Convenience versus Stratified Sampling for the Statistical Design of Data Collections

Jad Ramadan

Problem Report submitted to the
Eberly College of Arts and Sciences
at West Virginia University
in partial fulfillment of the requirements
for the degree of:

Master of Science
in Statistics

Defense Commitee:

Mark Culp, Ph.D., Commmittee Chair
Bojan Cukic, Ph.D.
E. James Harner, Ph.D.

Department of Statistics

Morgantown, West Virginia
2012

Keywords: Sampling; Stratified Sampling Example; Rank 1 Facial Recognition; Biometric Collections; Minimum Order Statistic

Abstract

Convenience versus Stratified Sampling for the Statistical Design of Data Collections

Jad Ramadan

This project is concerned with collection methods of biometric samples for research. Some issues to consider for such samples are:

1. The number of subjects to include in a sample.
2. How the subjects are chosen.

Typically, an arbitrary number of subjects that meets a project's financial constraints is agreed upon. The recruitment of subjects is based on local advertising and willingness of individuals to participate. In statistics, this approach of recruitment is called convenience sampling, which, consequently, leads to selection bias. If selection bias is present, the accuracy and risks associated with recognition performance estimated in ensuing biometric studies can be compromised. This report addresses the following questions:

1. How can a biometrics researcher use existing "large" data sets to generate stratified samples?
2. When SRS is used in biometric studies, what practical benefits result from minimizing selection bias?
3. Can we offer a cost effective strategy for using SRS sampling in future studies?

These questions will be answered based on the analysis of anonymous participant data using two samples (stratified and convenience) from recent large scale collections at WVU.

Contents

1 Introduction

This report is concerned with the current data collection methods used in biometric research. The main data frame of interest is an FBI Biometric Collection of People (BioCoP) conducted in 2008. Issues that must be considered when collecting any type of sample include (Schaeffer et al., 15):

- Selection of a *representative* and *random* sample.
- The sample size required.

One of the critical steps underlying the progress in biometric research is the collection of data (single or multiple modalities), such as face, iris, fingerprints, etc. Collection usually follows the examination of operational needs leading towards the design of scenarios, and the IRB approval process. Typically, a sample size that meets a project's financial constraints is agreed upon. The recruitment of subjects is based on local advertising and willingness of individuals to participate (Ortega-Garcia, 2010). This sort of approach is known as *convenience sampling*. Estimators that are based off of convenience samples (such as a sample with no random selection involved) are found to be *biased* (Cochran 12). If bias is present, the statistical analysis may give misleading results and conclusions. Hence, the accuracy and risks associated with estimated recognition performance in ensuing biometric studies can be compromised. This report addresses the following questions:

1. How can a researcher use existing large data sets to generate stratified samples?
2. When stratified sampling is used, what practical benefits result from minimizing selection bias?
3. Can a cost effective strategy using stratified sampling be used for future large scale collections?

2 Literature Review

2.1 Stratified Sampling and Allocation

In stratified sampling, the population is partitioned into a selected number, denoted L, of strata (classes). A sample is then selected by taking a simple random sample within each stratum. Selections in different strata are made independently, which means that the variance estimator for each individual stratum can be added together to obtain a variance estimator for the whole population. Since only the within-stratum variances factor into the variance of the estimator, the *principle of stratification* calls for the partitioning of the population in such a way so that the units within a stratum are as similar as possible (Thompson 101). Even if one stratum differs markedly from another, a stratified sample with the appropriate number of randomly selected units from each stratum will tend to be a representative sample of the entire population (Thompson 101).

Use of random stratified sampling dates back to 1934, when Jerzy Neyman advised the Polish Institute for Social Problems to employ this method in a study of the structure of Polish workers. It was already known that the method of random sampling provided a consistent estimate of an average, denoted X, regardless of the properties of a population (Neyman, 1934, p.586). However, there were, at times, difficulties in defining the "generally representative sample". A *representative method of sampling* was defined by Neyman to be one that "allows us to ascribe to every possible sample, denoted Σ, a confidence interval $X_1(\Sigma)$, $X_2(\Sigma)$ such that the frequency of errors in the statements

$$X_1(\Sigma) \leq X \leq X_2(\Sigma)$$

does not exceed a limit $1 - \epsilon$ prescribed in advance, *whatever the unknown properties of the population*" (Neyman, 1934, p. 585). The method of stratified random sampling is thus classified as a *representative method of sampling* (Neyman, 1934, p. 586).

Once the total sample size n is calculated, one must choose how to allocate it among the L strata. With the BioCoP data, the cost of obtaining an observation is equal, regardless of the observation's respective stratum. Since this is the case, the method of *Neyman allocation* will be used to allocate n to the L strata. Neyman allocation incorporates each stratum standard deviation (approximated by $s_1, s_2, \ldots, s_L$), as well as each stratum population size ($N_1, N_2, \ldots, N_L$), to determine the stratum allocation fractions ($a_1, a_2, \ldots, a_L$), i.e. (Schaeffer et al. 128)

$$a_i = \frac{N_i \sigma_i}{\sum_{k=1}^{L} N_k \sigma_k} \tag{1}$$

If stratum variance is unknown or cannot be estimated, proportional allocation can be used as an alternative. However, it was found almost invariably that the precision of the estimate is increased if Neyman's method is adopted in preference to the method of proportional allocation (Sukhatme 1966, p.371). The gain in precision when using Neyman allocation is even more considerable whenever the variability within each stratum is very different (Sukhatme 1966, p.371). Earlier investigations also showed that stratified random sampling, when properly used, nearly always results in a smaller variance of an estimator than one given by a comparable simple random sample (Sukhatme 1966, p.371).

Improper use of stratified sampling occurs when the sample allocation among the strata is careless or not optimized. In this case, the stratified sampling estimator may have a higher variance when compared to the simple random sample estimator. Another issue with stratified sampling is that bias in the estimator will be introduced if the strata sizes are not accurately known. An increase in sample size will decrease the variance component of the mean square error while the bias component remains unchanged. This bias term will eventually dominate the mean square error and cause the stratified sampling estimator to become less accurate than the simple random sample estimator (Sukhatme 1966, p.372).

2.2 Facial Recognition

A gallery $\mathcal{G}$ consists of a set of biometric samples, g_i, with one biometric sample per person (Li & Jain 553). Since our interest in the BioCoP dataset lies in the question "is the top match correct?," image matching for this data will be considered as an *open-set identification* task. In the open-set identification task, a system determines if a probe p_j corresponds to a person in a gallery $\mathcal{G}$. If the probe is determined to be in the gallery, then the algorithm identifies the person believed to be associated with the probe (Li & Jain 553). When a probe is presented to the system, it is assigned a similarity score, s_{ij}, with each g_i. Similarity between any pair of facial images can be calculated by finding the Euclidean distance, $(g_i - p_j)^2$, between their corresponding feature vectors. This distance measure is used to obtain a simple similarity score through the following equation:

$$s_{ij} = \frac{1}{1 + (g_i - p_j)^2}$$

Larger similarity scores indicate that the two biometric samples in comparison are more similar (Li & Jain 553). A similarity score is a *match* score if g_i and p_j are biometric samples of the same people. Now, let g^* denote the unique match of p_j, with similarity score s_{*j}. A probe p_j has rank n if s_{*j} is the n^{th} largest similarity score. Thus, rank 1 is sometimes called the top match (Li & Jain 553).

It has been found that the risks to experimental validity do not necessarily decrease as collection size increases (Bourlai, 2006). The risk may actually increase with sample size, as observed on several databases (XM2VTS, FRGC) (Bourlai, 2006). This is likely due to convenience sampling, but has not been studied in literature. Stratified random sampling, if used properly, is a technique designed to reduce selection bias, which offers to improve the validity of a study's conclusions. If this technique is performed properly, the experimenter can estimate the sample size necessary prior to sampling in order to possibly reduce some of the cost of collection.

3 Methodology

3.1 Necessary Distributional Derivations

To determine an appropriate sample size for any sort of experiment, the test statistic must first be defined. One case of interest is the BioCoP dataset, on which the Colorado State University Face Identification Evaluation System (Bolme 2003) will used to find similarity scores between 1,130 subjects. The observations, denoted as X_i, are assumed to be normally and independently distributed with mean Z and variance σ^2, i.e. $X_1, X_2, \ldots, X_n \sim N(Z, \sigma^2)$. Setting $Y_i = \frac{X_i - Z}{\sigma}$ normalizes each observation from the target photo, meaning that $Y_i \sim N(0,1)$. From this, $Y_i^2 \sim \chi^2_{(1)}$. It should be noted that Y_i^2 serves the same purpose as the Euclidean distance measure, $(g_i - p_j)^2$, that was previously introduced in section 2.2.

It seems natural to order the values of Y_i^2 from least to greatest. Intuitively, the minimum observation, denoted as $Y^2_{(1)}$, should have the best chance to achieve rank 1 and hence is an appropriate test statistic to use. Through repeated sampling, the next task will be to investigate exactly how small the distance score (denoted as t) needs to be so that the probability of achieving rank 1 is less than or equal to 0.9, i.e. $P(Y_i^2 \leq t) \geq 0.9$. Calculating the probability of a statistic of any sort relies on knowledge of its distribution. For the case of $Y^2_{(1)}$, the theory of order statistics for independent observations must be used (Hogg et al., 240). Thus,

$$\begin{aligned}
P(Y^2_{(1)} \leq t) &= 1 - P(Y^2_{(1)} \geq t) \\
&= 1 - \prod_{i=1}^{n} P(Y_i^2 \geq t) \\
&= 1 - \left[P(Y_i^2 \geq t)\right]^n \\
&= 1 - \left[1 - P(Y_i^2 \leq t)\right]^n \\
&= 1 - \left[1 - F(t)\right]^n \\
&= 1 - \left[1 - \int_{-\infty}^{t} \frac{1}{\sqrt{2}\sqrt{\pi}} x^{-1/2} e^{-x/2} dx\right]^n \qquad (2)
\end{aligned}$$

By definition of a cumulative distribution function, $P(Y_i^2 \leq t) = F(t)$. Before, it was stated that Y_i^2 followed a chi-square distribution with one degree of freedom, which allowed for the substitution in the last line of equation (2).

Solving the equation $1-[1-F(t)]^n \geq 0.9$ for n results in $n \geq \frac{\log(0.1)}{\log(1-F(t))}$. A sample size for the study now only depends on the choice of t. However, incorporation of σ^2 to the bound, by using an estimate from a previous experiment, would perhaps be a more appropriate bound to consider. If the approximation for σ^2 is found to be significantly larger than 1, $\sigma^2 Y_i^2 = (X_i - Z)^2$ (which now serves the exact same purpose as $(g_i - p_j)^2$ from section 2.2) will provide a more accurate estimate of the distance value required in order to achieve rank 1 90% of the time.

Since $Y_i^2 \sim \chi_1^2$, multiplying each Y_i^2 by σ^2 will have (from theory) $\sigma^2 Y_i^2$ following a Gamma distribution, i.e.

$$\sigma^2 Y_i^2 \sim \Gamma(\frac{1}{2}, 2\sigma^2). \tag{3}$$

The cumulative distribution function for such a statistic will have the form

$$\int_{-\infty}^{t} \frac{1}{\sqrt{\pi}(2\sigma^2)^{1/2}} x^{-1/2} e^{-\frac{x}{2\sigma^2}} dx \tag{4}$$

Now, consider the following u substitution:

$u = \frac{x}{\sigma^2}$ $\qquad$ $du = \frac{dx}{\sigma^2}$

$x = \sigma^2 u$ $\qquad$ $dx = \sigma^2 du$

$-\infty < x \leq t \implies -\infty < u \leq \frac{t}{\sigma^2}$

Through this, equation (4) may be re-written as:

$$\begin{aligned} \int_{-\infty}^{t} \frac{1}{\sqrt{\pi}(2\sigma^2)^{1/2}} x^{-1/2} e^{-\frac{x}{2\sigma^2}} dx &= \int_{-\infty}^{\frac{t}{\sigma^2}} \frac{1}{\sqrt{\pi}\sqrt{2\sigma^2}} (\sigma^2 u)^{-1/2} e^{-\frac{\sigma^2 u}{2\sigma^2}} \sigma^2 du \\ &= \int_{-\infty}^{\frac{t}{\sigma^2}} \frac{1}{\sqrt{2 * \pi}} u^{-1/2} e^{-\frac{u}{2}} du \\ &= F\left(\frac{t}{\sigma^2}\right) \end{aligned} \tag{5}$$

$F\left(\frac{t}{\sigma^2}\right)$ is also found to have the form of a chi-square distribution. With this result,

$$P(\sigma^2 Y_{(1)}^2 \leq t) = 1 - \left[1 - F\left(\frac{t}{\sigma^2}\right)\right]^n \tag{6}$$

Setting equation (6) ≥ 0.9, the solution to n when incorporating an estimated prior variance becomes $n \geq \frac{\log(0.1)}{\log\left(1-F\left(\frac{t}{\sigma^2}\right)\right)}$. Values of n at various selections of t and σ^2 are shown in Table 1 below:

t	0.01			0.001			$1 * 10^{-5}$			$1 * 10^{-8}$		
σ^2	1	3	8	1	3	8	1	3	8	1	3	8
n	28	49	81	91	157	257	912	1,580	2,580	28,857	49,984	81,624

Table 1: *Sample size calculations at various values of σ^2 and t*

Probability values such as 0.9999 will also be investigated. When dealing with large datasets that contain millions of observations, a misclassification rate of 1% would result in thousands of misclassifications, which could be a cause for concern. In order to decrease the probability of a type I or type II error for these large datasets, a very small error rate must be stipulated.

3.2 Understanding the Stratum Distributions

In biometrics, it is common to collect databases of images, but these databases are often underused. A stratified sample size estimation approach provides a first step in using this information. For the facial recognition dataset, several features of the observations, including gender, ethnicity, eye color, facial hair, etc., were recorded before photographs were taken. Some features, such as eye color, do not mask any facial features and present no difficulties in facial recognition. Such features will not be considered for stratification since there is no new information to be gained. On the other hand, features such as skin color, whether the subject has a mustache or beard, and make-up (if the subject is wearing some) can cause difficulties

in facial recognition. Stratification of these features may help magnify the ones that really make facial recognition more difficult. The more troublesome features will exhibit a higher variance. Thus, in future collections, stratified sampling will call for more observations from these troublesome strata. This will help the researcher to better understand how to identify, for example, an Asian Indian male with a goatee.

Just because a researcher knows that, say, Middle Eastern individuals give facial recognition techniques many difficulties, does not allow for that researcher to simply sample every Middle Eastern individual available and fill out the experiment with individuals of other ethnicities. Figure 1 below shows the gender and ethnicity distributions of three separate collections:

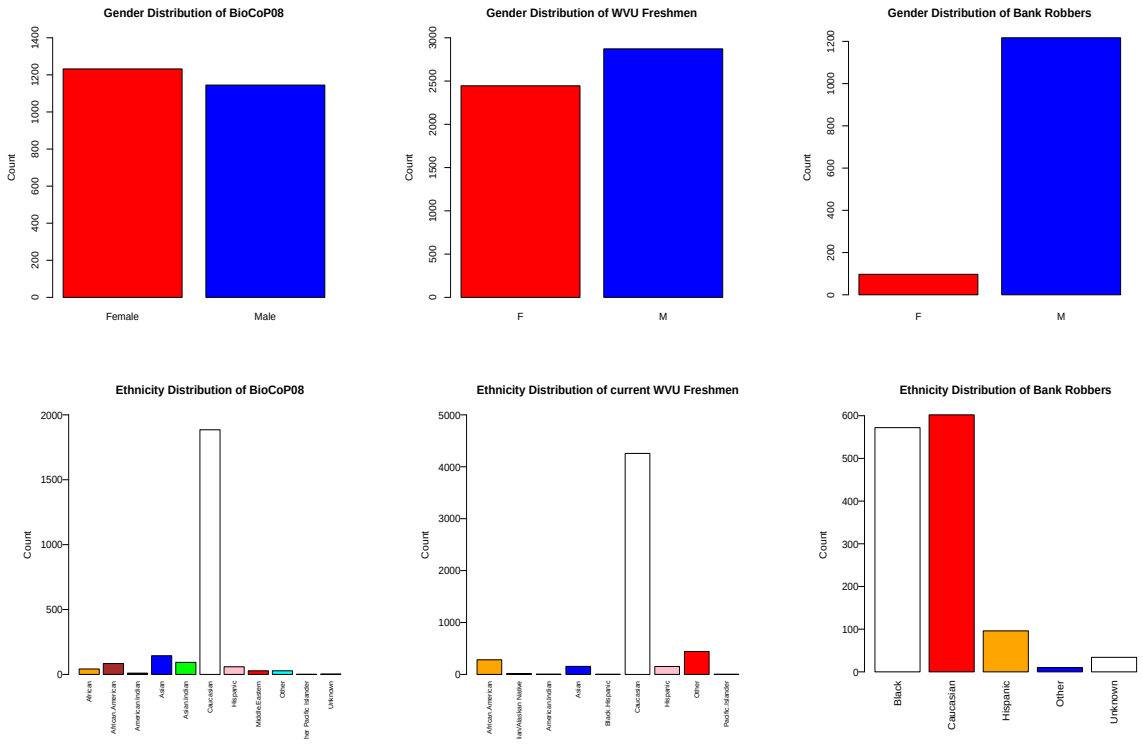

Figure 1: *(left panel) Characteristics of a WVU sample from the BioCoP of 2008. It is dominated by individuals of Caucasian ethnicity. (middle panel) The ethnicity distribution of current WVU freshmen is also heavily dominated by individuals of Caucasian ethnicity. (right panel) A sample of bank robbers from 2011 is shown to be predominantly male and has a more "representative" ethnicity distribution. Clearly, conclusions drawn from this sample should not be applied to the previous two.*

The collections from the left and middle panels of Figure 1 are suited for analysis with images of individuals of Caucasian ethnicity, but not so much for Middle Eastern individuals. With such a dominant Caucasian population at WVU, a random sample taken from this population would also be predominantly Caucasian, which explains why the characteristics of the FBI BioCoP sample from WVU are so similar to the characteristics of current WVU freshmen. An example of convenience sampling would be the case mentioned earlier, where a researcher simply "samples" all of the Middle Eastern individuals as part of the sample. Any conclusions drawn from such a sample will end up as biased and misleading. If future biometric studies depend on these conclusions, they too will be compromised.

Once the main facial recognition dataset of interest is obtained, one investigation will deal with how conclusions change if the underlying stratum distributions are changed. One example (right panel in Figure 1) that will be considered comes from a Bank Crime Statistics collection that was conducted from July 1, 2011 through September 30, 2011 (FBI.gov). Unfortunately, these conclusions can only be of the hypothetical type, as this population is not readily available for study.

4 Results

Since the dataset of interest has not yet been obtained, the focus turns to another dataset, one that was obtained from the official website of the National Basketball Association (NBA.com). The data consists of the height (in inches) and position of 434 NBA players who played at least one minute during a regular season game of the 2011-2012 season. The *principle of stratification*, when applied to a player's position, makes perfect sense for use with this dataset. Guards are similar in height compared to other guards, forwards are similar in height to other forwards, and centers are similar in height to other centers. The mean of these 434 players was found to be 79.04 inches while the variance was calculated to be 12.9. Thus, $X_i \sim N(79, 12.9)$. This estimated variance may be used to determine how

many players, perhaps from some other season, would need to be sampled to get an estimate of the average height from that season. A bound on the error of estimation, of course, must be specified. Figure 2 below compares the necessary sample sizes needed to satisfy a range of bounds for both a simple random sample and a stratified random sample:

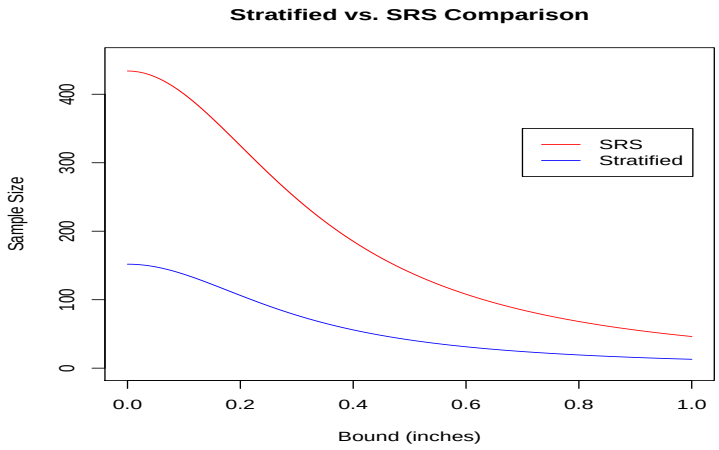

Figure 2: *Sample size comparison to estimate the average height of an NBA player. The dataset was stratified by position, and only bounds ranging from 0 to 1 inch were considered. In this case, stratified sampling is the more efficient method.*

At a bound of 1 inch, a simple random sample calls for 47 observations, while with a stratified sample, Neyman's sample size equation only requires 13 observations. Using *Neyman Allocation*, the sample allocates the 13 observations so that 7 guards, 4 forwards, and 2 centers are selected. Now the question becomes: what value of t (a height) satisfies the equation $P(X_{(1)} \leq t) \geq 0.9$? Here, $X_{(1)}$ is the minimum height generated by the stratified random sample of size 13. Using an earlier derivation from this paper, the cumulative distribution of $X_{(1)}$ with a sample size of 13 is $1 - [1 - \Phi(t)]^{13}$. Setting this equation ≥ 0.9 and solving t gives a value of 75.5 inches, meaning that there is a 90% chance that the minimum height from a stratified sample of size 13 will be less than or equal to 75.5 inches. Two cumulative distribution plots estimated from 50,000 minimums, obtained using both stratified sampling and simple random sampling, are shown in Figure 3.

Usually, an NBA roster consists of 14 or 15 total players. Since the overall sample

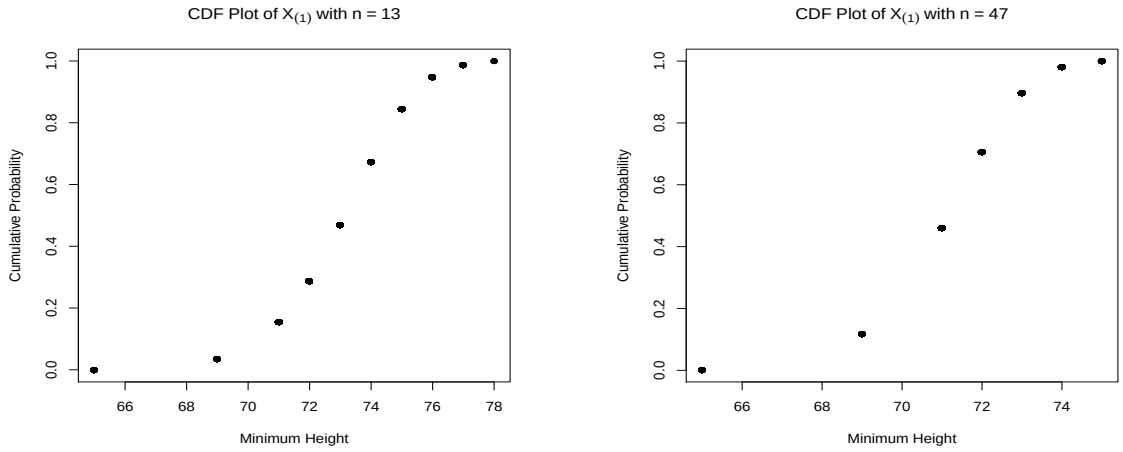

Figure 3: *(left panel) Results from this plot could be used to make conclusions on a typical 15 player NBA roster. (right panel) More samples taken increases the selection probability of a very short player. The minimum height is biased low in this case so these results should not be used for inference on a 15 man roster.*

size used for the cumulative density plot above was 13, one could argue that the shortest player on 90% of the 30 NBA teams will be less than or equal to 75 inches tall. The cumulative distribution function of $X_{(1)}$ changes fairly significantly, however, when a sample of 47 players is taken (as evidenced in Figure 3). Now, the height that satisfies the equation $P(X_{(1)} \leq t) \geq 0.9$ is calculated to be 73 inches. As the sample size increases, the probability of selecting a sample with a minimum height of 75 inches decreases. 47 players is far too many to make meaningful conclusions here, though, as no type of basketball team will have as many as 47 players on the roster. There were 8 teams whose final roster for the 2011-2012 season included a shortest player of 74 inches (NBA.com), which does not satisfy the p-value given by the CDF in the right panel of Figure 3. There were also no final rosters that included a shortest player of 76 inches, so perhaps a sample size of 13 is also too small for this sort of argument. However, a sample size of 13 provides a much more unbiased representation of a team's shortest player when compared to a sample size of 47.

A sample of 15 NBA players could be conveniently selected from one specific team. This also will provide a biased estimate of average NBA player height. The 2011-2012 roster of

the Washington Wizards, for example, had an average height of 79.6 inches, a shortest player of 75 inches, and a tallest player of 83 inches. A height of 75 inches, as seen in the left panel of Figure 3, is not very likely to be associated with the shortest player in a sample of this size. Management style differs from team to team, which means that a certain NBA roster may not necessarily serve as a representative sample of the entire population of NBA players. Because of this, if one team is used to make inferences on the average height of the population of NBA players, bias will be introduced into the estimate as well as the conclusions drawn from this estimate.

5 Conclusions

The NBA heights data answered all 3 questions posed in the introduction. An existing large dataset (that could have been much larger) was used to generate stratified samples by using ideas from the *principle of stratification* (Thompson 101). Players of similar heights were easily grouped together by their position. Stratifying by position came naturally. It was also shown that stratified random sampling provided more unbiased conclusions on the minimum order statistic. When too many players are included in the sample, the minimum height from that sample will not provide a very accurate depiction of the shortest player on an NBA team. Stratified sampling, when compared to simple random sampling, was also shown to greatly reduce the number of samples required to achieve an estimate of the average height of an NBA at every bound value. A stratified sample of size 13 could be taken from *any* NBA season and provide an estimate of the average height of that season with a bound of one inch.

The BioCop dataset is planned to be analyzed using the CSU Face Identification Evaluation System. Statistical analysis will proceed in the same manner as with the NBA heights dataset. First, observations with similar features will be grouped together to form the strata. The distribution of $Y_{(1)}^2 = \min (X_i - Z)^2$ will be investigated similarly to the distribution of

the minimum height statistic. The results will be used to better understand the similarity score and sample size required to satisfy the equation $P(Y^2_{(1)} \leq t) \geq 0.9$ and achieve rank 1. Rank curve generation will be used as it was in Figure 3 of [1]. The only difference will be that rank will be fixed at 1, while various values of sample size n as well as similarity scores s_{ij} will be investigated as independent variables.

Also, a blueprint for future large scale biometrics collections will be provided: one that might reduce the size of the data frame while emphasizing the inclusion of features that give facial recognition algorithms trouble. However, if a sensitive subject such as ethnicity is to be stratified, a researcher must be very careful. Many individuals, especially those in college, may lie about their ethnicity to not feel singled out. In more prestigious institutions, such as the Massachusetts Institute of Technology, many applicants lie about ethnicity in the hope that "minority status" increases their chance of acceptance or their chance to receive funding (College Confidential Forums). Some questions to be thought about when stratifying on a stratum such as ethnicity include:

1. What happens when strata are misleading?

2. Does it matter if an individual's lie is intentional or unintentional?

With the facial recognition dataset, optimal characteristics (ethnicity, facial hair, etc.) to stratify the data frame with will be determined. The two questions from above, however, may prove to be irrelevant if ethnicity is not found to lower the required sample size.

References

[1] Bolme, D., Teixeria, M., Beveridge, J., and Draper, B. (2003). The CSU Face Identification Evaluation System: its Purpose, Features and Structure. *Proceedings of 3rd International Conference on Computer Vision Systems*, pp. 304-311.

[2] Bourlai, T., Kittler, J., and Messer, K. (2006). Database Size Effects on Performance on a Smart Card Face Verification System, *Face and Gesture*.

[3] Cochran, W.G. (1977) *Sampling Techniques* (Third Edition). New York: Wiley.

[4] Hogg, R., McKean, J., Craig, A. (2005). *Introduction to Mathematical Statistics* (Sixth Edition). New Jersey: Pearson Prentice Hall.

[5] Li, S., and Jain, A. (2005). *Handbook of Facial Recognition* (Second Edition). Ney Jersey: Springer-Verlag.

[6] Neyman, J. (1934). On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, **97**, pp. 558-606.

[7] Ortega-Garcia, J., et al. (2010). The Multi-Scenario Multi-Environment BioSecure Multimodal Database (BMDB), *IEEE PAMI*.

[8] Schaeffer, R., Mendenhall III, W., Ott, R., and Gerow, K. (2012). *Elementary Survey Sampling* (Seventh Edition). Boston: Cengage.

[9] Sukhatme, P. (1966). Major Developments in Sampling Theory and Practice. *Research Papers in Statistics*. F.N. David (Ed.). New York: Wiley.

[10] Thompson, S. (1992). *Sampling* (Second Edition). New York: Wiley.