, and visa applications). For high-quality photos, FNMR was found to be low and it is fairly difficult to measure false negative differentials across demographics. Compared to high-quality application photos, the FNMR is higher for lower-quality border crossing images. Similar observations regarding image quality have been made by others, e.g. [4]. A measure of uncertainty is calculated for each demographic group based on a bootstrapping approach. In bootstrapping, the genuine scores are sampled 2000 times and the 95% interval is plotted providing bounds for each group. No method was presented to suggest when an algorithm might be “fair” under uncertainty. A notional approach might be to declare an algorithm fair if the intervals plotted overlap across all combinations of groups. This, however, does not fully address the possibility of Type I errors.

Cook et al. [4] examined the effect of demographic factors on the performance of the eleven commercial face biometric systems tested as part of the 2018 United States Department of Homeland Security, Science and Technology Directorate (DHS S&T) Biometric Technology Rally. Each participating

system was tasked with acquiring face images from a diverse population of 363 subjects in a controlled environment. Biometric performance was assessed by measuring both efficiency (transaction times) and accuracy (mated similarity scores using a leading commercial algorithm). The authors quantified the effect of relative facial skin reflectance and other demographic covariates on performance using linear modeling. Both the efficiency and accuracy of the tested acquisition systems were significantly affected by multiple demographic covariates including skin reflectance, gender, age, eyewear, and height, with skin reflectance having the strongest net linear effect on performance. Linear modeling showed that lower (darker) skin reflectance was associated with lower efficiency (higher transaction times) and accuracy (lower mated similarity scores) [4]. While statistical significance of demographic factors was considered based on a linear model of match scores, this approach may not be applicable for assessing commercial systems which operate at a fixed threshold.

de Freitas Pereira and Marcel [5] introduce the Fairness Discrepancy Rate (FDR) which is a summary of system performance accounting for both FNMR and FMR. Their approach uses a “relaxation constant” rather than trying to assess the sampling variation or statistical variation between FNMR’s from different demographic groups. Howard et al. [6] present an evaluation of FDR noting its scaling problem. To address this scaling problem, the authors propose a new fairness measure called Gini Aggregation Rate for Biometric Equitability (GARBE).

Other research has also performed extensive evaluations of face recognition across demographic groups, e.g. [12], but have not presented statistical methods as part of their work.

III. VARIANCE AND CORRELATION STRUCTURE OF FNMR

Statistical methods for estimation of FNMR’s are dependent upon the variance and correlation of matching decisions. In this section, we present the basic statistical structures for a single FNMR following [13]. This structure forms the basis for the statistical methods that we present in the next sections. Let D_{ij} represent the decision for the j^{th} pair of captures or signals collected on the i^{th} individual, where n is the number of individuals, $i = 1, \dots, n$ and $j = 1, \dots, m_i$. Thus, the number of sample pairs that are compared for the i^{th} individual is m_i , and n is the number of different individuals being compared. The use of m_i implies that we are allowing the number of comparisons made per individual to vary across individuals. We then define

$$D_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ pair of signals from individual } i \\ & \text{is declared a non-match,} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

We assume for the D_{ij} ’s that $E[D_{ij}] = \pi$ and $V[D_{ij}] = \pi(1 - \pi)$ represent the mean and variance, respectively. Thus, π represents the FNMR. We assume that we have a stationary matching process within each demographic group and implicit in this assumption is that we have a fixed threshold within each

group. Our estimate of π , the process FNMR, will be the total number of errors divided by the total number of decisions:

$$\hat{\pi} = \left[\sum_{i=1}^n \sum_{j=1}^{m_i} D_{ij} \right] / \left[\sum_{i=1}^n m_i \right]. \quad (2)$$

Following Schuckers [14], [15], we have the following correlation structure for the D'_{ij} 's:

$$\text{Corr}(D_{ij}, D_{i'j'}) = \begin{cases} 1 & \text{if } i = i', j = j' \\ \rho & \text{if } i = i', j \neq j' \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

This correlation structure for the FNMR is based upon the idea that there will only be correlations between decisions made on signals from the same individual but not between decisions made on signals from different individuals. Thus, conditional upon the error rate, there is no correlation between decisions on the i^{th} individual and decisions on the i'^{th} individual, when $i \neq i'$. The degree of correlation is summarized by ρ .

Then we can write the variance of $\hat{\pi}$, the estimated FNMR, as

$$V[\hat{\pi}] = N_\pi^{-2} \pi(1 - \pi) [N_\pi + \rho \sum_{i=1}^n m_i(m_i - 1)] \quad (4)$$

where $N_\pi = \sum_{i=1}^n m_i$. An estimator for ρ is given by:

$$\hat{\rho} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{j'=1}^{m_i} (D_{ij} - \hat{\pi})(D_{ij'} - \hat{\pi})}{\hat{\pi}(1 - \hat{\pi}) \sum_{i=1}^n m_i(m_i - 1)}. \quad (5)$$

Models like that found in (3) are known as intra-individual or intra-class models and have been studied extensively in the statistics literature, e.g. Fleiss et al. [16], Williams [17] or Ridout et al. [18]. The parameter ρ in the models above represents the intra-class correlation. This measures the degree of similarity between the decisions made on each individual. If the decisions on each individual are varying in a way that suggests that the decisions are not dependent upon the individual then ρ is zero, meaning that the observations are uncorrelated. Negative values of ρ are possible but such values suggest that decisions on signals from the same individual are less similar to each other than they are to all of the other decisions. This seems unlikely to be the case in the context of biometric authentication. Several authors, including Fleiss et al. [16], have suggested using the following alternative way of writing (4)

$$V[\hat{\pi}] = N_\pi^{-1} \pi(1 - \pi)(1 + (m_0 - 1)\rho) \quad (6)$$

where $m_0 = \frac{\sum_{i=1}^n m_i^2}{N_\pi}$. If $m_i = m$ for all i , then $N_\pi = nm$ and the variance of $\hat{\pi}$ from (6) becomes $V[\hat{\pi}] = (nm)^{-1} \pi(1 - \pi)(1 + (m - 1)\rho)$.

The intra-class correlation has a direct relationship with the variance of $\hat{\pi}$. As ρ increases, the variance in both cases increases. This is a consequence of the lack of independent information from each individual. If ρ is large, then each additional decision on a previously observed individual is providing little new information.

IV. STATISTICAL METHODS FOR MULTIPLE FNMR'S

To evaluate and assess if different FNMR's are *detectably* different¹, we need to understand the variation due to sampling. In equity studies across different demographic groups, we need to account for the sampling variation from each of the G groups. For what follows we will assume that there are G demographic groups across multiple dimensions. For example, if a study wants to compare four ethnic groups, five education levels, three genders and five age groups, then $G = 4 + 5 + 3 + 5 = 17$. Methods for comparisons of different demographic groups on their FNMR's generally involve comparing FNMR's across three or more categories. These methods are more advanced and more sophisticated than those for comparing two groups or for comparing a single group to a specific value. See [13] for methods involving one or two FNMR's. Below we present and discuss statistical methods for determining if there are detectable differences between FNMR's among G independent groups. This single methodology is preferable to testing multiple times which yields potentially higher rates of Type I errors. Below we begin with a bootstrap hypothesis test and that is followed by a simplified version that may be more easily understood by a broad audience.

A. Bootstrap Hypothesis Test

Since the individuals and decisions are independent between groups, we bootstrap each group separately to mirror the variability in the sampling process. As with an analysis of variance (ANOVA), we use a test statistic similar to the usual F-statistic and then we compare the observed value to a reference distribution composed of bootstrapped values. Formally, our hypotheses are: $H_0 : \pi_1 = \pi_2 = \pi_3 = \dots = \pi_G$, vs H_1 : at least one π_g is different.

1) Calculate

$$F = \frac{\left[\sum_{g=1}^G N_\pi^{(g)} (\hat{\pi}_g - \hat{\pi})^2 \right] / (G - 1)}{\left[\sum_{g=1}^G N_\pi^{(g)} \hat{\pi}_g (1 - \hat{\pi}_g) (1 + (m_0^{(g)} - 1)\hat{\rho}_g) \right] / (N - G)} \quad (7)$$

for the observed data where

$$\hat{\pi} = \frac{\sum_{g=1}^G N_\pi^{(g)} \hat{\pi}_g}{\sum_{g=1}^G N_\pi^{(g)}}, \quad \hat{\pi}_g = \frac{\sum_{i=1}^{n_g} \sum_{j=1}^{m_i^{(g)}} D_{ij}^{(g)}}{\sum_{i=1}^{n_g} m_i^{(g)}} \quad (8)$$

and $N = \sum_{g=1}^G N_\pi^{(g)}$. Here $\hat{\pi}$ is the (weighted) average of the FNMR's across the G groups.

2) For each group g , sample n_g individuals with replacement from the n_g individuals in the g^{th} group. Denote these selected individuals by $b_1^{(g)}, b_2^{(g)}, \dots, b_{n_g}^{(g)}$. For each selected individual, $b_i^{(g)}$, in the g^{th} group take all the $m_{b_i^{(g)}}$ non-match decisions for that individual. Call these

¹We are using detectably different here in place of significantly different. See [19].

selected decisions $D_{b_i^{(g)} b_i^{(g)} j}^{(g)}$'s with $j = 1, \dots, m_{b_i^{(g)}}^{(g)}$ and calculate

$$\hat{\pi}_g^b = \frac{\sum_{i=1}^{n_g} \sum_{j=1}^{m_{b_i^{(g)}}^{(g)}} D_{b_i^{(g)} b_i^{(g)} j}^{(g)}}{\sum_{i=1}^{n_g} m_{b_i^{(g)}}^{(g)}} - \hat{\pi}_g + \hat{\pi}. \quad (9)$$

3) Repeat the previous two steps some large number of times, K , each time calculating and storing

$$F_\pi = \frac{\left[\sum_{g=1}^G N_\pi^{(g)} (\hat{\pi}_g^b - \bar{\pi}^b)^2 \right] / (G-1)}{\left[\sum_{g=1}^G N_\pi^{(g)} \hat{\pi}_g^b (1 - \hat{\pi}_g^b) (1 + (m_0^{(g)b} - 1) \hat{\rho}_g^b) \right] / (N-G)}. \quad (10)$$

Here $\bar{\pi}^b$ represents the calculations given above applied to the bootstrapped matching decisions,

$$\bar{\pi}^b = \frac{\sum_{g=1}^G N_\pi^{(g)b} \hat{\pi}_g^b}{\sum_{g=1}^G N_\pi^{(g)b}}, \quad (11)$$

where $N_\pi^{(g)b} = \sum_{i=1}^{n_g} m_{b_i^{(g)}}^{(g)}$. The values for $\hat{\rho}_g^b$ and $m_0^{(g)b}$ are found by using the usual estimates for those quantities applied to the bootstrapped decisions from the g^{th} group.

4) Then the *p-value* for this test is $p = \frac{1 + \sum_{\zeta=1}^K I_{\{F_\pi \geq F\}}}{K+1}$.
5) We will conclude that at least one of the FNMR's is different from the rest if the *p-value* is small. When a significance level is designated, then we will reject the null hypothesis, H_0 , if $p < \alpha$.

We adjust our bootstrapped sample statistic, here $\hat{\pi}^b$, to center their distributions of the FNMR's in each group in accordance with the null hypothesis of equality between the G FNMR's. In this case we center with respect to our estimate of the FNMR, $\hat{\pi}$, assuming all of the FNMR's are identical.

B. Simplified Alternative Methodology for Broad Audience Reporting

The methods of the previous subsection may be difficult to explain to a broad, non-technical audience. Consequently, in this section, we propose a methodology for simplifying the testing of multiple FNMR's across demographic groups. That is, we will conclude that a particular subgroup g has a different FNMR if its observed FNMR is outside of the interval created by taking the average FNMR, $\hat{\pi}$, and adding and subtracting a margin of error, M . This methodology is more straightforward for explaining to decision makers and to wide audiences and takes advantage of the common usage of the margin of error.

In order to generate a single margin of error for all G groups, we bootstrap the differences of each FNMR from the overall FNMR, then use the distribution of the maximal absolute differences to obtain M . This approach is the following:

- 1) Calculate the estimated overall FNMR, $\hat{\pi}$ and the estimated FNMR for each group, $\hat{\pi}_g$, for $g = 1, \dots, G$.
- 2) Sample with replacement the individuals in each group following Step 2) of the bootstrap approach above and

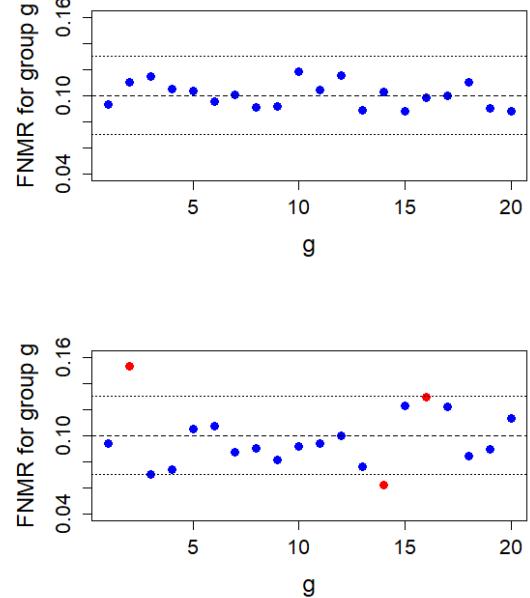


Fig. 1: The top subfigure has all $G = 20$ group FNMR's fall between the bounds (dotted lines) generated by adding and subtracting a margin of error, M , from the overall FNMR (dashed lines), while the bottom subfigure has three subgroups (in red) that fall outside of these bounds.

calculate $\hat{\pi}_g^b$, the scaled bootstrapped estimated FNMM for group g .

- 3) Calculate and store $\phi = \max_g |\hat{\pi}_g^b - \hat{\pi}|$ using the notation from Equation 9 of the bootstrap approach.
- 4) Repeat the previous two steps K times where K is large, say more than 500.
- 5) Determine M by finding the $1 - \alpha/2^{th}$ percentile from the distribution of ϕ .

The maximal differences, the ϕ 's, are calculated from each group FNMR which are scaled to (subtracted from) the overall mean, so the distribution for ϕ that assumes variation if all of the FNMR's are equal. From this approach a range, $(\hat{\pi} - M, \hat{\pi} + M)$, of acceptable variation from the overall estimated FNMR, is generated. The probability that a sample group FNMR would be outside of this interval by chance is $\alpha \times 100\%$ if all the groups are equal. To get a $100(1 - \alpha)\%$ interval, use $\alpha = 0.05$. Thus, if $M = 0.03$ is the 95^{th} percentile of the distribution of ϕ 's and $\hat{\pi} = 0.10$, the probability of a group g having an FNMR be within 3% of the overall FNMR is at least 90%. To illustrate this visually, consider Figure 1 where the top subfigure has no group FNMR's outside our interval while the bottom subfigure has three groups outside the generated bounds of 0.07 and 0.13. Thus, the practical use of this methodology is to produce an easily comprehensible range of values that would not be different from the overall FNMR and, likewise, yield a clearly delineated way to identify those groups with FNMR's that are statistical different from

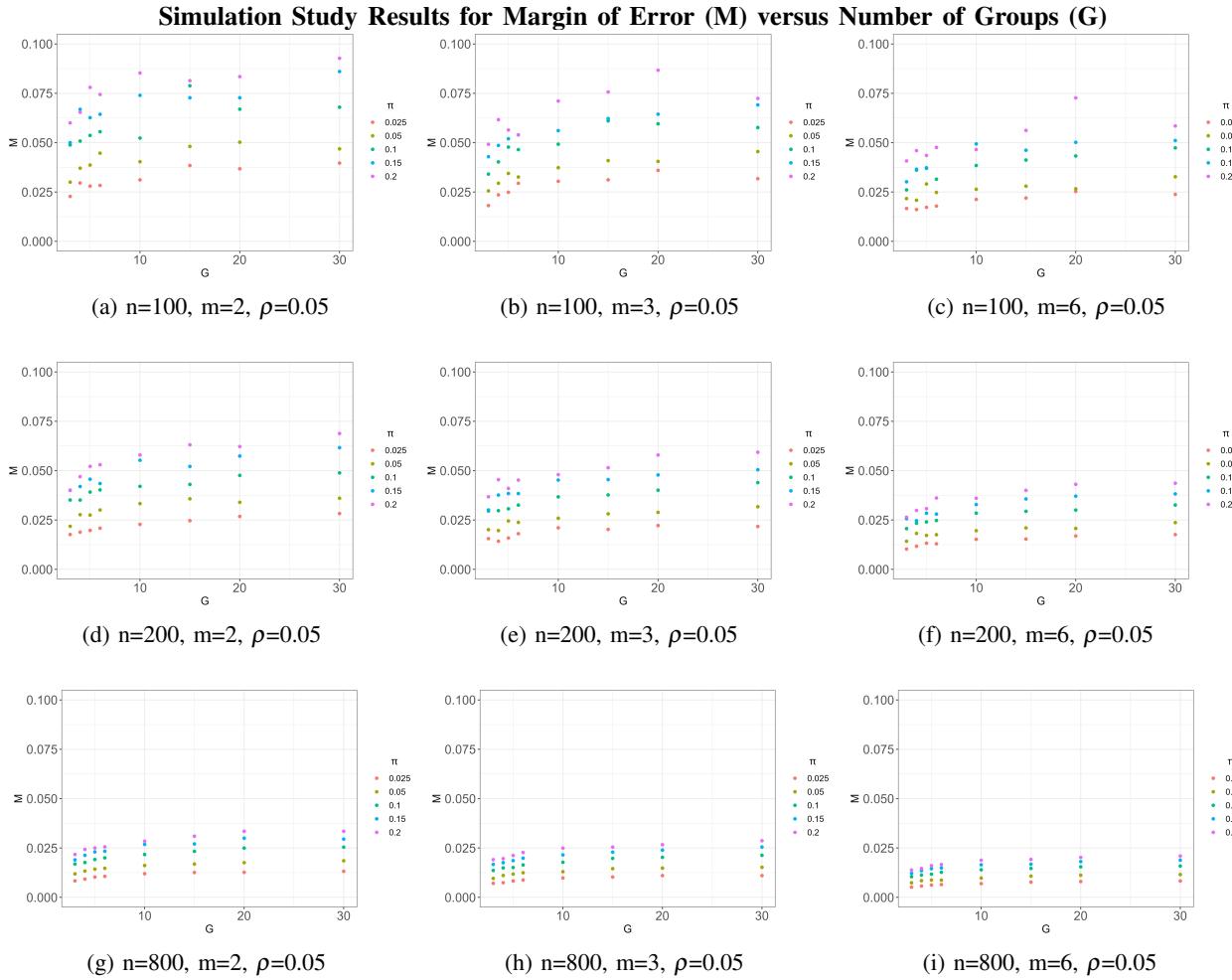


Fig. 2: Results of simulation study for margin of error (M) as a function of number of individuals (n), number of attempts (m), correlation between attempts (ρ), and FNMR (π). Subfigures are organized by columns where m increases from left to right and by rows where n increases from top to bottom. Each figure plots M versus G for fixed $\rho=0.05$ and with different values for π denoted by color.

the overall mean.

Simulation Study

To better understand how M depends upon the parameters of our model, we performed a simulation study. Given G groups, n subjects/group, m attempts per subject, π as the FNMR rate, and ρ as the correlation within subjects, we randomly generated false non match rates, $\hat{\pi}_g$, for each group 1000 times and calculated M as above. We ran all combinations of the following values for each parameter: $\pi = 0.025, 0.05, 0.1, 0.15, 0.20$, $\rho = 0.05, 0.15, 0.25, 0.35, 0.45$, $n = 100, 200, 400, 800$, $m = 2, 3, 4, 6, 10$ and $G = 3, 4, 5, 6, 10, 15, 20, 30$. Our values for ρ were selected to cover the values for estimated intra-individual correlations found in [13]. We fixed α at 0.05 for these simulations. Figures 2 and 3 show summaries of the results of these simulations with M rounded to three decimal places with $\alpha = 0.05$ in all cases. Figure 2 shows simulation results for various values of n , the subfigure rows, m , the subfigure columns and π , colors within each subfigure while

the intra-individual correlation ρ was fixed at 0.05 for these graphics. Within each subfigure, we have plotted M versus G and denoted different values of π by different colors. From each subfigure, we can see that M grows as G increases though the amount of increase in M slows as G gets larger than 10. Moving down subfigure rows, i.e. as n increases we see that M decreases. Similarly, going from left to right across subfigure columns, i.e. as m increases we see decreases in M . Within each subfigure we can see that M becomes smaller as π decreases. In Figure 3, we have plotted M versus G and varied a single parameter (denoted by different colors) in each subfigure at the values given above while fixing the other parameters at $n = 400$, $\pi = 0.10$, $\rho = 0.05$ and $m = 2$. From these values, we can see the impact of n , the number of individuals per group has the largest impact on M , followed by π , the overall FNMR, then m , the number of attempts per individual, and ρ the intra-individual decision correlation. Only ρ is negatively associated with the size of M . The impacts of m and ρ are tied together because of the nature of

Simulation Study Results for Margin of Error (M) versus Number of Groups (G)

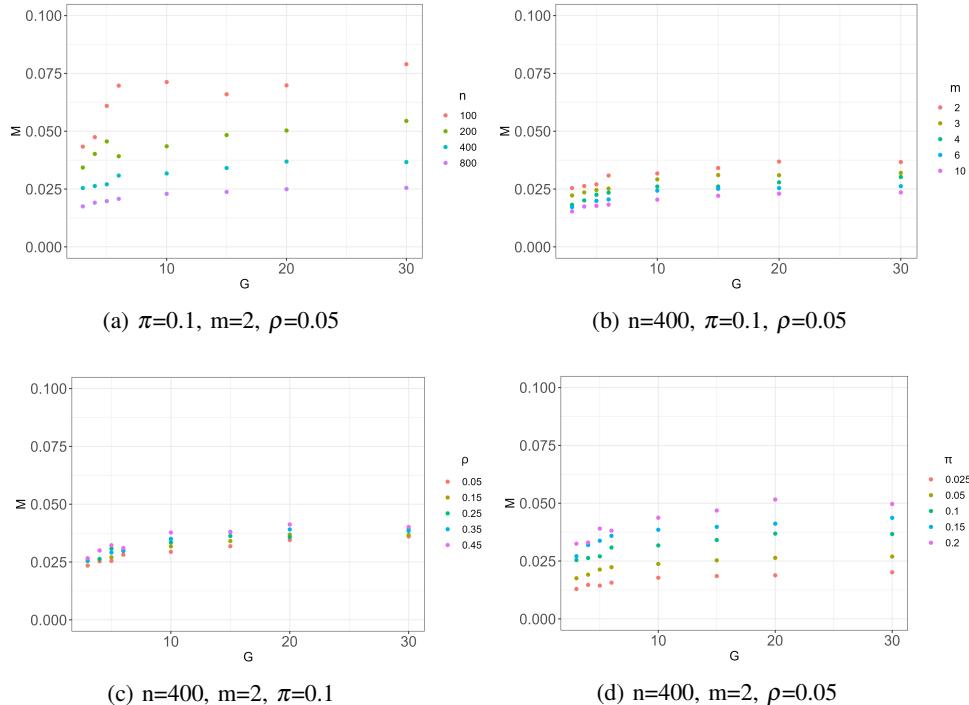


Fig. 3: Within each graph we vary different parameters from our simulation while fixing the others: each color is a varying n (subfigure a); varying m (b); varying ρ (c); and varying π (d)

FNMR data. While not shown in either of these figures, our simulations show that there is a positive, though not linear, relationship between ρ and M .

V. DISCUSSION

Equity and fairness in biometrics are important issues. The declaration of differences between demographic groups is a consequential one. Such conclusions about differences between groups need to be statistically sound and empirically based. In this paper, we have proposed two approaches for testing for statistically detectable differences in FNMR's across G groups. Our first approach uses the F-statistic as a metric and builds a reference distribution for that statistic via bootstrapping. As mentioned above, this methodology, while valid and appropriate, is not easy to explain. Our second approach attempts to remedy this drawback. The second approach is to bootstrap maximal differences among the FNMR's in the G groups assuming a known equal FNMR across all groups, then generates a margin of error, M , to be added and subtracted to the overall FNMR for delineating which groups or subgroups are statistically different from the overall mean. The latter approach has an advantage of being simpler and similar to other colloquial margins of error. From our simulation study of this simpler approach, we have confirmed that the number of groups, G , and the number of individuals tested, N , substantially impact the margin of error. Likewise though to a lesser extent the intra-individual correlation, ρ , and the number of attempts per individual, m , impact the size of M . Our simplified approach uses the maximal absolute

difference from the overall FNMR across the G groups. Using this distribution we generate an interval that is the overall FNMR plus and minus a margin of error M where M is based upon the distribution of the maximal absolute difference. Both of these methods, because they rely solely on thresholded decision data are applicable for testing commercial systems.

Both of our approaches in this paper have considered differences from the overall FNMR, but reasonable alternatives such as $\max_g(\frac{\hat{\pi}_g}{\hat{\pi}}, \frac{\hat{\pi}}{\hat{\pi}_g})$ might be of interest. The importance of being able to generate a reference distribution to allow for an appropriate comparison to the observed statistics is critical to any statistical evaluation regardless of the functional form of the variation.

This paper has looked at false non-match rate but similar methods and approaches exist for other common measures of bioauthentication performance including failure to enrol rates, failure to acquire rates and false match rates. See [13] for approaches for testing and comparing differences among multiple groups for these metrics.

ACKNOWLEDGMENT

This work was supported by grants from the US National Science Foundation CNS-1650503 and CNS-1919554, and the Center for Identification Technology Research (CITeR).

REFERENCES

- [1] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of Machine Learning Research, Conference on Fairness, Accountability, and Transparency*, 2018, pp. 1–15.

[2] J. A. Buolamwini, “Gender shades : intersectional phenotypic and demographic evaluation of face datasets and gender classifiers,” 2017, MSc Thesis; <http://hdl.handle.net/1721.1/114068>; Last accessed: July 10, 2022.

[3] P. Grother, M. Ngan, and K. Hanaoka, “Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects,” United States National Institute of Standards and Technology, Tech. Rep., 2019, NIST.IR 8280, <https://doi.org/10.6028/NIST.IR.8280>.

[4] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, “Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 32–41, 2019.

[5] T. de Freitas Pereira and S. Marcel, “Fairness in Biometrics: A Figure of Merit to Assess Biometric Verification Systems,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 19–29, 2022.

[6] J. J. Howard, E. J. Laird, Y. B. Sirotin, R. E. Rubin, J. L. Tipton, and A. R. Vemury, “Evaluating proposed fairness models for face recognition algorithms,” *arXiv preprint arXiv:2203.05051*, 2022.

[7] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” Tech. Rep., 2016, <https://doi.org/10.48550/arXiv.1610.02413>.

[8] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias,” *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.

[9] IBM, “AI Fairness 360,” IBM toolkit; <https://aif360.mybluemix.net/>, Last accessed: July 11, 2022.

[10] O. A. Osoba, B. Boudreaux, J. Saunders, J. L. Irwin, P. A. Mueller, and S. Cherney, “Algorithmic equity: A framework for social applications,” RAND Corporation, Tech. Rep., 2019.

[11] J. Hsu, *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC, 1996.

[12] K. S. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, “Issues related to face recognition accuracy varying based on race and skin tone,” *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 8–20, 2020.

[13] M. E. Schuckers, *Computational Methods in Biometric Authentication*. Springer, 2010.

[14] ———, “Theoretical statistical correlation for biometric identification performance,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[15] ———, “A parametric correlation framework for the statistical evaluation and estimation of biometric-based classification performance in a single environment,” *IEEE Transactions on Information Forensics and Security*, vol. 4, pp. 231–241, 2009.

[16] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*. John Wiley & Sons, Inc., 2003.

[17] D. A. Williams, “The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity,” *Biometrics*, vol. 31, pp. 949–952, 1975.

[18] M. S. Ridout, C. G. B. Demétrio, and D. Firth, “Estimating intraclass correlation for binary data,” *Biometrics*, vol. 55, pp. 137–148, 1999.

[19] R. L. Wasserstein and N. A. Lazar, “The ASA statement on p-values: Context, process, and purpose,” *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016. [Online]. Available: <https://doi.org/10.1080/00031305.2016.1154108>