N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

Statistical Methods for Testing Equity of False Non Match Rates across Multiple Demographic Groups[1]

Michael Schuckers[2] Kaniz Fatima[3] Sandip Purnapatra[4] Joseph Drahos[5] Daqing Hou[6]
Stephanie Schuckers[7]

Abstract: Biometric recognition is used for a variety of applications including authentication, identity proofing, and border security. One recent focus of research and development has been methods to ensure fairness across demographic groups and metrics to evaluate fairness. However, there has been little work in this area incorporating statistical variation. This is important because differences among groups can be found by chance when no difference is present or may be due to an actual difference in system performance. We extend previous work to consider when individuals are members of one or more demographics (age, gender, race). Our methodology is meant to be more comprehendable by a non-technical audience and uses a robust bootstrap approach for estimation of variation in false non-match rates. After presenting our methodology, we present a simulation study and we apply our approach to MORPH-II data.

Keywords: Fairness, Confidence Intervals, Demographics, Multiple Comparisons

1 Introduction

There has been significant attention to face recognition and artificial intelligence as a whole as it relates to equity. For example, the U.S. Federal Trade Commission released guidance on AI fairness, highlighting that "[i]t's essential to test your algorithm [for discrimination] based on race, gender, or other protected classes" [Ji21]. In a review of face recognition literature, demographic factors may have a significant influence on the performance of some biometric recognition algorithms, resulting in a lower biometric performance for demographic groups, such as females, dark-skinned, and/or youngest subjects [Dr20]. Research has shown that results differ depending on the specific algorithms, capture conditions, use cases, and a host of additional factors [HSV19, GZ19, Go21, We22, Yu22, CKG23] .

This paper develops statistical methods for determining if there are statistically distinguishable false non-match rates (FNMR's) simultaneously across multiple demographics each having more than one category. These methods are aimed at non-technical audience, such as policymakers, rather than the complicated analysis of variance and p-value approaches taken for similar circumstances by [Sc10] which can be problematic [WL16]. Building upon the concept of margins of error which are widely known in the public, we derive methods usable for each demographic group or for all demographic groups simultaneously. Specifically, we extend the work of [Sc22] who considered the case where all the

[1] This material is based upon work supported by the Center for Identification Technology Research and the National Science Foundation under Grant 1650503
[2] Math, CS & Stats, St. Lawrence University, Canton, NY, USA, schuckers@stlawu.edu
[3] Computer and Electrical Engineering, Clarkson University, Potsdam, NY, USA, fatimak@clarkson.edu
[4] Computer and Electrical Engineering, Clarkson University, Potsdam, NY, USA, purnaps@clarkson.edu
[5] Computer and Electrical Engineering, Clarkson University, Potsdam, NY, USA, drahosj@clarkson.edu
[6] Computer and Electrical Engineering, Clarkson University, Potsdam, NY, USA, dhou@clarkson.edu
[7] Computer and Electrical Engineering, Clarkson University, Potsdam, NY, USA, sschucke@clarkson.edu

demographic categories were non-overlapping. demographic groups. Here, we consider the case where individuals are members of multiple groups or categories in several demographics and we will refer to demographics as different dimensions while categories will be the values that each of those demographics can take. For example, our methods apply for simultaneously comparing individuals across racial, educational, and age demographics where each individual is classified into one category within each of those groups. In that instance for the demographic Age, an individual might be in the '25 to 40' category. Additionally, for practitioner flexibility, we present methods for both creating a single margin of error for all demographic groups or simultaneously creating intervals for each demographic separately. For the purposes of this paper, we think of fairness as meaning that the FNMR's are not statistically different across one or more demographic categories and we are motivated by an access application.

2 Related Work

Metrics for the assessment of fairness have been proposed in the literature. [dFPM22] 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. 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) [Ho22]. NIST scientists also propose the Inequality Rate (IR) metric [Gr21]. In addition, the ISO/IEC working draft 19795-10 [IS23] proposes several metrics for demographic performance differentials, including the error rate ratio in case of two groups, and the worst case error rate relative to the geometric mean in case of three or more groups.

While there are several metrics of fairness, there has been little research or use of statistical methods for fairness metrics. The United States National Institute for Standards and Technology (NIST) has performed the most extensive evaluation of biometric recognition as part of a technology evaluation [GNH19]. Results are continually updated at [NI]. Commercial software biometric algorithms are submitted to NIST for testing. Evaluation is performed across a variety of datasets including border, visa application, and mugshot images and for both identification (1:N) and verification (1:1). Performance is reported in terms of FNMR and FMR for verification and FNIR and FPIR for identification. Bootstrapping is provided as a measure of variability and presented throughout their analysis enabling the reader to assess differences, if any, in the context of its statistical variability. Some of the earliest work on the impact of demographics on biometric matching performance was done by [Gi04, Be08, Be09]. More recently, [Co19] look at the impact of demographics on facial recognition.

Bhatt et al. [Bh23] documented and explained the causal understanding of the gender gap problem in the popular deep learning-based facial recognition techniques. The authors claimed the gender gap problem is caused by the imbalance of the test dataset rather than the training set and sorting the images based on hairstyle can reduce the gender gap margin significantly. Other research has also performed extensive evaluations of face recognition

across demographic groups, e.g. [Zh17, Co19, Bu17, GNH19, Kr20, Gr21, Pa22, Te20, Yu22], but have not presented statistical fairness evaluation methods as part of their work.

A definitive methodology for statistical hypothesis testing of the equality of biometric error rates was given in [Sc10]. That approach used resampling methodology to create analysis of variance-like tests for comparing FNMR rates across groups equivalent to a single demographic here. As mentioned above, [Sc22] derived a statistical margin of error via bootstrapping for determining which, if any, FNMR's were different from the rest. However, that paper did not address the practical case when testing across multiple groups simultaneously. In this paper, we generalize their approach to handle the more general and more realistic case when individuals are classified into categories in one or more demographics. One obvious application of this work is the determination of fairness or statistically equal false non-match rates across demographic categories.

3 Methodology

The methods proposed here are motivated by an application where biometric devices are tested across multiple demographics and where each individual is classified into categories separately within each demographic. The aim here is to determine if any of the FNMR's from the categories within demographics are statistically different from the overall FNMR assuming a fixed decision threshold for all categories. Below we will provide methods for that determination within a single demographic or across all of the demographics. The techniques here are useful for assessing the equity of performance across demographics.

Our flexible approach is to bootstrap individuals across groups to obtain an understanding of the variation of the error rates in each category and use that variation to build a distribution of the maximal variation for the overall error rate. For our resampling, we follow the bootstrap methodology for FNMR of [Sc10]. Having obtained a reference distribution of the maximal variation, we then create intervals to determine if there are groups that are statistically different. It is important to note that this approach requires no distributional assumptions about the data. Here we present methods for both additive intervals and multiplicative intervals.

Denote the number of demographics by D and let G_d be the number of categories within each demographic d where $d = 1,\ldots,D$ and $k = 1,\ldots,G_d$. Let π represent a population FNMR and $\hat{\pi}$ represent the estimated FNMR from our sample. The estimated FNMR for category k within demographic d will be denoted by $\hat{\pi}_{dk}$. This is calculated by the total number of false non-matches divided by the total number of attempts of individuals in that category. The number of false non-matches for individual i will be denoted by y_i for $i = 1,2,\ldots,n$. We allow for a different number of attempts per individual which we denote by m_i for individual i. For a multiplicative interval, our equation for the weighted geometric mean FNMR is $\dot{\pi} = (\prod_d \prod_k \hat{\pi}_{dk}^{n_{dk}})^{1/(\sum_d \sum_k n_{dk})}$ where n_{dk} is the number of individuals in category k of demographic d.

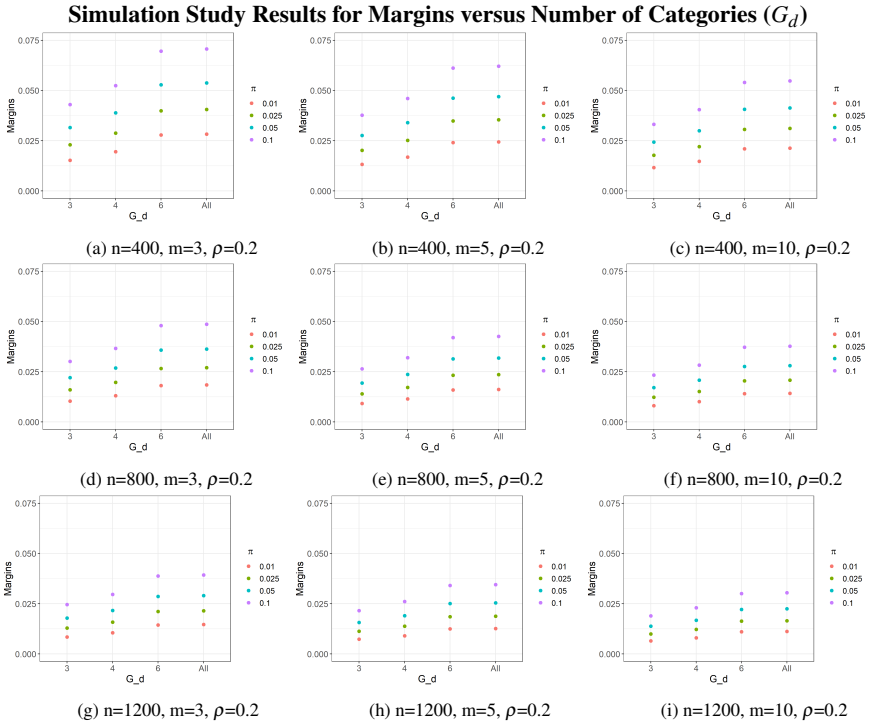
Here we propose two types of inferential intervals: additive and multiplicative. Additive intervals are the most commonly used in practice and involve an estimate plus or minus some margin of error (M). Multiplicative intervals are less common but involve ratios

and an estimate multiplied and divided by a ratio of error (R). We incorporate the latter approach since [IS23] is considering using ratios and geometric means for evaluating the fairness of a biometric device.

Below we present four different approaches to assessing fairness: an additive approach for comparing the FNMR's for all categories with a single interval, an additive approach for comparing FNMR's with each demographic separately, a multiplicative approach for comparing the FNMR's for all categories with a single interval and a multiplicative approach for comparing FNMR's with each demographic separately. The following are the steps for our algorithm.

1. Calculate the error rate, $\hat{\pi}$ and the error rate in each category k within demographic d, $\hat{\pi}_{dk}$. Likewise, calculate the weighted geometric mean for the entire test, $\dot{\pi}$, across the various categories k and demographics d.
2. Sample with replacement the n individuals. For the analysis below, carry along the corresponding demographic information (to which categories they belong) and the corresponding matching performance information (how many errors from how many attempts) for the selected individuals.
3. Calculate the bootstrapped category error rates. Denote them as $\hat{\pi}^b_{dk}$ for each category k in each demographic d.
4. Next calculate and store $\phi = \max_{dk} |\hat{\pi}^b_{dk} - \hat{\pi}_{dk}|$, $\phi_d = \max_k |\hat{\pi}^b_{dk} - \hat{\pi}_{dk}|$, $\psi = \max_{dk} \left(\hat{\pi}^b_{dk}/\hat{\pi}_{dk}, \hat{\pi}_{dk}/\hat{\pi}^b_{dk}\right)$, or $\psi_d = \max_k \left(\hat{\pi}^b_{dk}/\hat{\pi}_{dk}, \hat{\pi}_{dk}/\hat{\pi}^b_{dk}\right)$.
5. Repeat the previous three steps some large number of times, say B times.
6. Let M be the $1-\alpha/2^{th}$ percentile of the distribution of ϕ, let M_d be the $1-\alpha/2^{th}$ percentile of the distribution of ϕ_d, let R be the $1-\alpha/2^{th}$ percentile of the distribution of ψ, and let R_d be the $1-\alpha/2^{th}$ percentile of the distribution of ψ_d.
7. Having obtained values for M, M_d, R or R_d we can create additive intervals for each π_{dk} using $\hat{\pi} \pm M$ and $\hat{\pi}_d \pm M_d$, respectively, as well as multiplicative intervals for π and each π_{dk} using $\dot{\pi}R^{\pm 1}$ and $\dot{\pi}R_d^{\pm 1}$, respectively.

From the intervals derived in the last step of the above algorithm, we can use them to determine which, if any, groups have error rates that differ from the rest. Outstanding category FNMR's, $\hat{\pi}_{dk}$'s, will lie outside of the intervals calculated via the algorithm above. One use for this approach is to look at the equity of FNMR's across all of the demographics and determining which categories have FNMR's outside of the obtained intervals. *A priori* practitioners should decide if they are interested in differences across all demographic groups (using M or R) or in differences within each demographic group (using M_d or R_d for each d). Only one approach should be used since it is possible that there may be differences in the outcomes between the approaches and using multiple approaches induces issues with the familywise confidence level of the interval.



(a) n=400, m=3, ρ=0.2 (b) n=400, m=5, ρ=0.2 (c) n=400, m=10, ρ=0.2

(d) n=800, m=3, ρ=0.2 (e) n=800, m=5, ρ=0.2 (f) n=800, m=10, ρ=0.2

(g) n=1200, m=3, ρ=0.2 (h) n=1200, m=5, ρ=0.2 (i) n=1200, m=10, ρ=0.2

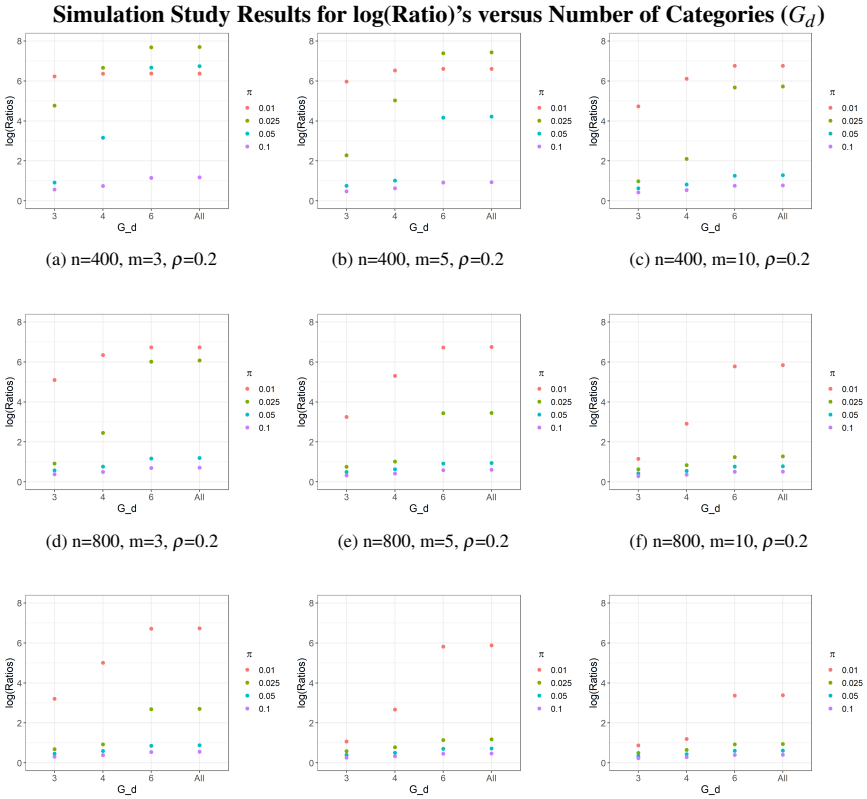
Fig. 1: Results of a simulation study for margins of error as a function of the number of individuals (n), number of attempts (m), the correlation between attempts (ρ), and FNMR (π). Subfigures are organized by columns where m increases from left to right and by rows where n increases from top to bottom. Each figure plots M versus G for fixed ρ=0.2 and with different values for π denoted by color.

4 Simulation Study

To explicate our methodology, we present a simulation study to understand how these performances will differ for different size demographic groups, for different overall error rates and for sample sizes. For a combination of parameters, we generated average values of M, M_d, R, and R_d in order to understand the impact of changes to the parameters on those quantities.

We have the following steps to our simulations having set values for the number of demographics (D), the number of categories in demographic d (G_d), the False Non-Match Rate (π), the intra-individual correlation (ρ), the number of individuals (n), and the number of attempts per individual (m).

1. Generate m attempts from n individuals with an FNMR of π and an intra-individual correlation of ρ.



(a) n=400, m=3, ρ=0.2 (b) n=400, m=5, ρ=0.2 (c) n=400, m=10, ρ=0.2

(d) n=800, m=3, ρ=0.2 (e) n=800, m=5, ρ=0.2 (f) n=800, m=10, ρ=0.2

(g) n=1200, m=3, ρ=0.2 (h) n=1200, m=5, ρ=0.2 (i) n=1200, m=10, ρ=0.2

Fig. 2: Results of simulation study for Ratios 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 natural logarithm of R_d and R versus G_d for fixed ρ=0.2 and with different values for π denoted by color.

2. For each individual $i, i = 1,2,\ldots,n$, and each demographic d, $d = 1,2,\ldots,D$, randomly select a category in $\{1,2,\ldots,G_d\}$ for demographic d.
3. Bootstrap individuals and their corresponding performance/matching measurements and their demographic categories using the algorithm given in the previous section.
4. Find and store M, M_d for each d, R, and R_d for each d.
5. Repeat the previous four steps some larger number of times, say $Z = 1000$.
6. Calculate the mean value for M, M_d, R and R_d.

For our simulation study we used $D = 3$ demographics and $G_1 = 3$, $G_2 = 4$ and $G_3 = 6$. We ran all combinations of the following values for each of the these parameters: $\pi = 0.01$,

Number of Categories, G_d	**Total Subjects,** n	**Percentiles** 80%	90%	95%	97.5%
3	400	0.0128	0.0153	0.0177	0.0200
4	400	0.0160	0.0191	0.0221	0.0250
6	400	0.0221	0.0262	0.0303	0.0345
All	400	0.0228	0.0268	0.0309	0.0350
3	800	0.0090	0.0107	0.0123	0.0137
4	800	0.0112	0.0132	0.0151	0.0169
6	800	0.0153	0.0179	0.0204	0.0229
All	800	0.0158	0.0183	0.0207	0.0232
3	1200	0.0073	0.0087	0.0099	0.0111
4	1200	0.0091	0.0107	0.0121	0.0135
6	1200	0.0123	0.0143	0.0162	0.0180
All	1200	0.0127	0.0146	0.0164	0.0182

Tab. 1: Percentiles from the distribution of M_d's and M's with parameters ρ=0.2, m=10 and π=0.025

Number of Categories, G_d	**Total Subjects,** n	**Percentiles** 80%	90%	95%	97.5%
3	400	2.00	2.30	2.66	3.08
4	400	2.59	3.21	8.20	15.22
6	400	42.06	128.27	290.14	536.41
All	400	42.46	130.29	307.77	557.58
3	800	1.60	1.74	1.88	2.03
4	800	1.86	2.07	2.30	2.54
6	800	2.48	2.93	3.47	4.11
All	800	2.58	3.04	3.58	4.23
3	1200	1.45	1.55	1.64	1.73
4	1200	1.63	1.76	1.90	2.04
6	1200	2.02	2.26	2.52	2.80
All	1200	2.07	2.31	2.57	2.85

Tab. 2: Percentiles from the distribution of R_d's and R's with parameters ρ=0.2, m=10 and π=0.025

0.025, 0.05, 0.10, $\rho = 0$,0.05, 0.1, 0.2, $n = 400$, 800, 1200, and $m = 1,3,5,8,10$. To generate to which category of demographic d an individual belonged, we used equal probability though the methodology could easily be extended to consider non-equal probabilities. We generated $Z = 1000$ datasets for each combination of parameters to ensure that our results were statistically robust. Note that the average number of match decisions or attempts per category was nm/G_d and, thus, the average number of errors was $nm\pi/G_d$. Thus, the number of observations and the number of errors per category decreased as G_d increased. In these simulations, if the number of errors in a given category was zero, we used a small value, $\varepsilon = 1.5/n_{dk}$, the midpoint of a Rule of 30 interval [JL97], to ensure a well-defined values for the ratio.

Tab. 3: FNMR Statistical Summaries for MORPH-II Analysis

	Race		Gender		Age		
	Black	White	Female	Male	17-30	31-45	45+
$\sum_i m_i$	41964	9885	7927	43922	23837	18781	9231
n_{dk}	10561	2599	2074	11086	6163	4657	2340
$\hat{\pi}_{dk}$	0.0241	0.0530	0.0566	0.0247	0.0347	0.0258	0.0242

4.1 Results for M and M_d

We start by considering results for M and M_d from our simulation study described above. Figure 1 shows the 95^{th} percentiles of the average error margin across sets of parameters. There the x-axis of the subfigures is the number of categories, G_d, except for the last category on the right which is labeled as 'All.' This category represents the values for ϕ which is based upon the maximal absolute value of the differences across all categories in all demographics. For the first the values on the x-axis in each subfigure, the quantity plotted is M_d. From each subfigure, we can see that the margin of error grows as G_d increases. Moving down subfigure rows, i.e. as n increases we see that M and M_d decrease. Similarly, going from left to right across subfigure columns, i.e. as m increases we see decreases in the margins of error. Within each subfigure, we can see that M becomes smaller as π decreases. Similar results with specific values can be found in Table 1 which give specific values for the percentiles of M_d's and M's for value of n, when $\rho = 0.2$, $m = 10$ and $\pi = 0.025$.

4.2 Results for R and R_d

Next, we discuss the results of our simulation study for the distributions of ratios that were generated. Figure 2 has the 95^{th} percentiles of the parameter combinations for R_d and R from our simulation study. Because of the large range of values, the y-axis is on a natural logarithmic scale. As we did above in Figure 1, Figure 2 varies n along the subfigure columns and varies m along the subfigure rows. This highlights one of the results of our simulation study which is the ratios generated by our simulation study were sometimes quite large. This was particularly the case when the expected number of errors per number of categories, $nm\pi/G_d$, was small. As above, as either n or m increased these average ratios generally decreased. Increases in π tended to result in decreased values for R_d and R. This pattern, increases in π, differs from the trend for additive intervals and is likely a function of the instability of ratios of small values of π. Table 2 presents the average percentiles for R_d's and R's for three values of n, when $\rho = 0.2$, $m = 10$ and $\pi = 0.025$. Here the same pattern of results as in Figure 2 and the impact of small errors on these ratios is clear as G_d increases when $n = 400$.

5 Illustration using MORPH-II Data

In this section, we apply our methodology to data from the MORPH-II dataset. The MORPH-II dataset is a longitudinal dataset consisting of mugshots images selected from repeat

offenders, taken over the course of 5 years. For our analysis of the MORPH-II mugshot dataset, we used a Resnet50 face recognition model pre-trained on the VGGFace2 dataset from an open-source code repository [He15, Ca18]. Using this model, we extracted the 512-dimensional embeddings from each sample within the dataset. Then we performed comparisons within each individual and computed FNMR. The comparison score was computed using the cosine similarity between two sample embeddings. We computed every permutation of genuine comparisons for each individual. For this analysis because of sample size considerations, we considered (D=3) three demographics: race, gender, and age. Race had two categories (black and white), gender had two categories (female and male), and age had three categories (young adults [17-30], middle-aged adults [30-45], and old-aged adults [45+]). The data analyzed for this project are from 13160 individuals resulting 51844 intra-individual comparisons. Table 3 has the summary for all categories across the various demographics. The total number of attempts per category, $\sum_i m_i$, is given by the first row. The second and third rows have the number of individuals, n_{dk}, and the FNMR, $\hat{\pi}_{dk}$, for demographic d and category k, respectively.

For this application, we set the False Match Rate to 0.10 and had an overall FNMR of $\hat{\pi} = 0.0296$ for all individuals and a weighted geometric mean of $\dot{\pi} = 0.0285$. As expected there is variation between the categories in the FNMR's. We applied our methods above to determine if those differences were statistically discernible. For this bootstrap, we did 5000 replications of the data and results for 95^{th} and 97.5^{th} percentiles can be found in Table 4.

If we want to have a single additive interval for all categories, we should start with the first row, M, and an 95% confidence rate would give an range of $\hat{\pi} \pm M = 0.0296 \pm 0.0094 = (0.0202,\ 0.0390)$. From this we would conclude that any category that fall outside this interval would be statistically different from the overall FNMR. In this case, that would mean that the FNMR's for Whites and Females were statistically larger than the FNMR for all groups. Likewise if we were using a multiplicative interval for all categories, we would find the appropriate interval by taking $\dot{\pi} \cdot R^{\pm 1} = 0.0285(1.252)^{\pm 1} = (0.0228,\ 0.0357)$. As above, our conclusions would be that the FNMR's for Whites and Females are larger than the overall FNMR.

Tab. 4: Bootstrap Percentiles for FNMR Intervals

		95^{th}	97.5^{th}		95^{th}	97.5^{th}
All	M	0.0082	0.0094	R	1.221	1.252
Race	M_1	0.0068	0.0079	R_1	1.141	1.161
Gender	M_2	0.0078	0.0090	R_2	1.152	1.170
Age	M_3	0.0049	0.0055	R_3	1.121	1.247

It is conceivable that the focus of an analysis will be on one specific demographic rather than across all demographics. In that case, the appropriate tool would be the intervals based upon the appropriate demographic. For example, if for the MORPH-II data we are solely interested in Gender, then we would make an additive interval via $\hat{\pi} \pm M_2$. So that a 90% interval would be $0.0296 \pm 0.0078 = (0.0218,\ 0.0374)$ and we would conclude that Females were discernibly different from average. A similarly constructed 90% multiplicative

interval for Age, $0.0285(1.121)^{\pm 1} = (0.0254, 0.0319)$ would find that individuals aged 17 to 30 had a detectably higher FNMR.

6 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 recognize the presence of sampling variation. In this paper, we have proposed interpretable methods for the determination of statistical differences in FNMR's in categories across multiple demographics based upon bootstrapping biometric match data. The first approach is an additive bootstrap-based one that extends previous work and deals with the dependence on the FNMR when individuals are classified in categories across multiple demographics. The second approach is similar to the first but uses a multiplicative methodology from ratios in order to generate ranges of values that are statistically similar. Both approaches yield intervals based on the sampling variation in the relevant metrics and can be used for the identification of demographic categories with FNMR's that are statistically discernible. Our resampling-based approach is focused on creating a simple interval that can be explained to a broad audience.

For the application here and the simulation study we have described above, each individual appeared in only one category for each demographic. However, the methodology is flexible enough to support the case where an individual is in (or selects) multiple categories within a demographic.

The simulation study illustrated that ratio-based confidence intervals are less stable than additive confidence intervals when the expected number of errors is small. This instability in the ratios is more pronounced as the overall error rate decreases.

As with any statistical intervals, the choice of $1-\alpha$, the confidence level, is important. Here the intervals chosen are derived to define family-wise error rates and control the effects of multiplicity. While we prefer the use of a single interval across all demographics for ease of interpretation, we have provided methods for the creation of demographic specific intervals.

To illustrate the utility of our methodology, we have applied our approach to MORPH-II data. In this application, we note that percentiles for M and R are larger than values for any of the M_d and R_d, respectively. This is expected since the former quantities account for variation across all demographics, rather than across a single set of demographic categories. Additionally, we see that the variation within a demographic depends upon the error rate within each category and upon the group with the smallest number of match decisions. This is because biometric error rates are inherently binary. The application of our bootstrap methodology to the MORPH-II data took less than one minute to complete using the `R` programming language on a standard laptop.

The focus of this paper has been on false non-match rates since we are motivated in fairness in access but it is possible to extend the work here to false match rates though the variance structure of false match rates requires a more complicated bootstrap resampling structure, see Schuckers [Sc10].

References

[Be08] Beveridge, J. Ross; Givens, Geof H.; Phillips, P. Jonathon; Draper, Bruce A.; Lui, Yui Man: Focus on quality, predicting FRVT 2006 performance. In: 2008 8th IEEE International Conference on Automatic Face Gesture Recognition. pp. 1–8, 2008.

[Be09] Beveridge, J. Ross; Givens, Geof H.; Phillips, P. Jonathon; Draper, Bruce A.: Factors that influence algorithm performance in the Face Recognition Grand Challenge. Computer Vision and Image Understanding, 113(6):750–762, 2009.

[Bh23] Bhatta, Aman; Albiero, Vítor; Bowyer, Kevin W; King, Michael C: The Gender Gap in Face Recognition Accuracy Is a Hairy Problem. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 303–312, 2023.

[Bu17] Buolamwini, Joy Adowaa: , 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.

[Ca18] Cao, Qiong; Shen, Li; Xie, Weidi; Parkhi, Omkar M.; Zisserman, Andrew: , VGGFace2: A dataset for recognising faces across pose and age, 2018.

[CKG23] Cheong, Jiaee; Kalkan, Sinan; Gunes, Hatice: Causal Structure Learning of Bias for Fair Affect Recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 340–349, 2023.

[Co19] Cook, Cynthia M.; Howard, John J.; Sirotin, Yevgeniy B.; Tipton, Jerry L.; Vemury, Arun R.: 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, 1(1):32–41, 2019.

[dFPM22] de Freitas Pereira, Tiago; Marcel, Sébastien: Fairness in Biometrics: A Figure of Merit to Assess Biometric Verification Systems. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(1):19–29, 2022.

[Dr20] Drozdowski, Pawel; Rathgeb, Christian; Dantcheva, Antitza; Damer, Naser; Busch, Christoph: Demographic bias in biometrics: A survey on an emerging challenge. IEEE Transactions on Technology and Society, 1(2):89–103, 2020.

[Gi04] Givens, Geof H.; Beveridge, J. Ross; Draper, Bruce A.; Bolme, David: Using a Generalized Linear Mixed Model to Study the Configuration Space of a PCA+LDA Human Face Recognition Algorithm. In: Articulated Motion and Deformable Objects. volume 3179 of Lecture Notes in Computer Science, pp. 1–11, 2004.

[GNH19] Grother, P.; Ngan, M.; Hanaoka, K.: Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Technical report, United States National Institute of Standards and Technology, 2019. NIST.IR 8280, https://doi.org/10.6028/NIST.IR.8280.

[Go21] Gong, Sixue: Face Recognition: Representation, Intrinsic Dimensionality, Capacity, and Demographic Bias. Michigan State University, 2021.

[Gr21] Grother, P: Demographic differentials in face recognition algorithms. Virtual Events Series–Demo-Graphic Fairness in Biometric Systems, 2021.

[GZ19] Guo, Guodong; Zhang, Na: A survey on deep learning based face recognition. Computer vision and image understanding, 189:102805, 2019.

[He15] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: , Deep Residual Learning for Image Recognition, 2015.

[Ho22] Howard, John J; Laird, Eli J; Sirotin, Yevgeniy B; Rubin, Rebecca E; Tipton, Jerry L; Vemury, Arun R: Evaluating Proposed Fairness Models for Face Recognition Algorithms. arXiv preprint arXiv:2203.05051, 2022.

[HSV19] Howard, John J; Sirotin, Yevgeniy B; Vemury, Arun R: The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In: 2019 ieee 10th international conference on biometrics theory, applications and systems (btas). IEEE, pp. 1–8, 2019.

[IS23] ISO/IEC 19795-10: , Information technology – Biometric performance testing and reporting — Part 10: Quantifying biometric system performance variation across demographic groups (draft), 2023.

[Ji21] Jillson, Elisa: , Aiming for truth, fairness, and equity in your company's use of AI, 2021. https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai; Last accessed: July 7, 2022.

[JL97] Jovanovic, B. D.; Levy, P. S.: A Look at the Rule of Three. The American Statistician, 51(2):137–139, may 1997.

[Kr20] Krishnapriya, K. S.; Albiero, Vítor; Vangara, Kushal; King, Michael C.; Bowyer, Kevin W.: Issues Related to Face Recognition Accuracy Varying Based on Race and Skin Tone. IEEE Transactions on Technology and Society, 1(1):8–20, 2020.

[NI] NIST: , NIST FRVT Demographics. https://pages.nist.gov/frvt/html/frvt_demographics.html. Accessed: 2023-04-13.

[Pa22] Pahl, Jaspar; Rieger, Ines; Möller, Anna; Wittenberg, Thomas; Schmid, Ute: Female, white, 27? bias evaluation on data and algorithms for affect recognition in faces. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 973–987, 2022.

[Sc10] Schuckers, Michael E.: Computational Methods in Biometric Authentication. Springer, 2010.

[Sc22] Schuckers, Michael; Purnapatra, Sandip; Fatima, Kaniz; Hou, Daqing; Schuckers, Stephanie: Statistical Methods for Assessing Differences in False Non-Match Rates Across Demographic Groups. In: Pattern Recognition. ICPR International Workshops and Challenges. IEEE, Montreal, QC, pp. 207–216, 2022.

[Te20] Terhörst, Philipp; Fährmann, Daniel; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Beyond identity: What information is stored in biometric face templates? In: 2020 IEEE international joint conference on biometrics (IJCB). IEEE, pp. 1–10, 2020.

[We22] Wehrli, Samuel; Hertweck, Corinna; Amirian, Mohammadreza; Glüge, Stefan; Stadelmann, Thilo: Bias, awareness, and ignorance in deep-learning-based face recognition. AI and Ethics, 2(3):509–522, 2022.

[WL16] Wasserstein, Ronald L.; Lazar, Nicole A.: The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician, 70(2):129–133, 2016.

[Yu22] Yucer, Seyma; Poyser, Matt; Al Moubayed, Noura; Breckon, Toby P: Does lossy image compression affect racial bias within face recognition? In: 2022 IEEE International Joint Conference on Biometrics (IJCB). IEEE, pp. 1–10, 2022.

[Zh17] Zhang, Xiao; Fang, Zhiyuan; Wen, Yandong; Li, Zhifeng; Qiao, Yu: Range loss for deep face recognition with long-tailed training data. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5409–5418, 2017.