TESTING HYPOTHESIS OF THE SAMPLE WEIGHTED MEAN OF SUPERPOPULATION PARAMETERS

Carlos N. Bouza, Departamento de Matemática Aplicada, Facultad de Matemática y Computación, Universidad de La Habana

ABSTRACT
A linear regression superpopulation model is assumed and the behavior of a linear function of the unit's parameters is studied. A predictor is analyzed when the interest variable has a Bernoulli distribution. Different test statistics are proposed. Under suitable conditions, well known nonparametric tests implement the proposed testing procedures.

Key words: survey weights, Two Sample Location tests, K-Sample Location tests, consistency.

RESUMEN
Es asumido un modelo superpoblacional descrito por una regresión lineal para estudiar el comportamiento de una función lineal de los parámetros de las unidades. Un predictor es analizado cuando las variables de interés siguen una distribución de Bernoulli. Diferentes estadísticos de prueba son propuestos. Bajo condiciones adecuadas, conocidas pruebas no paramétricas implementan los procedimientos de decimasía.

MSC: 62-DO5.

1. INTRODUCTION

A finite population U = {1,…,N} is going to be studied by analyzing the behavior of a function of the parameter $Y^t = [Y1,\ldots,Y_N]$. We assume that a superpopulation model

$$Y_i = \mu_i + \Delta_i \qquad (1.1)$$

characterizes the variable of interest. μ_i is an unknown parameter and β_i is a random vector such that $E[\Delta_i\Delta_j] = 0$ $\left[V_i^2\right]$ if $i \neq j$ [i = j]. A sample of n units is selected with the purpose of predicting.

$$\mu = \sum_{i\in s} w_i\mu_i / w$$

where w_i is a weight attached to unit i and

$$w = \sum_{i\in s} w_i$$

The Decision Maker [DM] wants to infer on the mechanism that generates Y. Using this framework Pothoff, Woodbury and Manton [1992] derived procedures for estimating confidence intervals [CI's] which depend of the "equivalent degrees of freedom" [EDF]. They are defined in Section 2. Bouza [1995] developed linear rank tests for the hypothesis $H_0:\mu = \mu_0$. The present paper is devoted to the development of tests for μ which do not depend of the EDF. Some properties of the superpopulation model are discussed [Section 2]. The paired sample problem is analyzed [Section 3]. The involved variable Y is considered a Bernoulli random variable. In this case the test statistic proposed permit to derive, as particular cases the Sign-Test [ST] and Wilcoxon-Signed Test [WST] statistics.

The individuals may be clustered. A stratification scheme is proposed [Section 4] for studying the related problem. A K-Sample procedure is proposed. It yields, a statistic similar to the Extension of the Median Test [EMT], if the weights are equal, and to WST when they are ranks.

Section 5 is devoted to the analysis of an example. The behavior of the tests is characterized by the proportion of rejects of the true hyphotesis when compared to the fixed value of the size of the test α.

2. THE SUPERPOPULATION FRAMEWORK

Pothoff, Woodbury and Manton [1992] derived approximate procedures for CI estimation using a superpopulation model set up. In many applications historical data may be analyzed and the a priori distribution is characterized by the model [1.1]. A sample s of n units is selected for the purpose of predicting μ. The weight w_i attached to i measures its importance. This problem arises in the study of different systems where s is fixed and observed repeteadly. That is the case in the designing of a Decision Support System for controlling the demand of electricity.

Muralidhar-Tretter [1989] proposed classic sampling schemes for computing the needed input parameters. In this problem is possible to use the available information for characterizing the a priori distribution by certain relationships and properties of the a priori expectations and covariances. Pothoff, Woodbury and Manton [1992] defined that

$$n_e = \sum_{i \in s} w_i^2 / w^2 = w[2] / w^2$$

as the EDF and used $p_i = n_e w_i / w$ as a transformed weight. Note that

$$\sum_{i \in s} p_i = \sum_{i \in s} p_i^2 = n_e$$

Then

$$\mu = \sum_{i \in s} p_i \mu_i / n_e = \sum_{i \in s} p_i^* \mu_i$$

and a naive predictor of it is

$$\hat{\mu} = \sum_{i \in s} p_i^* Y_i \ .$$

It is model unbiased and its mean squared error is

$$V(\hat{\mu}) = \sum_{i \in s} W_i^2 V_i^2 / w^2 = W[2] V^2 / w^2 = V^2 / n_e$$

and

$$S^2 = \sum_{i \in s} p_i [y_i - \hat{\mu}]^2 / [n_e - 1]$$

is a consistent estimator of

$$V^2 = \sum_{i \in s} p_i^2 V_i^2 / n_e = w^2 V^* / w[2]$$

because

$$E[S^2] = V^2 + \left[\sum_{i \in s} p_i [\mu_i - \mu]^2 - \sum_{i \in s} \left[p_i - p_i^2 \right] V_i^2 \right] / [n_e - 1]$$

tends to V^2 if $n_e \to \infty$.

To construct a confidence interval we consider the method of test inversion. We let the null hyphotesis $H_0: \mu = \mu_0$ be accepted when

$$Z^2 = \left[(\hat{\mu} - \mu_0) / \left(S / n_e^{1/2} \right) \right]^2 = z_{1-\alpha/2}^2 \qquad (2.1)$$

where $z_{1-\alpha/2}$ is the α-th percentile of the standard normal distribution. The approximate normallity may be valid, because S^2/n_e is a consistent estimator of the mean squared error of $\hat{\mu}\left[MSE(\hat{\mu})\right]$, when N and n go to infinity, see Sen [1988]. Francisco-Fuller [1991] developed CI's of the class defined by (3.1). Gains may be achieved by using Z instead of Taylor based CI's when n is relatively small.

3. THE TWO PAIRED SAMPLES PROBLEM

Suppose that we measure a variable X repeatedly in the units $i \in s$. Our inferential interest is related with the increase of X and $D_i = X_{2i} - x_{1i}$ is computed. Take $Y_i = 1$ [0] if $D_i > 0$ [$D_i < 0$]. Then μ_i = Prob [$Y_i = 1$] is the superpopulation parameter and Var [Y_i] = μ_i[1- μ_i]. The properties of $\hat{\mu}$ hold for this particular case. We will derive inferential procedures for certain classes of weights.

Take $w_i = 1$ for any $i \in s$. Then is easily derived that $p_i = 1$, $n_e = n$ and

$$\hat{\mu} = \sum_{i \in s} Y_i / n$$

Note that

$$S^2 = \left[\sum_{i \in s} Y_i^2 - n\hat{\mu}^2 \right] / [n - 1]$$

has expectation

$$E[S^2] = V^2 + \sum_{i \in s} [\mu_i - \mu]^2 / n - 1$$

When H_0: $\mu_i = \mu_0$ for any $i \in s$ the estimator S^2 is unbiased.

Commonly we are interested in testing the hypothesis

$H_0: \mu_i = 0.5$ for any $i \in s$ (3.1)

In this case $V_i^2 = 0.25$ and after some algebraic work we obtain that Z, defined in (2.1), becomes

$$Z_s = \frac{\hat{\mu} - 0.5}{[0.25/n]^{1/2}} = \frac{n\bar{y} - 0.5n}{0.5n^{1/2}} \qquad (3.2)$$

which is the normal approximation of the Sign Test statistic.

We can take w_i = rank [$|D_i|$]. In this case the parameters involved in $\hat{\mu}$ have the expressions W = n[n+1]/2 and W[2] = n[n+1] [2n+1]/6. Then

$$\hat{\mu} = 2\sum_{i\in s} w_i y_i / n[n+1] = 2T^+ / n[n+1]$$

Note that T^+ is the Wilcoxon's statistic and

$$V^* = 6\sum_{i\in s} w_i^2 \mu_i[1-\mu_i] / n[n+1][2n+1]$$

As n_e = 3n[n+1]/2[2n+1] under (3.1)

$$\hat{\mu} - \mu_0 = \frac{4T^+ - n[n+1]}{2n[n+1]}$$

Hence the test statistic Z is equivalent to

$$Z_W = \frac{4T^+ - n[n+1]}{[n(n+1)(2n+1)/1.5]^{1/2}} \qquad (3.3)$$

It is a WST statistic under the normal approximation.

Therefore the analyzed procedure generalizes the Sign Test and the Wilcoxon Signed Test. In both cases the test does not depend of the EDF because the normal approximation of Z is derived from the hypothesis of large N and n.

4. STRATIFIED SAMPLING

We will assume that U is divided into k strata of size N_j and a sample s_j of n_j units is selected. The parameter of interest is the sum of the weighted parameters of the superpopulations

$$\mu_s = \sum_{j=1}^{k} \sum_{i\in s_j} w_{ij}\mu_{ij} / w$$

where

$$w = \sum_{j=1}^{k} \sum_{i=1}^{n_j} w_{ij}$$

Take y_{ij} = 1 [0] if $D_{ij} = X_{2ij} - X_{1ij} > 0$ [$D_{ij} < 0$] as generated by the superpopulation model $Y_{ij} = \mu_{ij} + \Delta_{ij}$. Then the counterpart of the parameters derived in Section 3 are

$p_{ij} = w_{ij}n_e/w$

$$p_{.j} = \sum_{i\in s_j} p_{ij}$$

$$n_{se} = w^2 / \sum_{j=1}^{k} \sum_{i \in s_j} w_{ij}^2 = w^2 / w_s[2]$$

$$\mu_{.j} = \sum_{i \in s_j} \mu_{ij} p_{ij} / p_{.j}$$

Then we can write

$$\mu_S = \sum_{j=1}^{k} p_{.j} \mu_{.j} / n_e$$

which is a linear function of the stratum's parameter.

Mimicking the results of Section 3 a predictor of it is

$$\hat{\mu}_s = \sum_{j=1}^{k} \sum_{i \in s_j} w_{ij} y_{ij} / w = \sum_{j=1}^{k} p_{.j} \hat{\mu}_{.j} / n_{se}$$

The corresponding error, when the strata samples are independent, is given by

$$V_s = V[\hat{\mu}_s] = \sum_{j=1}^{k} \sum_{i \in s_j} w_{ij}^2 \mu_{ij} [1 - \mu_{ij}] / w^2$$

A biased estimator of it is

$$\hat{V}_s = \sum_{j=1}^{k} \sum_{i \in s_j} p_{ij} [y_{ij} - \hat{\mu}_{.j}]^2 / [n_{se} - n^*] = \sum_{j=1}^{k} \sum_{i \in s_j} p_{ij} V_{ij} / m_e$$

where

$$n^* = \sum_{j=1}^{k} \sum_{i \in s_j} p_{ij}^2 / p_{.j}$$

Its expectation is

$$E[\hat{V}_s] = V_s + \sum_{j=1}^{k} \sum_{i \in s_j} \left[p_{ij} - p_{ij}^2 / p_{.j} \right] / \left[m_e - p_{ij}^2 / n_{se} \right] V_{ij} + \sum_{j=1}^{k} \sum_{i \in s_j} p_{ij} \left[\mu_{ij} - \mu_{.j} \right]^2 / m_e \qquad (4.1)$$

When the null hypothesis

$H_0: \mu_{ij} = \mu_{0j}$, for any $i = 1,\ldots,n_j$ and $j = 1,\ldots,k$ (4.2)

is valid the third term in the right hand side of (4.1) is equal to zero. If

H_0:$\mu_{ij} = \mu_0$, for any i = 1,…,n_j and j = 1,…,k (4.3)

$\hat{V}_s$ is unbiased.

The test inversion method allows us to use the result

$$Z^2 = \left[\frac{\hat{\mu}_s - \mu_{0s}}{[V_s / m_e]^{1/2}} \right]^2 < z^2_{1-/2}$$

for deducing CI's and to test

H_0:$\mu_{ij} = \mu_{0ij}$ for any i = 1,…,n_j and j = 1,…,k. (4.4)

If the DM uses equal weights, $w_{ij} = 1$ for any i,j, we have that $p_{ij} = 1$, and $p_{.j} = n_j$. Then if (4.2) holds

$$\mu_{0s} = \sum_{j=1}^{k} \sum_{i \in s_j} \mu_{ij} / n = \sum_{j=1}^{k} n_j \mu_{0j} / n$$

Let us assume that (4.4) is true. Then

$$V[\hat{\mu}_s] = \sum_{j=1}^{k} \sum_{i \in s_j} \mu_{0ij}[1 - \mu_{0ij}] / n$$

Note that now $\hat{\mu}_s$ is the sample weighted mean

$$\hat{\mu}_s = \sum_{i=1}^{k} n_j \bar{y}_j / n$$

Generally we are interested in testing if Prob[$D_{ij} > 0$] = 0.5.

Then $V[\hat{\mu}_s]$ = 0.25 and μ_{0s} = 0.5. Hence we derive from Z_s that chi-square test statistic is given by

$$u = \sum_{i=1}^{k} [n\bar{y}_j - 0.5n]^2 / 0.25n .$$

Under a set of mild conditions $[n\bar{y}_j - 0.5n] / [0.25n]^{1/2}$ has a standard normal distribution. Hence u follows, approximatelly, a chi-squared distribution, with k degrees of freedom, because of the independence of the y_{ij}'s within and between strata. We have impossed in our hypothesis that the D_{ij}'s have a common median equal to zero. Then u is similar to the test statistic used in the Extension of the Median Test [EMT], which has a chi-squared distribution with k-1 degrees of fredom, because the common median is estimated.

Now we will combine the n observations into a single ordered sequence and w_{ij} = rank [$|D_{ij}|$]. Then

$p_{ij} = 2w_{ij}n_{se}/n[n+1]$

$p_{.j} = 2n_{se} \sum_{i \in s_j} w_{ij}/n[n+1]$

$n_{se} = 3n[n+1] / 2[2n+1]$

$$\mu_{.j} = 3\sum_{i \in s_j} w_{ij}\mu_{ij}/[2n+1]$$

and

$$\mu_s = 2\sum_{j=1}^{k}\sum_{i \in s_j} w_{ij}\mu_{ij} / n[n+1]$$

Defining

$$w_{.j} = \sum_{i \in s_j} w_{ij}$$

under (4.2)

$$\mu^*_{0s} = 2\sum_{i=1}^{k} w_{.j}\mu_{0j} / n[n+1]$$

and

$$V[\hat{\mu}_s] = 4\sum_{j=1}^{k}\mu_{0j}[1-\mu_{0j}]\sum_{i \in s_j} w_{ij}^2 / [n(n+1)]^2$$

Therefore the corresponding statistic is

$$Z^* = \frac{\sum_{j=1}^{k}\left[\sum_{i \in s_j} w_{ij}y_{ij} - w_{.j}\mu_{0j}\right]}{\left[\sum_{j=1}^{k}\mu_{0j}[1-\mu_{0j}]\sum_{i \in s_j} w_{ij}^2\right]^{1/2}}$$

If (4.3) holds the test statistic is

$$Z^{**} = \frac{2\sum_{j=1}^{k}\sum_{i \in s_j} w_{ij}y_{ij} - n[n+1]\mu_0}{2[\mu_0[1-\mu_0]\, n(n+1)(2n+1)/6]^{1/2}}$$

The most usual problem is given by establishing the hyphotesis that the k strata have a common median. That is that $\mu_{ij} = 0.5$ for any $i = 1,\ldots,n_j$ and $j = 1,\ldots,k$. In this case $n[n+1]\mu_0 = n[n+1]$ 0,5 and

$$Z^{**} = \frac{4\sum_{j=1}^{k}\sum_{i\in s_j} w_{ij}y_{ij} - n[n+1]}{[2n(n+1)(2n+1)/3]^{1/2}}$$

is the WST statistic. Then the proposed test can be treated as an extension of the WST for the k-sample problem.

5. BEHAVIOR OF THE TESTS

An ideal neighborhood with 10.000 electricity consumers was constructed. Ten percent of them were "large consumers" [hospitals, factories, etc.]. Thirty percent were "medium consumers" [medium size business, primary schools, etc.]. The rest of the consumers were mainly "families". They defined the strata.

Ten measurements of the consumption were generated and μ_i was computed for each $i \in U$. Samples were randomly generated and the interest variable were measured in the selected consumers. A set of 100 samples were examined for each sample size. The hypothetical μ was fixed by computing its true value. The proportion of rejects of H_0 was computed for each experiment

We used equal ranks weights. Two procedures were analyzed: Simple Random Sampling and Stratified Random Sampling with Proportional Allocation [$n_j=nN_j/N$]. The results of the experiments are given in Tables 5.1 and 5.2.

Note in Table 5.1 that for n = 1000 the true proportion is close to the fixed size of the test. In general the use of the rank based weights [RBW] generates better results than the other criteria. When $n \geq 500$ the assumed percent of rejects susttains that the assumed approximations are adequate for both sets of weights.

The analysis of Table 5.2 yields a result similar to those derived from Table 5.1: RBW statistic converges faster than the Equal Based Weights [EBS] statistic. For a sample fraction of 0.1 the behavior of the test are close to the expected percent of rejects of the true hypothesis. In general it is larged when stratified sampling is used. This result suggests that the approximation to the normallity is more questionable in this case.

The initial data base was transformed for ensuring that $\mu_i = 0.5$ for any $i \in U$. The median Mi of the Di´s was computed using the ten measurements of the consumption. Then $D_i^* = D_i - M_i$ was calculated. The described sampling procedure was used for analyzing the results of testing the corresponding null hypothesis. Tables 5.3 and 5.4 present the results of the tests.

The test for EBW seems to have an acceptable behavior for $n \geq 50$. The same conclusion holds for RBW for $n \geq 30$.

The analysis of the stratified case results are given in Table 5.4. Note that the perfomance of the chi-squared approximation has a better behavior than the WST statistic. Even for n = 30 it exhibits a percent of rejections close to the expected α.

.

Table 5.1. Proportion of rejects of H_0:$\mu_S = \mu_0$:
Simple Random Sampling
for three values of α

n	Equal weights			Rank weights		
	0.1	0.05	0.01	0.1	0.05	0.01
30	0.23	0.09	0.07	0.29	0.12	0.13
100	0.16	0.06	0.04	0.25	0.07	0.09
500	0.12	0.08	0.04	0.13	0.05	0.01
1000	0.09	0.05	0.02	0.11	0.04	0.01

Table 5.2. Proportion of rejects of H_0:$\mu_S = \mu_0$:
Stratified Random Sampling
for three values of α

n	Equal weights			Rank weights		
	0.1	0.05	0.01	0.1	0.05	0.01
30	0.44	0.15	0.13	0.35	0.14	0.09
100	0.35	0.09	0.05	0.29	0.10	0.07
500	0.18	0.07	0.02	0.16	0.08	0.02
1000	0.12	0.06	0.01	0.11	0.05	0.02

Table 5.3. Proportion of rejects of H_0:$\mu_S = 0.5$;
Simple Random Sampling
for three values of α

n	Equal weights			Rank weights		
	0.1	0.05	0.01	0.1	0.05	0.01
30	0.21	0.19	0.09	0.14	0.08	0.07
100	0.14	0.07	0.06	0.15	0.06	0.02
500	0.13	0.05	0.03	0.12	0.06	0.02
1000	0.11	0.04	0.01	0.09	0.05	0.01

The results of this analysis suggest that the best procedure is to stratify and to use the chi-squared approximation

Table 5.4. Proportion of rejects of H_0:$\mu_{S=}$ 0.5;
Stratified Random Sampling for three values of α

n	χ^2 Approximation			EST Approximation		
	0.1	0.05	0.01	0.1	0.05	0.01
30	0.11	0.07	0.03	0.12	0.09	0.07
100	0.11	0.07	0.04	0.14	0.08	0.03
500	0.09	0.04	0.03	0.13	0.07	0.03
1000	0.08	0.05	0.01	0.10	0.06	0.01

.

ACKNOWLEDGEMENTS

This paper is a result of a research supported by a project with the Institut für Mathematik of Humboldt - Universität zu Berlin.

REFERENCES

BOUZA, C.N. [1995]: "Linear rank tests derived from a superpopulation model", **Biometrical Journal**, 38, 497-506.

FRANCISCO, C.A. and W.A. FULLR [1991]: "Quantile estimation with a complex design", **The Ann. of Statistics**, 19, 454-469.

MURALIDHAR, K. and M.J. TRETTER [1989]: "The impact of special requirements on the estimation of the electrical demand", **Manag. Sc.** 37, 368-373.

POTHOFF, R.F.; M.A. WOODBURY, M.A. and K.G. MANTON [1992]: "Equivalent sample size and equivalent degrees of freedom refinements for inference using survey weights under superpopulation models", **J. American Stat. Ass.**, 87, 383-396.

SEN, P.K. [1988]: "Asymptotics in finite population sampling", In **Handbook of Statistics** (P.R. Krishnaiah and C.R. Rao, eds.) 6, 291-331.