Is the Urn Well-Mixed?
Uncovering False Cofactor Homogeneity Assumptions in Evaluation

Ross J. Micheals
Image Group, Information Access Division, Information Technology Lab
National Institute of Standards and Technology
Gaithersburg, MD, USA
rossm@nist.gov

Terrance E. Boult
Vision and Software Technology Lab
University of Colorado at Colorado Springs
Colorado Springs, CO, USA
tboult@cs.uccs.edu

October 21, 2004

Abstract

Measuring system performance is conceptually straightforward; it is the interpretation of the results and their use as predictors of future performance that are the exceptional challenges in system evaluation, and experimentation in general. Good experimental design is critical in evaluation, but there have been very few techniques that a scientist may use to check their design for either overlooked associations or weak assumptions. For biometric and vision system evaluation, the complexity of the systems make a thorough exploration of the problem space impossible. This lack of verifiability in experimental design is a serious issue. In this paper, we present a new evaluation methodology that aids the researcher in discovering false assumptions about the homogeneity of cofactors — when the data is not "well mixed." The new methodology is then applied in the context of a biometric system evaluation.

Keywords — biometrics; evaluation; variance estimation; performance characterization; sample design

1 Introduction

Conceptually, measuring system performance is straightforward — run the system over some input data, then, combine the output and ground truth reporting some metric. With the data volume computational requirements inherent to vision systems, the mechanics and logistics involved in running a system evaluation is considerable (consider [1] or the forthcoming NIST Fingerprint Vendor Technology Evaluation 2003 [1], for example). Furthermore, a more significant difficulty in performance evaluation is the interpretation of the results. Results from an evaluation are most valuable if they can be used as predictors of performance on compatible data. But what is meant precisely by *performance prediction* and *compatible data*?

From a statistical perspective, system performance prediction is often viewed as a parameter estimation problem. For biometric systems, this might be a system's identification rate, or some point on an ROC (receiver operator characteristic). Traditionally, the performance parameter is estimated by executing an experiment and computing the performance metric. A single measurement alone, however, gives limited insight into the expected range of performance values should the experiment be run again. For this, we need to estimate the variance of the potential results, and ideally, their corresponding likelihoods of occurring. Various vision papers have looked at computing performance with added statistical data or confidence intervals [2–8]. However, implicit in each of these analyses are simplistic sample designs that may not always properly reflect the nature of the data.

The concept of *compatible data* is more nebulous. For vision systems, the mapping from real-world predicates to the data that impacts the system is very complex. If the mapping from these cofactors to the actual data produces an unpredicted clustering, the observations within an experiment may not be compatible. On the subject of homogeneity and the implications of clustered data, Kish [9] wrote

> The correspondence with the "well-mixed urn," inherent in the assumption of independence, is negated; and formulas that depend on that assumption fail to apply.

Therefore, consider the following definition. Given a pair of observations (X, Y), where X is a vector of cofactors and Y is some value(s) of interest, we consider a cofactor X_i to be *sufficiently homogeneous* if and only if across trials of an experiment the X_is are independent and identically distributed, or *iid*. (This includes the trivial case in which X_i is constant). This is compatible with the desire in experimental design that cofactors be either (a) constrained or (b) sufficiently randomized. In either case, the cofactor would be sufficiently homogeneous, and might be analyzed as an (a) fixed or (b) random effect in some statistical model [10]. Homogeneity of all cofactors is not a necessary condition for Y to be *iid*, since it is possible that a cofactor has a nominal effect on the observations. Note that such a model requires explicitly culling out each cofactor X_i.

A proposed model that departs from empirical data calls into question the cofactor homogeneity assumptions proposed by an experimenter. For example, suppose for

[1] `http://fpvte.nist.gov`

a given experiment, the statistic of interest is σ_θ^2, the variance of some performance metric θ across multiple trials. Let $\hat{\sigma}_\theta^2$ represent the point estimator of σ_θ^2 computed from a single trial. Then, if $\bar{\sigma}_\theta^2$, the average point estimator of the sample variance and, $V[\hat{\theta}]$, the sample variance of the parameter of interest do not converge — that is, if $\bar{\sigma}_\theta^2 = V[\hat{\theta}]$ does *not* hold as the number of trials goes to infinity,[2] then this indicates that at least one of the assumptions required for this equality to hold must be false. One of the most common model departures, addressed in this research, is the violation of the *iid* assumption — that is, one or more of the cofactors X_i or observations Y_i are not sufficiently homogeneous.

Statistical literature is rich with research regarding the effects of cofactors that do not meet the experimenter's expectations once they are known [11]. It is the authors' conjecture that there is much less research regarding the identification, selection, and incorporation of influential cofactors when it comes to analysis ([12, 13] for example) because of the intimate domain dependence of any cofactor judgment. While many researchers consider cofactors in their own experiments, it seems there as been little discussion within the performance evaluation literature as a whole (particularly within computer vision and biometrics) with respect to the discovery of influential cofactors that remain unmodeled.

In this paper we present a methodology that can be used to help identify false cofactor homogeneity assumptions. The new methodology is not a panacea, and is no substitute for thoughtful experimental design. It can, however, help guide an evaluator in determining the validity of various cofactor assumptions. Unlike other biometric- or classifier-oriented methods [2–4], a distinguishing characteristic of this research is the recognition that influential cofactors may each have their own sample designs, and that this sample design may have a large influence on the resultant variance and therefore, confidence intervals.

The new methodology has some unique features worth separate consideration:

- **Iterative**. The new methodology is iterative in nature. This allows information gained from one analysis to be used as feedback to the next iteration. In this manner experimenters can adjust assumptions until they fail to reject a proposed sampling design. Given the often large amounts of data/computation needed in vision this iterative feature is well suited to vision system evaluation.

- **Hierarchical RVs**. In its current form, the methodology is best suited, but not limited to, for random variables (RVs) having distributions that are hierarchical in nature. (Hierarchical RVs will be reviewed in the following section).

- **Binary**. The new methodology exploits some unique relationships between survey sampling and binary data. These properties are at the core of the methodology's capability to provide evidence against false cofactor homogeneity assumptions.

The methodology is primarily data-driven, — i.e., the data of interest itself is used to explore the cofactors of interest. Therefore, the methodology is most convenient for

[2] $E[\bar{\hat{\theta}}] = \theta$ by definition since we assumed that $\hat{\theta}$ is unbiased

use in studies where the experiment can be easily repeated. Computational experiments and simulations that allow for the generation of large amounts of data are a particularly good fit to the proposed method.

Fundamentally, it is the goal of this paper to provide a methodology that may be used to not only select an appropriate variance estimator for an experiment, but also provide an estimator that is empirically justified. Ideally, the methodology would provide a way to prove the correctness of the estimator, however, this is far beyond the scope of this paper. Regardless, using an emperically justified variance estimate is a significant improvement from the common practice of selecting a traditional variance estimator (along with its requisite assumptions) out of convenience.

This paper is organized as follows. In Section 2, Background, we provide the necessary information on hierarchical distributions, repeated measures, sample design, infinite populations and their relationships to binary statistics. Next, in Section 3, Methodology Outline, we show the basic steps that are to be iterated until valid set of homogeneity assumptions are found. In Section 4, Experimentation, we show how the methodology can be applied to a biometric system evaluation. Section 6, Conclusions, closes the paper.

2 Background

In this section, we will consider hierarchical distributions of random variables and how they relate to survey sampling statistics. A hybridization of both methods serves as the basis of the new methodology.

The distribution of a random variable is *hierarchical* if that random variable is realized through sampling of a cascade of two or more component distributions, or *stages*. A typical mixture distribution (such as a mixture of Gaussians commonly used in modeling vision systems) is a two-stage hierarchical distribution since each realization requires two separate samplings. In hierarchical models of two levels, the high-level stage or *hyperdistribution* is effectively a distribution of distributions.

Many system evaluations are concerned with random variables having distributions that are hierarchical in nature. For example, consider the distribution of the input to a classifier system. The first stage corresponds to the selection of a particular class and the second stage a particular instance of the selected class. Hierarchical distributions can be problematic with respect to cofactor homogeneity when they are sampled in non-designed ways. Specifically, when data contains repeated measures, there can be groups of data that share an influential cofactor in an unbalanced fashion. From a distributional standpoint, repeated measures can occur whenever one level of the hierarchy is constrained in a manner inconsistent with other observations. For example, in a classifier experiment in which m instances of each of n classes are observed, the classifier label (certainly an important cofactor) is not well-mixed in the *iid* sense.

The statistics of survey sampling are particularly well-suited for repeated measurements [11, 14]. However, traditional survey sampling is primarily concerned with modeling finite populations, or more generally, with observations made by a sampling process that eventually terminates. The finite population correction terms found in survey sampling estimators ensure that if, in each trial,n *all* of the population elements are se-

lected and selection occurs without replacement, then the variance of a statistic across different trials will be zero.

The vast majority of vision and biometric system evaluations, however, are concerned with modeling infinite quantities of input-output pairs. The potential *values* that these pairs may take on may be finite, but the process by which these values are generated does not terminate in the same sense of sampling from finite population without replacement. Likewise, sampling a finite population with replacement is not a process that *must* terminate.

As noted in [15], [16], and [17], traditionally, random variable-oriented research lacks robustness when it comes to describing data with complex sample designs (i.e., non-*iid* data). Specificially, it is not a general rule that survey sampling estimators can be applied as is to infinite populations, without breaking down. For example, consider the finite population stratified sampling estimator from [14](p. 92)

$$[M\,(M-m)]/(n^2 m)\sum_{i}^{N} S_i^2 \tag{1}$$

where M is the total elements per stratum, m is the number of sampled elements per stratum, N is the number of strata and S_i^2 is the element variance of stratum i. Clearly, this estimator breaks down as M approaches infinity. Therefore, what is needed is a collection of unbiased estimators for various sample designs as they correspond to infinite population models. This requires mapping the sample design concept from survey sampling to random variables.

We have not yet considered accommodating *how* repeated measures are generated given a hierarchical distribution. In survey sampling, the *sample design* is the mapping from elements in the population to their respective probabilities of being selected. With random variables of hierarchical distributions, we do not have quite the same concept — observations are not viewed as having an *a priori* determined value. Philosophy aside, we will discuss three survey sampling designs — simple random sampling, stratified sampling, and cluster sampling — with respect to hierarchical distributions where the sampling process at each stage of the distribution does not terminate.

2.1 Variance Estimators

It is common to find discussions within the computer vision community of what is considered to be a *repeatable experiment*. We propse the definition that a repeatable experiment is one that can be modeled statistically and verified empirically. While most experiments have a potentially infinite number of parameters that can be modeled, our focus will be on the minimal modeling needed for prediction — i.e. on estimating variance. The ultimate goal of the method is quite simple — seek a sample design in which the variance estimate computed from a single trial agrees with the average sample variance from repeating experiments. When this is *not* true, the assumed sample design is rejected.

In this section, we derive unbiased estimators of the variance of the mean of group means for simple random sampling (*srs*), stratified sampling (*st*), and cluster sampling (*cl*).

2.1.1 Simple Random Sampling

We begin with the relationship between simple random sampling and *iid* data. In survey sampling, *srs* implies that every population element has the same (uniform) probability of being selected. We can make an analogous construction for random variables. Consider a discrete population of n elements, where each element x_i has a relative frequency f_i. Then, it is trivial to construct an index set over the interval $[0, 1)$ and a mapping F^{-1} from indexes to outcomes such at F^{-1} is the quantile function of F, the CDF of all f_is. Then, a uniform selection over the index set generates *iid* samples of F [18]. A similar construction can be used for the continuous case.

2.1.2 Stratified Sampling

In survey sampling, stratified sampling involves partitioning the population into n non-overlappying groups, or *strata* , and then systematically selecting m elements via *srs* within each stratum. Given the duality of *srs* and *iid* data, it follows that for infinite populations, stratified sampling is tantamount to selecting, across trials, the *same* n hyperdistributions. Notice that having a hyperdistribution that is effectively discrete is a necessary (but not sufficient) condition for stratified sampling. A hyperdistribution might cover an infinite population, but if the sampling process results in a deterministic and repeatable selection of particular mixture components, then stratified sampling results, since the same hyperdistributions (or strata) are selected every trial.

It can be demonstrated that an unbiased estimator of $V[E[\mu_i]]$, the variance of the expected value of the stratum means is $\bar{S_i^2}/nm$. Let X_{ij} be the jth element of stratum i. Then

$$V[\bar{\bar{X}}_{st}] = \frac{1}{n^2m^2}\sum_i^n\sum_j^m V[X_{ij}]. \quad (2)$$

But X_{ij} and $X_{ij'}$ share the same distribution F_i for $j \neq j'$, so

$$V[\bar{\bar{X}}_{st}] = \frac{1}{n^2m^2}\sum_i^n\sum_j^m V[X_{i\cdot}] \quad (3)$$

However, by definition of the stratified sampling process, samples are *iid* within a stratum, so

$$V[\bar{\bar{X}}_{st}] = \frac{1}{n^2m^2}\sum_i^n mV[X_{i\cdot}] \quad (4)$$

$$= \frac{1}{n^2m}\sum_i^n \sigma_i^2 \quad (5)$$

where σ_i^2 is the variance the statistic of interest over stratum i. Therefore, an unbiased estimator of $V[\bar{\bar{X}}_{st}]$ is $\bar{S_i^2}/nm$, since $E[\bar{S_i^2}] = \sigma_i^2/n$. This form is similar, but not quite the same as the survey sampling estimator for finite populations.

2.1.3 Cluster Sampling

In survey sampling, cluster sampling involves two stages of *srs*. First, *srs* is performed such that each sample selects a subpopulation. Second, *srs* is performed over each selected subpopulation. From a random variable standpoint, this is tantamount to first, realizing a distribution — e.g. fixing a hyperdistribution parameter — and second, sampling *iid* data from that constrained distribution. Note the subtle, but important difference from stratified sampling, where the same subpopulations are selected every trial.

We derive the variance of the cluster sample mean for doubly infinite populations by first considering the case in which the mean is computed from m samples of one hyperdistribution sample. This case is easily generalized to n hyperdistributions since, for *iid* Xs, $V[\sum^n X] = nV[X]$.

We make use of the conditional variance formula,

$$V[X] = E[V(X|\theta)] + V[E(X|\theta)] \tag{6}$$

where, in our case, θ corresponds to fixing the hyperdistribution. From Equation (6),

$$\begin{aligned} V[\bar{X}] &= \frac{1}{m^2}\left(V[E(\sum_{\theta} X|\theta)] + E[V(\sum_{\theta} X|\theta)]\right) &(7)\\ &= V[E(X|\theta)] + \frac{1}{m}E[V(X|\theta)] &(8)\end{aligned}$$

Therefore, for n repetitions of the sample process

$$\begin{aligned} V[\bar{\bar{X}}_{cl}] &= \frac{1}{n}V[E(X|\theta)] + \frac{1}{nm}E[V(X|\theta)] &(9)\\ &= \frac{1}{n}V[\mu_i] + \frac{1}{nm}E[\sigma_i^2] &(10)\end{aligned}$$

However, we still need an unbiased estimator of $V[\bar{\bar{X}}]$. We might use a linear combination of $\bar{S}_i^2 = \sum S_i^2/n$ for $E[\sigma_i^2]$ and $S_{\bar{i}}^2 = \sum(\mu_i - \bar{\mu})^2/(n-1)$ for $V[\mu_i]$. Although $E[\bar{S}_i^2] = E[\sigma_i^2]$, $E[S_{\bar{i}}^2] = V[\bar{X}_i]$, suggesting another application of the conditional variance formula. This yields

$$E[S_{\bar{i}}^2] = V[\mu_i] + \frac{1}{m}E[\sigma_i], \tag{11}$$

which implies $S_{\bar{i}}^2/n$ should be used for an unbiased estimator of $V[\bar{\bar{X}}_{cl}]$. Comparing this with Equation (5), we can see that the subtle difference between stratified and cluster sampling leads to significantly different variance estimators.

2.2 Binary Data

In the introduction it was mentioned that a key component of the new methodology stems from reducing the response variable of interest into binary data. In performance evaluation this is common — we produce experiment data that is classified as either

a success or a failure. We may group this into higher level constructs, such as an ROC curve, but the inherent binary nature has important implications on the statistical models we can use. Here, we will examine how our estimators interact with binary data, and present an important inequality used in the main result of the paper.

For a two-stage hierarchical distribution in which the final observations are binary, the hyperdistribution consists of the space of all single-dimensional distribution functions over $[0, 1]$. (For simplicity, we restrict ourselves to only distributions with finite first and second moments.) It follows that the second stage is a Bernoulli distribution with mean θ (probability of success) and variance $\theta(1-\theta)$. Let μ_θ and σ_θ^2 represent the true mean and variance of θ, as determined by the hyperdistribution.

For cluster sampling, we combine the identity

$$\begin{aligned} E[\theta - \theta^2] &= E[\theta] - E[\theta^2] + E^2[\theta] - E^2[\theta] && (12) \\ &= \mu_\theta(1-\mu_\theta) - \sigma_\theta^2 && (13) \end{aligned}$$

with Equation (9), to get

$$\begin{aligned} V[\bar{\bar{X}}_{cl}] &= \frac{V[\theta]}{n} + \frac{E[\theta-\theta^2]}{nm} && (14) \\ &= \frac{\sigma_\theta^2}{n} + \frac{\mu_\theta(1-\mu_\theta) - \sigma_\theta^2}{nm} && (15) \\ &= \frac{(m-1)\sigma_\theta^2}{nm} + \frac{\mu_\theta(1-\mu_\theta)}{nm} && (16) \end{aligned}$$

Observe, that for *srs*, $m = 1$, so $V[\bar{\bar{X}}_{srs}] = \mu_\theta(1-\mu_\theta)/n$. Therefore, $V[\bar{\bar{X}}_{srs}]$ from $n' = nm$ samples is always less than or equal to $V[\bar{\bar{X}}_{cl}]$.

For stratified sampling, each stratum has its own probability of success, θ_i. From Equation (9),

$$V[\bar{\bar{X}}_{st}] = \frac{1}{nm}\sum_i^n \theta_i(1-\theta_i). \quad (17)$$

It can be shown ([19]) that Equation (17) is equivalent to

$$V[\bar{\bar{X}}_{st}] = \frac{\mu_\theta(1-\mu_\theta)}{nm} - \frac{\sum_i^n \theta_i - \mu_\theta^2}{n^m}, \quad (18)$$

which is always less than or equal to $V[\bar{\bar{X}}_{srs}]$ for nm samples.

In conclusion, for two-stage hierarchical distribution of binary outcomes, the following inequality always holds true

$$V[\bar{\bar{X}}_{st}] \leq V[\bar{\bar{X}}_{srs}] \leq V[\bar{\bar{X}}_{cl}]. \quad (19)$$

In the next section, we will show from an evaluation perspective, the great utility of this inequality. These variances are directly parameterized by: n, the number of selected distributions in the hierarchy — m, the samples per distribution, and finally — the sample design. The effect of different kinds of groupings, and the reconciliation of the observed versus expected variances provide insight into the effect of the cofactors that define those groupings. This is discussed in greater detail in the next section.

3 Methodology Outline

With the terminology defined, we can now proceed with an overview of the new methodology. The overall process consists of three steps, that are repeated as long as a violation of assumptions is detected. They are:

A. **Dichotomization**. Reduce the statistic of interest to a binary value.

B. **Assumptions**. Propose a data partition and sample design.

C. **Analysis**. Compare the sample (empirical) statistic to the point estimate average. If the statistics are not considered compatible, reject the proposed assumptions and iterate.

We now consider each step in more detail.

3.1 Dichotomization Step

The first step of the methodology is to reduce the system output to a binary value. In evaluations where the metric of interest is a rate of success or failure, this usually involves a simple thresholding of a result.

Dichotomizing the performance measures allows for the exploitation of some unique interactions between sample design and binary data. However, it is not without disadvantages. For scalar data, the binary requirement is nominal, since it can be thresholded and the analysis performed over a wide variety of thresholds. Multidimensional data or scalar data depending on multiple cofactors may require a complex transformation or an analysis in a high-dimensional space. This view could make the application of the new methodology difficult. While it might be tempting to view this as a preprocessing step and try to have a methodology that simply uses the dichotomized data, including this in the iterative process is important because a poor projection into the binary space may itself cause unmodeled clustering of the data that impacts predictability. As illustrated in [19], the thresholding of performance measures can be problematic when it forces the results into the tails of the distribution.

3.2 Assumptions Step

The second step of the methodology is to assume a set of influential cofactors along with a sample design. The grouping involves selecting a partition P based on cofactors X that the experimenter believes divides the observations Y into one of the equivalence classes such that

1. For each set, all observations within an equivalence class are mutually *iid*.

2. The observations Y are partially exchangeable[3] according to P.

[3] If, for arbitrary n, the (symmetric) distribution function $F(.,.,\ldots,.)$ of y_i is the same for all orders of the observations y, then the observations are considered *exchangeable*. If the observations are exchangeable only within an equivalence class, then the observations are considered *partially exchangeable*. Clearly, partial exchangeability is a weaker constraint than conditional *iid* [20].

Standard survey sampling theory (e.g. [14]) can be used to show that the estimators derived in Section 2.1 hold for Item 1 in the list above. Sugden [21] shows that the estimators hold for the more relaxed Item 2.

In addition to proposing a partition, an experimenter must assume a sample design. At each level of the hierarchy, the experimenter must determine the sample design — in this methodology we restrict ourselves to comparing three designs, stratified sampling (*st*), simple random sampling (*srs*), and cluster sampling (*cl*). If an influential cofactor is selected in a deterministic, repeatable fashion, this suggests that stratified sampling is in use. Cofactors that are randomly selected, but constrained for sets of observations, suggest that a cluster sampling process is present. Cluster sampling collapses to simple random sampling when the number of elements is one. That is, observation is made via the same process, and the level of the cofactor is not dependent on the trial number.

3.3 Analysis Step

Finally, in the analysis step, the sample variance of the statistic of interest is compared with the average point estimate of the sample variance. Only the means of V_{cl} and $\bar{\hat{\theta}}$ are of interest in the analysis phase, since $V[\hat{\theta}]$ can be made arbitrarily small by increasing the number of trials per experiment.

If and only if the statistic of interest is known to be of a particular distribution, then the analysis step could be made more formal by using a hypothesis test to decide whether or not to reject the proposed grouping and sample design. We recommend that should a hypothesis test be used, it be used only in the very last iteration of the methodology so that an experimenter does not reach conclusions about any particular assumptions prematurely.

The desired output of the analysis stage is not only a partition that is not rejected, but also insight into the effect of the cofactor of interest and how it is sampled. For example, consider the case where for a particular cofactor, cluster sampling is assumed but the sample variance is more similar to simple random sampling. Then, an evaluator may wish to perform another experiment in which a cofactor previously suspected as influential is demoted to a nuisance parameter.

The analysis phase might also suggest that additional randomization is needed. Suppose that no estimator — i.e., neither V_{st}, V_{srs}, nor V_{cl} — is compatible with a particular partition. The divergence of the sample variance and the point estimators could be due to a basic lack of stationarity required for obtaining a repeatable experiment. We investigate this phenomenon with a real experiment in the next section.

4 Experimentation

In this section, we apply the new methodology to the biometric system evaluation installation which we will refer to as the *Photohead* experiments. The main goal of this section is to show a systematic application of the new methodology for a vision system evaluation.

The Photohead system is a testbed designed to evaluate the influence of environmental and sensor effects have on the performance of face recognition systems. Col-

lecting sufficient data for many different subjects, over all times of day and weather conditions, is simply not feasible. Furthermore, the inherent variance in pose and subject would be confounded with the sensor and environmental effects, but separating the cofactors would require a very large amount of data. The basic function of the Photohead system is to isolate the effects of sensors/weather by displaying a sequence of facial imagery, and 'recapturing' the image via another sensor located at some distance away from the display. Such a re-imaging would effectively isolate the degradation of image quality due to environmental and sensor effects. Because the Photohead system uses a displayed image and not a live subject, the recognition results are potentially confounded with cofactors due to the imaging display. However, the effect of that confounding is much simpler than subject variations over time or pose variation. Therefore, although the Photohead system is not a replacement for using actual human subjects, it does provide a much more practical, and repeatable, mechanism for obtaining the vast amount of outdoor data required for such an analysis and also helps isolated particular variables of interest.

Functionally, the Photohead system takes a set of *original* images as input, and produces a set of *degraded images*, or *reimaged data*. A *data collection* refers to the generation of a particular set of degraded images. The software controlling the Photohead executes data collections autonomously, and was implemented with a client-server architecture. The server side of the software controlled the image display, while the clients were in charge of capture, where each client corresponded to a different sensor.

At fixed intervals throughout the day, a data collection would begin by logging the time of day, and downloading, from the National Weather Service website, the weather conditions at the nearest airport. [4] After shuffling the order in which the set of original images would be displayed (the reason for this is mentioned later), the server would display an image, and indicate that an image was ready for capture. Upon receiving the signal, each client would capture an image from their assigned sensor and signal back to the server that they had captured an image. The server, after receiving a signal from each client, would repeat the process, until every image in the original set had been recaptured.

The Photohead data was collected over several months, and across all times of day The images were degraded from a large variety of weather effects including snow, rain, fog, and wind. Although the experimental setup is straightfoward, the changes in the acquired images could hardly be considered trivial.

The Photohead data includes multiple images of each subject so a standard FERET-style [22] probe/gallery test is possible — each probe differs from the gallery by a change in some variable of interest. In a test like FERET, the same image would *never* be used in both the gallery and the probe set because not only does this not makes sense operationally, but it also changes the face recognition problem into a trivial image matching problem. An important idea in the Photohead project, a form of self-matching, is somewhat contrary to what one would do in a more traditional evaluation. Suppose, instead of a probe-gallery pair consisting of two *different* images of the same person, we used the *same* image, except that the probe version of the image undergoes

[4] Lehigh Valley International Airport in Allentown, PA

some form of digital or analog processing such as imaging through the weather. This is in keeping with the spirit of good experimental design, one parameter of interest is varied between probe and gallery; the fact the that face is in exactly same pose and lighting helps to isolate effects of weather and sensor cofactors from other cofactors. Thus the self-matching experiments, any deviation from ideal performance can therefore be attributed solely to the analog degradation or image processing.

The facial imagery used for the Photohead experiments also came from the well-known FERET database. The source data consists of 1,024 images, four each of 256 subjects. Let S represent the set of all 1,024 images, S_1 consist of the all of the first images of the 256 subjects, S_2 the second, and so on. During a single data collection, all 1,024 images are displayed, and recaptured. Let S^t represent the set of recaptured images from the data collection started at time t. Then, all 1,024 recaptured images S^t are compared via a commercial face recognition algorithm to all of the original 1,024 images S, producing a similarity matrix. Since we do not perform any normalization [1] virtual experiments can be extracted. Since we are interested in isolating the sensor and atmospheric effects, we will primarily be interested in the experiments in which S_i is a gallery and S_i^t. If re-imaging has a nominal effect, then our identification rate should be 100 percent.

4.1 Analysis

Our analysis involves computing collections of point estimators, and comparing their mean to a set of sample statistics. The evaluations consist of nested loops that group, sample, and calculate the various statistics of interest. The nested loops correspond to different *trials*, *runs*, and *experiments*. An experiment may consist of several trials. The experiment may itself be repeated in another run. Point estimators are generated at the trial level — sample (empirical) variances are determined per run. Each experimental run has an associated block of data; the particular block depending on the conditions under consideration. It is easiest to visualize the data as a large table, where each row of the table is a group (i.e., a cluster or stratum), the columns as different data collections, and within each cell is a set of scores generated by a particular probe. This set of scores can transformed to a summary statistic. To simplify the analysis, we choose recognition rank (as defined by the FRVT 2002 protocol [1]) because it is unidimensional.

Within a run, each experiment uses a random subset of the total available data, which is at most, one-half of the data. Therefore, across experiments, data is sometimes reused. This is not as desirable as having new data for each experiment, however, we save the availability of new data for intra-experiment use. That is, the statistics across experiments are primarily qualitative — where the statistics are use quantitatively, within an experiment, different trials do *not* share data.

As discussed in section 3.2, in these experiments we restrict ourselves to comparing three designs, stratified sampling (*st*), simple random sampling (*srs*), or cluster sampling (*cl*). Each design has an associated variance estimator which will be compared to the point estimator described above. For each iteration, we identify if the cofactors of interest are ignored (i.e., considered nuisance parameters), fixed, or randomized.

In out first experiment, we consider stratified sampling over the (rank one) identification rate for clear days, where the data is partitioned by subject. The corpus of data

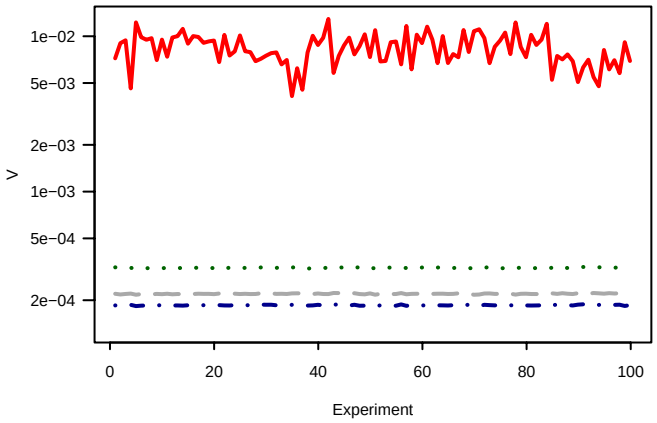

Figure 1: Baseline sample variance and average point estimator of that variance. For this case, all estimators are severely overdispersed (underestimate the true variance). The solid (red) line corresponds to the sample variance, the dotted (green) line corresponds to the mean cluster sampling point estimator, the dashed (gray) line corresponds to the mean simple random sampling point estimator, and the dot-dash (blue) line corresponds to the stratified sampling estimator.

has 1,440 such collections (as defined in the previous section), so the total experimental data is a $256 \times 1,440$ array. For each experimental run, 180 columns were selected, without replacement, from the full data array. Then, each experiment is broken into a multitude of trials, or $180/m$ matrices of size $256 \times m$ where m is the number of elements per trial. ($m = 4$ is chosen out of convenience, but other values work as well [19]).

The results of this initial partition are shown in Figure 1. In these experiments, all trials for a given experiment have the same set of fixed subject, each group within a trial has the same fixed subject, and and all other potential cofactors are ignored. As evident in the graph, the empirical variance (red, solid line) is grossly underestimated, regardless of the class of estimators, which are orders of magnitute away (the y-axis is logarithmic). This indicates that the data, used as is, fails to capture a large amount of variance. Therefore, the data is not yet well-mixed. One way of resolving this would be to model an additional level in the heiarchical distribution — i.e., try cluster stratified sampling or three-stage cluster sampling. While this might incorporate the discrepancy numerically, it would neither give us insight into *why* such a difference in variance is present, nor would it allow us to compare sample variances and point estimators, since we would be reducing our entire dataset into a single point estimator. It also suggests if we had just blindly applied a statistical analysis of the sample of convenience — e.g. using Bernoulli as a basis for hypothesis testing — we would not obtain meaningful results.

To find the source of this variance a series of lattice plots were generated (not shown). On visual inspection, it was apparent that the most influential cofactor was

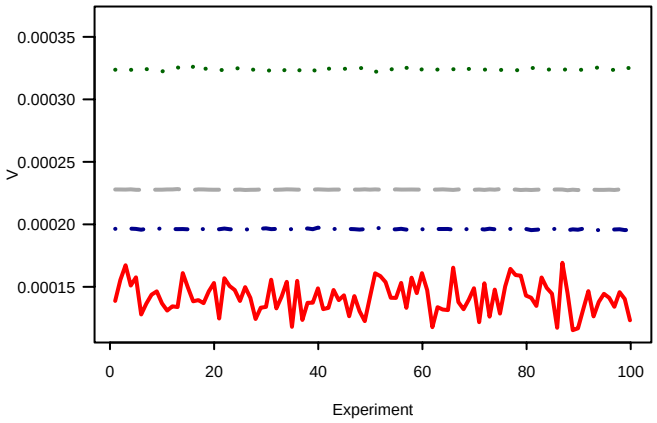

Figure 2: Sample variance (solid red line) and average point estimators of that variance when randomizing over time. (Cluster in dotten green, simple random in dashed grey, and statified in dot-dash blue.) This is a clear improvement over the previous case.

time. Therefore, to better accomodate any time-dependent variance, we introduce a randomizaiton step at the beginning of the experiment phase. If we shuffle all of the data within a row, then we simulate the condition where, within a stratum, images are taken at a random time. By shuffling each row independently, and uniformly, we reduce the dependence caused by all of the nth images within a group sharing the same timestamp. This effectively, helps homogenize the cofactor of time.

The results after temporal randomization are shown in Figure 2. Here, we have a fixed set of subjects across trials, a fixed subject per group, but unlike the previous graph, we now randomize over time within each group. This is certainly an impovement of the previous graph — the point estimates are no longer orders of magnitudes away from the sample variance. However, we have now overestimated the sample variance, for stratifed sampling, simple random sampling, and cluster sampling. This suggests we still have not selected the best criterion for our data partition.

Recalling our experimental setup, consider a partition based on degraded images corresponding to the same original image. That is, we have 1,024 instead of 256 groups since there are four original images per subject. These results are shown in Figure 3. Here, we have the a fixed set of subjects across trials, a fixed *original image* per group, but we still randomize over time per group. This graph gives strong evidence that for this Photohead installation, stratified sampling given temporal randomization is the most appropriate variance estimator. A more formal hypothesis test might be devised to test this assumption; this is a subject of future research.

These three graphs demonstrate importance of sampling design for estimation, and the methodology in action. While for this data stratified sampling variance estimator was the best match that is not always the case, different designs yield different results. Consider the case where we use different subjects per trial. Figure 4 shows the results when partitioning on original image with temporal randomization. In this case, the

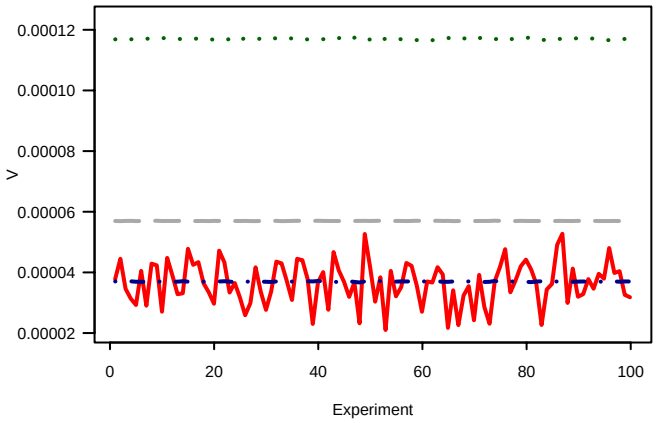

Figure 3: Sample variance (solid red line) and average point estimator of that variance when randomizing over time and grouping by (original) image. (Cluster in dotten green, simple random in dashed grey, and statified in dot-dash blue.) Finally we have agreement between the the sample variance and a class of estimator, with stratified sampling being most appropriate for this experiment.

set of images per trial is randomized, each group within a trial has a fixed original image, and time is randomized within each group. This graph demonstrates the critical nature of considering the proper sample design. Given the same underlying data, we have vastly different variances given different selection mechanisms — i.e, if we use stratified or cluster sampling.

In summary, this section showed how the new methodology can be applied to a real biometric experiment. By rejecting homogeneity assumptions that were clearly incorrect, an estimator compatible with the empirical variance was achieved. Understanding the variance and the nature of the sampling establishes the foundation for the application of techniques such as BRR or Fay's Method [5, 23], which allow for variance estimation of more general, non-linear statistics, such as the degrees of freedom estimate required for confidence intervals.

5 Acknowledgements

The authors would like to thank Jin Chu Wu (NIST) and Richard Lazarick (Transportation Security Labs) for their reviews of this paper.

6 Conclusions

In this paper we showed a new methodology for rejecting incorrect homogeneity assumptions in evaluation. We illustrated how the methodology in the context of a real biometric system evaluation. Given randomization over time, and partitioning by orig-

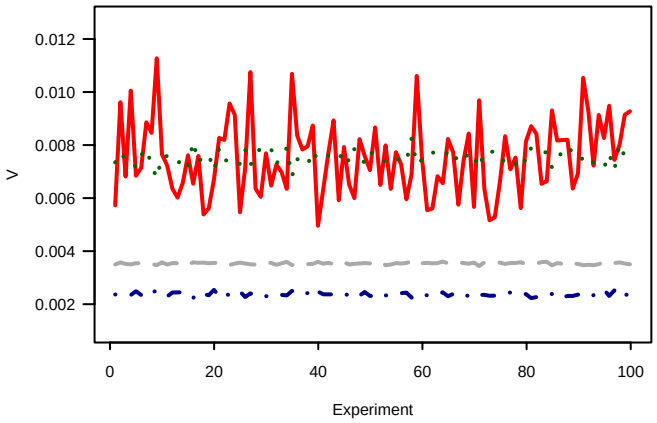

Figure 4: Sample variance (solid red line) and average point estimator of that variance when randomizing over time and grouping by (original) image. (Cluster in dotten green, simple random in dashed grey, and statified in dot-dash blue.) Between trials, different subjects are used. This time the results are compatible with cluster sampling. The sample variance point estimator shows a high variation due to the reduced number of trials per experiment — changing subjects each trial without overlap severely limits the number of independent repetitions (per experiment) that can be performed.

inal image, the Photohead data was compatible with stratified or cluster sampling, depending on whether or not the same (stratified sampling) or different (cluster sampling) original images we selected between trials.

Having a valid variance estimator is the most critical component in obtaining confidence intervals. As shown in [19], the variance estimates developed using this methodology, may be transformed into confidence intervals by combining a resampling method known as *balanced repeated replication* with the Satterthwaite approximation.

A feature that is both a strength and weakness of the new methodology is the requirement for what could potentially be a considerable quantity of data. As always, a large amount of data provides a great deal of support for a set of assumptions that are not rejected. However, the resources required for a single experiment may simply not allow enough data to be collected for a thorough application of the new methodology accross all possible grouping or hypothesized cofactors. As is common in experimental design, a pilot study using the proposed methodology could be applied to test assumptions before embarking on a more extensive data collection.

Finally, the methodology presented here is a better generalization than previous work where stratified sampling was explicitly assumed [5], or where the use of the binomial model presumes a well mixed urn [2, 24].

References

[1] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "Face Recognition Vendor Test 2002 (FRVT 2002)," National Institute of Standards and Technology (NIST), Tech. Rep. NISTIR 6965, March 2003, also available at http://www.frvt.org.

[2] J. R. Beveridge, K. She, B. Draper, and G. H. Givens, "A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, Hawaii, December 11–13 2001.

[3] R. M. Bolle, N. K. Ratha, and S. Pankanti, "Confidence interval measurement in performance analysis of biometrics systems using the bootstrap," in *Proceedings of the Third IEEE Workshop on Empirical Evaluation Methods in Computer Vision*, 2001, cVPR 2001 Workshop.

[4] D. D. Jensen, "Induction with randomization testing: Decision-oriented analysis of large data sets," Ph.D. dissertation, Washigton University, May 1992.

[5] R. J. Micheals and T. E. Boult, "Efficient evaluation of classification and recognition systems," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, Hawaii, December 11–13 2001.

[6] P. Meer and B. Georgescu, "Edge detection with embedded confidence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, pp. 1351–1365, December 2001.

[7] K. Cho, P. Meer, and J. Cabrera, "Performance assessment through bootstrap," *IEEE PAMI*, vol. 19, no. 11, pp. 1185–1198, November 1997.

[8] B. Matei, P. Meer, and D. Tyler, *Empirical Evaluation Techniques in Computer Vision*. IEEE Computer Society, 1998, ch. Performance Assessment by Resampling: Rigid Motion Estimators.

[9] L. Kish, "Confidence intervals for clustered samples," *American Sociological Review*, vol. 22, no. 2, pp. 154–165, April 1957.

[10] D. C. Montgomery, *Design and Analysis of Experiments*, 5th ed. John Wiley & Sons, 2001.

[11] L. Kish, *Survey Sampling*. John Wiley & Sons, 1965.

[12] D. Draper, J. S. Hodges, C. L. Mallows, and D. Pregibon, "Exchangeability in data analysis," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 156, no. 1, pp. 9–37, 1993.

[13] W. A. Shewhart and W. E. Deming, *Statistical Method From the Viewpoint of Quality Control*. United States Department of Agriculture, 1939.

[14] W. G. Cochran, *Sampling Techniques*, 3rd ed. John Wiley & Sons, 1977.

[15] L. Kish and M. R. Frankel, "Inference from complex samples," *Journal of the Royal Statistical Society B*, vol. 36, no. 1, pp. 1–37, 1974.

[16] I. Traat and K. Meister, "Distributional assumptions for the inference in survey sampling," in *Proceedings of the 52nd Session of the International Statistical Institute*, 1999.

[17] C.-M. Cassel, C.-E. Sarndal, and J. H. Wretman, *Foundations of Inference in Survey Sampling*. John Wiley & Sons, 1977.

[18] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. McGraw-Hill, 1991.

[19] R. J. Micheals, "Biometric system evaluation," Ph.D. dissertation, Lehigh University, 2003.

[20] B. de Finetti, *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons, 1970.

[21] R. A. Sugden, "Exchangeability and the foundations of survey sampling," Ph.D. dissertation, University of Southampton, August 1978.

[22] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, October 2000.

[23] R. Valliant, A. H. Dorfman, and R. M. Royall, *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley & Sons, 2000.

[24] J. R. Beveridge, B. Draper, K. She, and G. H. Givens, "Parametric and nonparametric methods for the statistical evaluation of human id algorithms," Colorado State University, Tech. Rep., 2001.