A Design-Based Approach for Multivariate Hypothesis Testing in Complex Surveys

Stalyn Guerrero
Universidad Nacional de Colombia, Bogota, Colombia - syguerrerog@unal.edu.co

Mario Pacheco
Universidad de Cordoba, Monteria, Colombia - mariopachecolopez@gmail.com

Guillermo Martinez
Universidad de Cordoba, Monteria, Colombia - gmartinezflorez11@gmail.com

Leonardo Trujillo*
Universidad Nacional de Colombia, Bogota, Colombia - ltrujilloo@unal.edu.co

Abstract

Classic multivariate statistical methods of inference for the mean vector and the covariance matrix are commonly based in the multivariate normal distribution and simple random sampling with replacement. However, in practice, it is common to use different complex survey designs in order to get the sample implying that the parameter estimation should incorporate the sampling weights. In this paper, an extension of the classical multivariate inference methods for the mean vector and the covariance matrix to complex survey designs is considered under a design based approach and specific density distributions for the data. The expressions for the estimators of the parameters of interest under any arbitrary sampling design are presented assuming normal and lognormal finite populations. Once these expressions are obtained, we find their asymptotic density, their corresponding confidence regions and we establish hypothesis test of the parameters. At the end, an application with actual data from a lognormal distribution is shown in order to present the advantages of the methodology using a probability proportional to size sampling design.

Keywords: Confidence regions, covariance matrix, mean vector, sampling weights.

1. Introduction.

Multivariate statistics deals with observations made on several variables of study. The aim is to study how the variables are related to each other as opposed to univariate processes with the analysis of single variables at a time. The traditional presentation in classical books of multivariate analysis (Anderson, 1984; Jobson, 1992; Johnson and Wichern, 1998) consider basically three types of methods: firstly, methods concerning the theory of estimation and testing, i.e. inferential methodds over parameters such as mean vector and covariance matrices in order to reinforce prior convictions; a second group of methods in order to investigate the dependence among variables for the purpose of prediction (e.g. Multiple Regression Analysis and Multivariate Analysis of Variance) and thirdly, data reduction or structural simplification without sacrificing valuable information that will make interpretation easier (e.g. Principal Components, Simple and Multiple Corrrespondence, Factor, Discriminant or Cluster Analysis). Most of the last methods underlie in multivariate normal assumptions of the observations and also under the assumption that the observations constitute a random sample implying that measurements taken on different items are unrelated to one another and the joint distribution of all the p variables considered remains the same for all items. In the context of survey sampling, this assumption corresponds to the assumption of a simple random sample with replacement (Cassel, Sarndal and Wretman, 1977).

Some of the methods for the second and third group above such as the estimation of coefficients of regression models (Sarndal, Swensson and Wretman, 1992), Principal Component Analysis (Skinner, Holmes and Smith, 1986), Multiple and Simple Correspondence Analysis (Ramirez and Martinez, 2010), Factor Analysis (Skinner, 1986) and the Discriminant Analysis (Canizares and Lera, 2001) have been extended to any arbitrary complex sampling design and incorporating the sampling weights. However, in our modest opinion, little has been done in order to incorporate the survey weights for the methods of the first group concerning inference about multivariate parameters under a design based approach. In this sense, Koch and Lemeshow

(1972); Koch, Freeman and Freeman (1975); Skinner (1983) have proposed model-based estimators for the mean vector and covariance matrices of a finite population. However, regarding to the inference of them and in our modest knowledge, this has not been completely extended to consider any arbitrary complex sampling design. Only some weighted likelihood methods have been proposed in order to assign unequal weights to the sample elements (Markatou, 1997, 1998). A weighted version of multivariate concepts such as likelihood ratio, Wald and score statistics was developed by Agostinelli and Markatou (2001) and they are asymptotically equivalent to those likelihood based tests that serve as reference for the multivariate estimation of parameters under some complex sampling designs as Gutierrez (2009).

2. Multivariate Inference in Survey Sampling.
This section presents the estimation of the parameters of interest when assuming that the observed numerical values are obtained using a probability sampling design. As a starting point, we will consider the existence of a finite population partially specified with the p-multivariate normal distribution.
Inference for the Multivariate Normal Distribution.
Let $\vec{\boldsymbol{X}} = \{X_1, X_2, \dots X_N\}$ being a finite population of N independent random variables assumed as $p-$variate normally distributed all with mean $\boldsymbol{\mu}$ and unknown covariance matrix Σ. The population size N could be known or unknown. The aim is to make estimations and inference of the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ according to a probability sample.
Theorem. Let a probability sample $\vec{\boldsymbol{X}}_s = \{X_1, \dots, X_k, \dots, X_n\}$,with fixed sample size n, obtained from $\vec{\boldsymbol{X}}$ according to the non-informative sampling design $p(\cdot)$ with first order and second order inclusion probabilities, π_k and π_{kl} respectively. Then, the corresponding estimators for $\boldsymbol{\mu}$ and Σ can be expressed as:

$$\bar{\boldsymbol{X}}_{\ell\pi} = \frac{1}{\hat{N}} \sum_{i=1}^{n} \frac{1}{\pi_k} X_k \quad \text{and} \quad S_{\ell\pi} = \frac{1}{\hat{N}} \sum_{k=1}^{n} \frac{1}{\pi_k} \left(X_k - \bar{\boldsymbol{X}}_{\ell\pi}\right) \left(X_k - \bar{\boldsymbol{X}}_{\ell\pi}\right)^T \tag{1}$$

with $\hat{N} = \sum_{k=1}^{n} \pi_k^{-1}$. Given that $\vec{\boldsymbol{X}}_s$ is a probability sample from a $p-$variant normal distribution, the pseudo-likelihood function (Skinner, Holt and Smith, 1989, chapter 3; Gutierrez, Trujillo, Silva, 2014, section 2.2) is given by

$$\hat{\ell}\left(\boldsymbol{\mu}, \Sigma | \vec{\mathbf{X}}_s\right) = -\frac{1}{2} \sum_{k=1}^{n} \frac{1}{\pi_k} \ln|\Sigma| - \frac{1}{2} \sum_{k=1}^{n} \frac{1}{\pi_k} \left(X_k - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(X_k - \boldsymbol{\mu}\right). \tag{2}$$

and then the main result follows. We now present some properties of these estimators:
Theorem. Let $\bar{\boldsymbol{X}}_{\ell\pi}$ and $S_{\ell\pi}$ being the estimators of $\boldsymbol{\mu}$ and Σ based in a probability sample $\vec{\boldsymbol{X}}_s$. Then, it follows that

1. $\bar{\boldsymbol{X}}_{\ell\pi}$ is an unbiased estimator of the mean vector $\boldsymbol{\mu}$.
2. The approximate covariance matrix of the vector $\bar{\boldsymbol{X}}_{\ell\pi}$ is

   $$\text{Cov}\left(\bar{\boldsymbol{X}}_{\ell\pi}\right) \doteq \mathbb{P}_\pi \Sigma$$

   with

   $$\mathbb{P}_\pi = \frac{(N-1)}{N^3} \sum_{k,l \in U} \sum \frac{\Delta_{kl}}{\pi_k \pi_l} + \frac{1}{N}$$

3. An unbiased estimator for the covariance matrix Σ is

   $$\hat{\Sigma}_{\ell\pi} = \frac{\hat{N}}{\hat{N} - \hat{N}^{(2)}} \sum_{k=1}^{n} \frac{1}{\pi_k} \left(X_k - \bar{\boldsymbol{X}}_{\ell\pi}\right) \left(X_k - \bar{\boldsymbol{X}}_{\ell\pi}\right)^T \tag{3}$$

   with $\hat{N}^{(2)} = \sum_{k=1}^{n} \pi_k^{-2}$.
4. The mean vector is such that $\bar{\boldsymbol{X}}_{\ell\pi} \sim N_p\left(\boldsymbol{\mu}, \mathbb{P}_\pi \Sigma\right)$

Example. Stratified Simple Random Sampling. Let $\vec{\boldsymbol{X}}_s = \left(\vec{\boldsymbol{X}}_{s1}, \vec{\boldsymbol{X}}_{s2}, \ldots, \vec{\boldsymbol{X}}_{sH}\right)$, a stratified simple random sampling obtained from $\vec{\boldsymbol{X}} = \left(\vec{\boldsymbol{X}}_{s1}, \vec{\boldsymbol{X}}_{s2}, \ldots, \vec{\boldsymbol{X}}_{sH}\right)$ now divided into H strata, with $\vec{\boldsymbol{X}}_{sh}$ is the sample obtained in the $h-$th stratum and $h = 1, 2, \ldots, H$ with corresponding inclusion probabilities $\pi_k = n_h/N_h$ and $\pi_{kl} = n_h\left(n_h - 1\right)/N_h\left(N_h - 1\right)$. Equation 2 can be expressed as

$$\ell_h\left(\boldsymbol{\mu}, \Sigma|\vec{\mathbf{X}}_h\right) \quad \overset{\circ}{=} \quad -\frac{N_h}{2}\ln|\Sigma| - \frac{1}{2}\frac{N_h}{n_h}\sum_{k=1}^{n_h}\left(X_{hk} - \boldsymbol{\mu}\right)^T\Sigma^{-1}\left(X_{hk} - \boldsymbol{\mu}\right)$$

Then,

$$\begin{aligned}\ell\left(\boldsymbol{\mu}, \Sigma|\vec{\mathbf{X}}_s\right) &= \sum_{h=1}^{H}\ell_h\left(\boldsymbol{\mu}, \Sigma|\vec{\mathbf{X}}_h\right) \\ &= -\sum_{h=1}^{H}\frac{N_h}{2}\ln|\Sigma| - \sum_{h=1}^{H}\frac{1}{2}\frac{N_h}{n_h}\sum_{k=1}^{n_h}\left(X_{hk} - \boldsymbol{\mu}\right)^T\Sigma^{-1}\left(X_{hk} - \boldsymbol{\mu}\right)\end{aligned}$$

The estimators of $\boldsymbol{\mu}$ and Σ obtained after maximizing the weighted log-likelihood function for $\bar{\boldsymbol{X}}_\pi$ and S_π can be expressed as

$$\bar{\boldsymbol{X}}_\pi = \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k=1}^{n_h}X_{hk} \quad \text{and} \quad S_\pi = \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k=1}^{n_h}\left(X_{hk} - \bar{\boldsymbol{X}}_\pi\right)\left(X_{hk} - \bar{\boldsymbol{X}}_\pi\right)^T.$$

These results differ with the traditional maximum likelihood estimators of $\boldsymbol{\mu}$ and Σ (under a simple random sampling with replacement).

Hypothesis Test and Confidence Regions for the Mean Vector.
This section develops inferential methods as confidence regions and hypothesis test for the mean vector in q normal $p-$variant populations. The results for the particular cases of known and equal covariance matrices, unknown but equal covariance matrices, known but unequal covariance matrices are obtained. For space issues, we only present here the particular case of unknown and unequal covariance matrices.
Theorem. Unknown and unequal covariance matrices Let $\vec{\boldsymbol{X}}_{s_1}, \vec{\boldsymbol{X}}_{s_2}, \ldots, \vec{\boldsymbol{X}}_{s_q}$ $q-$independent and identically distributed probability samples with distribution $N_p\left(\boldsymbol{\mu}_i, \Sigma_i\right)$, with unknown Σ and sizes n_i, $i = 1, \ldots, q$. Let $\bar{\boldsymbol{X}}_{\ell\pi_i}$ be the mean vector estimator $\boldsymbol{\mu}$ from the $i-$th sample. Then,

$$\Lambda_{\ell\pi} \quad = \quad -\frac{1}{2}\left(\prod_{i=1}^{q}\left[\hat{N}_i\left(\ln|S_{\ell\pi_i}^{-1}| - p - Q_i\right)\right] - \prod_{i=1}^{q}\left[\hat{N}_i\left(\ln|S_{\ell\pi_i}^{-1}| - p\right)\right]\right) \sim \chi^2_{(p)}$$

with $\hat{N}_i = \sum_{k=1}^{n_i}\pi_{ik}^{-1}$, $Q_i = \left(\bar{\boldsymbol{X}}_{\ell\pi_i} - \boldsymbol{\mu}_i\right)^T S_{\ell\pi_i}^{-1}\left(\bar{\boldsymbol{X}}_{\ell\pi_i} - \boldsymbol{\mu}_i\right)$. The proof for this theorem can be obtained using the maximum likelihood ratio criterion and properties of the weighted log-likelihood functions (see Agostinelli and Markatou, 2001). In that case when it is desirable to make hypothesis tests conditioning to unknown and different covariance matrices, we obtain the following rejection region

$$\Lambda_{\ell\pi} \quad = \quad -\frac{1}{2}\left(\prod_{i=1}^{q}\left[\hat{N}_i\left(\ln|S_{\ell\pi_i}^{-1}| - p - Q_i\right)\right] - \prod_{i=1}^{q}\left[\hat{N}_i\left(\ln|S_{\ell\pi_i}^{-1}| - p\right)\right]\right) \geq \chi^2_{(\alpha,p)}$$

with $Q_i = \left(\bar{\boldsymbol{X}}_{\ell\pi_i} - \boldsymbol{\mu}_0\right)^T S_{\ell\pi_i}^{-1}\left(\bar{\boldsymbol{X}}_{\ell\pi_i} - \boldsymbol{\mu}_0\right)$

Hypothesis Test over the Covariance Matrix.
Likelihood ratio tests are used in order to make tests of covariance matrices in a similar way to the one studied for mean vectors in the last section above.

Theorem. Let $\vec{\boldsymbol{X}}_s$ a probability sample with n observations belonging to a population $N_p(\boldsymbol{\mu}, \Sigma)$, with Σ a positive definite matrix. Then, the likelihood ratio statistic has the form

$$\Lambda_{\ell\pi} \quad = \quad -\sum_{k=1}^{n} \frac{1}{\pi_k} \left(\ln |\Sigma S_{\ell\pi}^{-1}| + \operatorname{tr} \left\{ \Sigma^{-1} S_{\ell\pi} \right\} - p \right) \sim \chi^2_{(\alpha;\, p(p+1)/2)} \tag{4}$$

This statistic $\Lambda_{\ell\pi}$ follows a χ^2 distribution with $p(p+1)/2$ degrees of freedom, being the number of different terms in Σ. If we want to make a hypothesis test over the covariance matrix, this could be stated out in the following way:

$$H_0 : \Sigma = \Sigma_0 \quad \text{against} \quad H_1 : \Sigma = \Sigma_0 \tag{5}$$

and the rejection region would correspond to those points where $\Lambda_{\ell\pi} \geq \chi^2_{(\alpha;\, p(p+1)/2)}$. We must note that the test consists on comparing Σ_0, the theoretical value with $S_{\ell\pi}$ being the estimated value according to the metric of the determinant and the trace. The hypothesis about the independence and homoscedasticity of the variables, is expressed as $H_0 : \Sigma_0 = \sigma^2 \boldsymbol{I}$, where σ^2 is the common and unknown variance. In particular, in order to test $H_0 : \Sigma_0 = \boldsymbol{I}$, a reduced statistic is given by

$$\Lambda_{\ell\pi} \quad = \quad -\sum_{k=1}^{n} \frac{1}{\pi_k} \left(-\ln |S_{\ell\pi}| + \operatorname{tr} \left\{ S_{\ell\pi} \right\} - p \right)$$

Also, sometimes it is required to test for the equality of covariance matrices prior the application of some particular multivariate techniques such as the comparison of means in two or more populations, Hotelling T^2 statistics, Multivariate Analysis of Variance, discriminant analysis, among others. Using a different approach that the one considered here, Layard (1972) have proposed a similar kind of tests. We will consider a design-based approach.

Theorem. Let $\vec{\boldsymbol{X}}_{s_1}, \vec{\boldsymbol{X}}_{s_2}, \ldots, \vec{\boldsymbol{X}}_{s_q}$, q probability samples of size n_i, from independent $N_p(\boldsymbol{\mu}_i, \Sigma_i)$ populations. Then,

$$\Lambda_{\ell\pi} \quad = \quad -\frac{1}{2^{q-1}} \left(\prod_{i=1}^{q} \left[\ln \left(\frac{|S_{\ell\pi}^{-1}|}{\exp(p)} \right)^{\hat{N}_i} \right] - \left[\ln |\Sigma^{-1}| - \operatorname{tr} \left\{ \Sigma^{-1} S_{\ell\pi_i} \right\} \right]^q \prod_{i=1}^{q} \hat{N}_i \right)$$

where

$$S_{\ell\pi_i} = \frac{1}{\hat{N}_i} \boldsymbol{A}_{\pi i} \quad \text{and} \quad S_{\ell\pi} = \frac{1}{\hat{N}} \boldsymbol{A}_{\pi}$$

with

$$\boldsymbol{A}_{\pi i} = \sum_{k=1}^{n_i} \frac{1}{\pi_{ki}} \left(X_{ki} - \bar{\boldsymbol{X}}_{\ell\pi_i} \right) \left(X_{ki} - \bar{\boldsymbol{X}}_{\ell\pi_i} \right)^T \quad \text{and} \quad \boldsymbol{A}_{\pi} = \sum_{i=1}^{q} \boldsymbol{A}_{\pi i}$$

and

$$\hat{N}_k = \sum_{k=1}^{n_i} \frac{1}{\pi_{ki}} \quad \text{and} \quad \hat{N} = \sum_{i=1}^{q} \hat{N}_i$$

follows a χ^2 distribution with degrees of freedom equal to the difference of the dimensions on the space of the parameters.

Inference for the Multivariate Lognormal Distribution.

We are also considering the particular case when the population follows a lognormal distribution. Let $\vec{\boldsymbol{Y}}_s = Y_1, \ldots, Y_n$ a probability sample with size n, obtained from the population $LN_p(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$ through a sampling design $p(\cdot)$ and let $\vec{\boldsymbol{X}}_s = X_1, \ldots, X_n$ where X_k is given for the transformation $X_k = \ln Y_k = (\ln y_{k1}, \ldots \ln y_{kp})$ for all $Y_k \in \vec{\boldsymbol{Y}}_s$. Then, X_k follows a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Then,

$$\hat{\boldsymbol{\mu}}_{\ell\pi} = \frac{\sum_{k=1}^{n} \pi_k^{-1} \ln Y_k}{\sum_{k=1}^{n} \pi_k^{-1}} \quad \text{and} \quad \hat{\Sigma}_{\ell\pi} = \frac{\sum_{k=1}^{n} \pi_k^{-1} \xi_{\pi k}}{\sum_{k=1}^{n} \pi_k^{-1}}$$

with $\xi_{\pi k} = \left(\ln Y_k - \bar{X}_\pi\right)\left(\ln Y_k - \bar{X}_\pi\right)^T$. Note that

$$\hat{\mu}_{\ell\pi Y_j} = \exp\left[\hat{\mu}_{\ell\pi j} + \frac{1}{2}\hat{\sigma}_{\ell\pi jj}\right]$$

and

$$\hat{\sigma}_{Y_{\pi jj'}} = \left\{\exp\left[(\hat{\mu}_{\ell\pi j} + \hat{\mu}_{\ell\pi j'}) + \frac{1}{2}\left(\hat{\sigma}_{\ell\pi jj} + \hat{\sigma}_{\ell\pi j'j'}\right)\right]\right\} \times \left\{\exp\left(\hat{\sigma}_{\ell\pi jj'}\right) - 1\right\}$$

with $j, j' = 1, \ldots, p$, $\hat{\mu}_{\ell\pi j}$ being the $j-$th component in the mean vector $\hat{\boldsymbol{\mu}}_\pi$ and $\hat{\sigma}_{Y\pi jj'}$ is an element in the covariance matrix $\hat{\Sigma}_{\ell\pi}$. In particular, when $j = j'$ we have the $j-$th element over the diagonal of the matrix

$$\hat{\sigma}_{Y_{\pi jj}} = \left\{\exp\left[2\hat{\mu}_{\ell\pi j} + \hat{\sigma}_{\ell\pi jj}\right]\right\} \times \left\{\exp\left(\hat{\sigma}_{\ell\pi jj}\right) - 1\right\}.$$

These individual estimators are unbiased for $\hat{\mu}_{Y_j}$ and $\hat{\sigma}_{Y_{jj'}}$.

Hypothesis Test and Confidence Regions.

Let $\vec{\boldsymbol{Y}}_s = Y_1, \ldots, Y_n$ a probability sample with size n, obtained from a population with multivariate lognormal distribution with mean vector $\boldsymbol{\mu}_Y$ and covariance matrix Σ_Y according to a sampling design $p(\cdot)$ and let $\vec{\boldsymbol{X}}_s = X_1, \ldots, X_n$ with X_k given by the transformation $X_k = \ln Y_k = (\ln y_{k1}, \ldots \ln y_{kp})$ for all $Y_k \in \vec{\boldsymbol{Y}}_s$. Then, X_k follows a $N_p(\boldsymbol{\mu}, \Sigma)$ distribution. Let $\boldsymbol{\Theta}$ be the parametric space for $\boldsymbol{\theta} = \left(\mu_{Y_1}, \mu_{Y_2}, \ldots \mu_{Y_p}, \Sigma_{Y_{jj'}}\right)$ where $\Sigma_{Y_{ij}}$ represents the $(p(p+1)/2)$ different elements in the matrix Σ_X, and we wish to test the hypothesis

$$H_0 \quad : \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \tag{6}$$

In order to make this test, Brownstein and Pensky (2008) show that if $Y = t(X)$ where X has a simple inference and t is a monotonous transformation not depending on any unknown parameters. Then, Y also has a simple inference.

3. Actual Application and Simulation.

Consider the database *calif* available at the library `pps` (Gambino, 2005) in R (R Development Core Team, 2009), considering only the third stratum with 291 observations and a bivariate lognormal distribution for the variables White (Y_1) and Amerindian (Y_2). We will use logarithm of the number of inhabitants as an auxiliary variable.

Population Parameters.

Let $\vec{\mathbf{X}} = (X_1, X_2)^T$, with $X_j = \ln Y_j = (\ln y_{j1}, \ldots \ln y_{jN})$ for $j = 1, 2$ are the values of the bivariate normal finite population with mean vector $\boldsymbol{\mu}$ and the unknown covariance matrix Σ. The corresponding scatterplot and contourns are shown in Figure 1. The normality assumption for $\vec{\mathbf{X}}$ was tested according to Royston's test (Royston, 1982), available in the library `MVN` (Korkmaz, 2014) in R, obtaining a p-value=0.5236, indicating that the population can be considered as bivariate normal distributed (this means, Y can be considered as bivariate lognormal distributed). Assuming the multivariate classical inference and considering the values of $\boldsymbol{\mu}$ and Σ maximizing the corresponding likelihood in the population, we obtain the values:

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} = \begin{bmatrix} 8.238006 \\ 4.005645 \end{bmatrix} \text{ and } \hat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \begin{bmatrix} 2.7534 & 2.3152 \\ 2.3152 & 2.7136 \end{bmatrix}$$

Then, we build the confidence region for the mean vector $\boldsymbol{\mu}$, formed by the order pairs $\boldsymbol{m} = (x_1, x_2)$ such that

$$T^2_{MSP} = N\left(\bar{\boldsymbol{X}}_{MSP} - \boldsymbol{m}\right)^T S^{-1}_{MSP}\left(\bar{\boldsymbol{X}}_{MSP} - \boldsymbol{m}\right) \leq T^2_{(\alpha, 2.290)}k \tag{7}$$

Estimation of the Parameters Ignoring the Sampling Design.

We will now consider a probability sample $\vec{\boldsymbol{X}}_s$ obtained by using a stratified pps without replacement sampling design. The considered auxiliary variable was *log of population*. The calculated sampling size was $n = 46$ observations with sample stratum sizes of 21, 12 and 13 respectively and according to proportional allocation.

After obtaining the probability sample $\vec{\boldsymbol{X}}_s$ the normality assumption was tested again according to the H Royston's test obtaining a $p\ value$=0.4188, suggesting that the set of observations in the sample $\vec{\boldsymbol{X}}_s$ can be considered bivariate normal. Under the traditional theory (not taking into account the sampling design), the estimated values were:

$$\bar{X}_{MAS} = \begin{bmatrix} 8.212732 \\ 3.918573 \end{bmatrix} \quad \text{and} \quad S_{MAS} = \begin{bmatrix} 2.463579 & 1.936044 \\ 1.936044 & 2.268236 \end{bmatrix}$$

From the classic theory of multivariate analysis, if Σ is unknown and we are considering data obtained from a random sample, the T^2 Hotelling statistic is used in order to obtain a confidence region. This region consider all the ordered pairs $\boldsymbol{m} = (x_1, x_2)$ such that

$$T^2_{MAS} \quad = \quad n\left(\bar{\boldsymbol{X}}_{MAS} - \boldsymbol{m}\right)^T S^{-1}_{MAS}\left(\bar{\boldsymbol{X}}_{MAS} - \boldsymbol{m}\right) \leq T^2_{(\alpha,2.45)} \tag{8}$$

Estimation of the Parameters Incorporating the Sampling Weights.
Under the proposed methodology, we get the confidence region

$$\Lambda = \hat{N}\log\left(1 + \mathbb{P}_\pi T^2_{\ell\pi}\right) \quad \leq \quad \chi^2_{(\alpha,2)} \tag{9}$$

where

$$T^2_{\ell\pi} \quad = \quad \mathbb{P}^{-1}_\pi\left(\bar{\boldsymbol{X}}_{\ell\pi} - \boldsymbol{m}\right)^T S^{-1}_{\ell\pi}\left(\bar{\boldsymbol{X}}_{\ell\pi} - \boldsymbol{m}\right)$$

Then, we get the values

$$\bar{\boldsymbol{X}}_{\ell\pi} = \begin{bmatrix} 8.257559 \\ 3.915200 \end{bmatrix} \quad , \quad S_{\ell\pi} = \begin{bmatrix} 1.824516 & 1.420047 \\ 1.420047 & 1.852431 \end{bmatrix}$$

and

$$\begin{aligned} \mathbb{P}_\pi \quad &= \quad \frac{N-1}{N^3}\sum_{j=1}^{3}\left(\sum\sum_{k,l\in U_j}\frac{\Delta_{kl}}{\pi_k\pi_l}\right) + \frac{1}{N} \\ &= \quad \frac{291-1}{291^3}\left(117.9879\right) + \frac{1}{291} \\ &= \quad 0.003436483 \end{aligned}$$

Figure 2 plots the confidence regions T^2_{MSP} using all the data in the finite population, T^2_{MAS} using the information in $\vec{\boldsymbol{X}}_s$ ignoring the sampling design and $\Lambda_{\pi PT}$ also using the information in $\vec{\boldsymbol{X}}_s$ but considering sampling weights. We can observe that T^2_{MAS} is the region with biggest size and contains the other two regions T^2_{MSP} and $\Lambda_{\pi PT}$. These last two regions have a matching area. Also, suppose the interest is to test the hypothesis, $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ where $\boldsymbol{\mu}_0 = [8, 3.6]^T$, we have the two alternatives; first, using the traditional methodology (for both $\vec{\boldsymbol{X}}$ and $\vec{\boldsymbol{X}}_s$) with Equations 7 and 8 and second, using the proposed methodology according to Equation 9. The obtained results are presented in Table 1.

Table 1: Comparison of Mean Vectors

N	n	$\mathbb{P}^{-1}_\pi$	T^2_{MSP}	T^2_{MAS}	$\Lambda_{\pi PT}$
291	46	290.9952	22.01247	2.256857	15.70721

The tabulated value T^2 for the Hotelling distribution with 2 and 46 degrees of freedom and α=0.05 was obtained using the approximation $T^2_{\alpha/2} = (np/(n-p+1))\, f_{(\alpha/2,p,n-p+1)}$=8.210 where $f_{(\alpha,p,n-p+1)}$ corresponds to the F(2,45) distribution. If we compare this value with T^2_{MAS} the null hypothesis is not rejected whereas is actually rejected when T^2_{MSP} is compared with $T^2_{\alpha/2}$=7.498 and $\Lambda_{\pi PT}$ is compared with $\chi^2_{\alpha/2}$=7.377.

4. Conclusions.
This paper has developed the theory of weighted maximum likelihood estimation for the mean vector and the covariance matrix under different multivariate models when a probability sample has been obtained. Different statistical inference procedures were established such as confidence regions and hypothesis tests. We considered in particular, the case of a multivariate normal and lognormal distribution under complex survey designs. We illustrate the proposed methodology with an actual example and highlight the possibility of getting different conclusions from sampling designs with equal and unequal inclusion probabilities (see also Pfeffermann, 1996). The application also suggest that a bad choice for the inference method (either maximum likelihood or weighted maximum likelihood) can produce erroneous conclusions. The results of the simulation are omitted here for space issues but they will be presented in the conference showing similar results when ignoring and when incorporating sampling weights. The results here can be easily extended for other elliptically contoured distributions.

References

1. Agostinelli, C., & Markatou, M. (2001). Test of hypothesis based on the weighted likelihood methodology. Statistica Sinica, 11: 499-514.
2. Anderson, T.W. (1984). An Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons.
3. Brownstein N. & Pensky, M. (2008). Application of transformations in parametric inference. Journal of Statistical Education, 16 (1), http://www.amstat.org/publications/jse/v16n1/brownstein.html
4. Canizares, M., & Lera, L. (2001). The effect of sampling design in the discriminant analysis for two groups. Biometrical Journal, 43(3): 343-356.
5. Cassel, C.M, Sarndal, C.E. and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: Wiley.
6. Gambino, J.G. (2005). pps: Functions for pps sampling. R Package version 0.94.
7. Gutierrez, A. (2009). Estimacion en encuestas por muestreo: Un enfoque multiparametrico. Revista Colombiana de Estadistica, 32: 76-97.
8. Gutierrez, A., Trujillo, L. & Silva, P.N. (2014). The estimation of gross flows in complex surveys with random nonresponse. Survey Methodology, 40(2): 285-321.
9. Jobson, J.D. (1992). Applied Multivariate Data Analysis. Volumes I and II. New York: Wiley.
10. Johnson, R. & Wichern, D.W. (1998). Applied Multivariate Statistical Analysis. New Jersey: Prentice Hall Inc.
11. Koch, G.G., Freeman, D.H. & Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. International Statistical Review, 59-78.
12. Koch, G.G & Lemeshow, S. (1972). An application of multivariate analysis to complex sample survey data. Journal of the American Statistical Association, 67: 780-782.
13. Korkmaz, S. (2014). MVN: Multivariate normality tests. R Package version 1.0.
14. Layard, M.W.J. (1972). Large sample tests for the equality of two covariance matrices. Annals of Mathematical Statistics, 43: 123-141. 15. Markatou, M., Basu, A. & Lindsay, B.G. (1997). Weighted likelihood estimating equations: The discrete case with application to logistic regression. Journal of Statistical Planning and Inference, 57: 215-232.
16. Markatou, M., Basu, A. & Lindsay, B.G. (1998). Weighted likelihood equations with bootstrap root search. Journal of the American Statistical Association. 93: 740-750.
17. Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. Statistical Methods in Medical Research, 5: 239-261.
18. Ramirez, J. & Martinez, G. (2010). Analisis de correspondencias a partir de una muestra probabilistica. Revista Colombiana de Estadistica, 33(2): 273-293.
19. R Development Core Team (2009). R: A language and environment for statistical computing.
20. Royston, J. (1982). An extension of Shaprio and Wilk's W test for normality to large samples. Applied Statistics, 115-124.
21. Sarndal, C.E., Swensson, B. & Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.
22. Skinner, C.J. (1983). Multivariate prediction from selected samples. Biometrika, 70: 289-292.
23. Skinner, C.J. (1986). Regression estimation and post-stratification in factor analysis. Psychometrika, 51: 346-356
24. Skinner, C.J., Holmes, D.J. & Smith, T.M.F. (1986). The effect of sample design on principal component analysis. Journal of the American Statistical Association. 81: 789-798.
25. Skinner, C.J., Holt, D. & Smith, T.M.F. (1989). Analysis of Complex Surveys. John Wiley and Sons.