Journal of Statistics: Advances in Theory and Applications
Volume 25, Number 2, 2021, Pages 51-59
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/jsata_7100122198

INFERENCES ON A NORMAL MEAN WITH AN AUXILIARY VARIABLE

JIANQI YU

College of Science
Guilin University of Technology
Guilin, Guangxi 540041
P. R. China
e-mail: jxy2416@yahoo.com

Abstract

Inferential procedures for a normal mean with an auxiliary variable are developed. First, the maximum likelihood estimation of the mean and its distribution are derived. Second, an F statistic based on the maximum likelihood estimation is proposed, and the hypothesis testing and confidence estimation are outlined. Finally, to illustrate the advantage of using auxiliary variable, Monte Carlo simulations are performed. The results indicate that using auxiliary variable can improve the efficiency of inference.

2020 Mathematics Subject Classification: 62H12, 62H15.
Keywords and phrases: auxiliary variables, maximum likelihood estimators, powers, bivariate normal.
This work is supported by NSFC(11961015).
Received April 9, 2021; Revised June 7, 2021

© 2021 Scientific Advances Publishers

This work is licensed under the Creative Commons Attribution International License (CC BY 3.0).
http://creativecommons.org/licenses/by/3.0/deed.en_US

Open Access

1. Introduction

Auxiliary information is very common and an important problem in practice. Making full use of auxiliary information can effectively improve the accuracy of inference. For instance, we use the sample mean to estimate the population mean, but when there is auxiliary information, there are other better estimates. Cochran [1] proposed the ratio estimation of the population mean in simple random sample survey, and pointed out that the ratio estimation reached the best when the research variables and auxiliary variables were highly positively correlated and the regression line passed through the origin. The product estimation was first proposed by Robson [5] and rediscovered by Murthy [4], which is suitable for the situation where the research variables and auxiliary variables are highly negatively correlated. The regression estimation proposed by Watson [8] is suitable for the case that the regression line of the research variable and auxiliary variable does not pass through the origin. In later years, many scholars proposed various methods to improve the estimation of population mean in Simple Random Sampling (SRSWOR). For details, see Singh and Tailor [6], Singh et al. [7], Yan and Tian [9], Khan et al. [3], and Kadilar [2], etc.

Unlike the above researchers who consider the problem in Simple Random Sampling, we consider the problem with normal assumption. Normal assumption is reasonable, since real data are usually normal or nearly normal. Moreover, normal population is easy to handle mathematically.

Let the research variable be y, the auxiliary variable be x, and the expectation of x is known. Together they satisfy

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N_{q+p}\left(\begin{pmatrix} c \\ \mu \end{pmatrix}, \sum = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right),$$

where c is a known constant. Suppose that there are n independent samples on both y and x, and in addition, there are m extra observations on x solely. In other words, we have a random sample as follows:

$$x_1, \ldots, x_n \quad x_{n+1}, \ldots, x_{n+m}, \quad y_1, \ldots, y_n. \tag{1.1}$$

We consider the problems of estimation and hypothesis testing of μ:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu \neq \mu_0, \tag{1.2}$$

where μ_0 is a given value of μ.

This article is organized as follows. In the following section, a procedure for estimation and hypothesis testing of the mean with auxiliary information is developed. In Section 3, to illustrate usefulness of auxiliary information, Monte Carlo simulations are conducted to compare the powers of the testing in this paper with those of the test without using auxiliary variables.

2. Inference on μ

2.1. Maximum likelihood estimation of μ

Partition the data in (1.1) as follows:

$$\mathbf{D}_1 = \begin{pmatrix} x_1, \ldots, x_n \\ y_1, \ldots, y_n \end{pmatrix},$$

$$\mathbf{D}_2 = (x_1, \cdots, x_n, \ldots, x_{n+m}). \tag{2.1}$$

Let $\bar{\mathbf{D}}_1 = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$ and $\mathbf{S} = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix}$ denote the sample mean vector and the sum of squares and sum of products matrix respectively based on $\mathbf{D}_1$. Similarly, let $\bar{D}_2$ and V denote respectively the sample mean and the sums of squares based on $\mathbf{D}_2$.

Consider the density function of data in (1.1). We note that the density of x and y can be written as the marginal density of x times the conditional density of y given x (we indicate the density of normal distribution by $f(\cdot)$ here), that is,

$$f(x,\ y|\ c,\ \mu,\ \mathbf{\Sigma}) = f(x|\ c,\ \sigma_{11})\, f(y|\ \mu_{2.1} + B_{2.1}x,\ \sigma_{2.1}),$$

where

$$B_{2.1} = \sigma_{21}\sigma_{11}^{-1},\ \mu_{2.1} = \mu - B_{2.1}c,\ \sigma_{2.1} = \sigma_{22} - B_{2.1}\sigma_{12}. \tag{2.2}$$

The likelihood function can be written as

$$L(\mu,\ \mathbf{\Sigma}) = \prod_{i=1}^{n+m} f(x_i|\ c,\ \sigma_{11}) \prod_{i=1}^{n} f(y_i|\ \mu_{2.1} + B_{2.1}x_i,\ \sigma_{2.1}). \tag{2.3}$$

The maximum likelihood estimates of σ_{11}, μ, $B_{2.1}$, $\sigma_{2.1}$ are those values that maximize (2.3). To maximize (2.3) with respect to σ_{11}, we maximize $\prod_{i=1}^{n+m} f(x_i|\ c,\ \sigma_{11})$. This procedure gives us the usual maximum likelihood estimates of the parameters of a normal distribution based on $n + m$ observations, namely,

$$\hat{\sigma}_{11} = \frac{1}{n+m} \sum_{i=1}^{n+m} (x_i - c)'\,(x_i - c). \tag{2.4}$$

To maximize (2.3) with respect to $\mu_{2.1}$, $B_{2.1}$ and $\sigma_{2.1}$, we maximize the second term of the right hand side of (2.3). This gives the usual estimates of regression parameters, namely,

$$\hat{B}_{2.1} = S_{xy}S_{xx}^{-1},\ \ \hat{\mu}_{2.1} = \bar{y} - \hat{B}_{2.1}\bar{x},\ \ \hat{\sigma}_{2.1} = (S_{yy} - \hat{B}_{2.1}S_{yx})/n. \tag{2.5}$$

It is easy to see that the maximum likelihood estimates of μ, σ_{12}, σ_{22} are obtained by solving (2.2), where $\mu_{2.1} = \hat{\mu}_{2.1}$, $B_{2.1} = \hat{B}_{2.1}$, and $\sigma_{2.1} = \hat{\sigma}_{2.1}$. Hence, we have

$$\hat{\mu} = \bar{y} - \hat{B}_{2.1}(\bar{x} - c) \text{ with } \hat{B}_{2.1} = S_{xy}S_{xx}^{-1}. \tag{2.6}$$

It is obvious that $\hat{\mu}$ is determined solely by $\mathbf{D}_1$ and it is same as regression estimator in sample survey.

2.2. Hypothesis test and confidence interval for μ

We consider conditional distribution of $\hat{\mu}$ first. After some complicated calculation, we have

$$\hat{\mu}|(x_1, \ldots, x_n) \sim N(\mu, [\frac{1}{n} + \frac{(\bar{x} - c)^2}{S_{xx}}]\sigma_{2.1}). \tag{2.7}$$

Meanwhile, we change the estimator of $\sigma_{2.1}$ in (2.5) into an unbiased estimator

$$\hat{\sigma}_{2.1} = (S_{yy} - \hat{B}_{2.1}S_{yx})/(n - 1). \tag{2.8}$$

Then we have

$$(n - 2)\hat{\sigma}_{2.1}|(x_1, \ldots, x_n) \sim \sigma_{2.1}\chi^2(n - 2).$$

Moreover, $\hat{\sigma}_{2.1}$ and $\hat{\mu}$ are independent conditional on $(x_1, \ldots, x_n)$.

Define

$$Q = \frac{(\hat{\mu} - \mu)^2}{(1/n + (\bar{x} - c)^2/S_{xx})\hat{\sigma}_{2.1}}, \tag{2.9}$$

then Q is an F statistics with

$$Q|(x_1, \ldots, x_n) \sim F(1, n - 2).$$

Since this conditional distribution is free of $(x_1, \ldots, x_n)$, we have

$$Q \sim F(1, n-2). \tag{2.10}$$

Thus, we have the testing and confidence set of the mean μ as follows:

(I) For a given level α and an observed value Q_0 of Q, the null hypothesis $\mu = \mu_0$ will be rejected whenever the p-value

$$P(F_{1-\alpha}(1, n-2) > Q_0 | H_0) < \alpha, \tag{2.11}$$

where $F_{1-\alpha}(1, n-2)$ is the $(1-\alpha)$-th quantile of the $F(1, n-2)$ distribution.

(II) An $1-\alpha$ confidence interval for μ is the set of values of μ that satisfy

$$\hat{\mu} - \sqrt{(\frac{1}{n} + \frac{(\bar{x}-c)^2}{S_{xx}})\hat{\sigma}_{2.1}F_{1-\alpha}(1, n-2)}$$

$$\leq \mu \leq \hat{\mu} + \sqrt{(\frac{1}{n} + \frac{(\bar{x}-c)^2}{S_{xx}})\hat{\sigma}_{2.1}F_{1-\alpha}(1, n-2)}. \tag{2.12}$$

3. Sizes and Power Comparison

To illustrate the advantage of using auxiliary variables, we compare the sizes and powers of two tests using and not using auxiliary variables.

First, if not using auxiliary variables, the test statistics for $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$ is

$$\bar{Q} = n(n-1)(\bar{y} - \mu_0)^2 / S_{yy}. \tag{3.1}$$

For a given level α and an observed value $\bar{Q}_0$ of $\bar{Q}$, the null hypothesis $\mu = \mu_0$ will be rejected whenever the p-value

$$P\left(F(1, n-1) > \bar{Q}_0 \,|\, H_0\right) < \alpha. \tag{3.2}$$

Second, let $\hat{T} = (\hat{\mu} - \mu_0)^2 / \hat{\sigma}_{2.1}$, then $Q = \dfrac{\hat{T}}{1/n + (\bar{x} - c)^2 / S_{xx}}$. T is the distance between the true value μ and the null value μ_0 divided by $\sigma_{2.1}$, the residual variance while regressing y on x. So, $\hat{T}$ can be thought as an estimate of the distance between μ and μ_0 adjusted by the auxiliary variable x.

Without loss of generality, it can be assumed that $\sum$ be a correlation matrix and $c = 0$. Each simulation result is based on 100,000 runs. In each run, we generate random data following the normal distribution with mean μ and covariance matrix $\sum$, and then test the hypothesis $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$. The proportion of times of rejecting H_0 in 100,000 runs is used as the estimates of the power. The estimated powers are presented in Table 1.

Table 1. Monte Carlo estimates of powers of the tests using and no-using auxiliary data (in parentheses); $\sum = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, $\alpha = 0.05$

	$(\mu - \mu_0, \rho)$					
n	(0.0,0.5	(0.3, – 0.2)	(– 0.5, – 0.4)	(0.75,0.6)	(1.0, – 0.8)	(– 1.5,0.9)
8	0.051(0.051)	0.105(0.113)	0.231(0.233)	0.544(0.449)	0.942(0.683)	0.999(0.949)
12	0.050(0.051)	0.152(0.160)	0.372(0.352)	0.797(0.662)	0.997(0.883)	1.000(0.997)
16	0.050(0.051)	0.194(0.199)	0.501(0.464)	0.920(0.801)	0.999(0.962)	1.000(0.999)
20	0.050(0.051)	0.245(0.247)	0.612(0.562)	0.971(0.889)	1.000(0.989)	1.000(1.000)
25	0.050(0.050)	0.302(0.303)	0.724(0.670)	0.992(0.949)	1.000(0.998)	1.000(1.000)
32	0.050(0.049)	0.380(0.378)	0.836(0.783)	0.999(0.984)	1.000(1.000)	1.000(1.000)

Firstly, we observe that the sizes of the test using auxiliary variable is more close to the nominal level. Secondly, when the sample size n is small, μ is close to μ_0, and y is little related to $\boldsymbol{x}$ which indicating large $\sigma_{2.1}$, the powers of two tests are about the same. However, when the sample size n becomes large, μ is far from μ_0, and y is highly related to $\boldsymbol{x}$, the powers of the test using auxiliary data are much higher. Hence, we conclude that the tests using auxiliary data is much better.

4. Concluding Remarks

In this paper, inferences on a normal mean with an auxiliary variable are considered. First, we derive the maximum likelihood estimation, confidence estimation, and hypothesis testing of the normal mean. We found that the additional observations solely on the auxiliary variable are of no use for inference on the mean of the research variable. The reason is that the expectation of the auxiliary variable is already known, and so this extra part of data have no longer useful information on the research variable. Hence, we can simply delete this part of data. Secondly, we compare the sizes and powers of two tests using and not-using auxiliary information through Monte Carlo simulations. When the sample size and the adjusted distance T between the true value μ and hypothesized value μ_0 are small, two tests are actually about the same. But when the sample size and T become large, the powers of the test using auxiliary variables are much higher. This indicates that auxiliary information can improve the efficiency of the inference for the normal mean.

References

[1] W. G. Cochran, The estimation of the yields of the cereal experiments by sampling for the ratio gain to total produce, Journal of Agriculture Science 30(2) (1940), 262-275.

DOI: https://doi.org/10.1017/S0021859600048012

[2] G. O. Kadilar, A new exponential type estimator for the population mean in simple random sampling, Journal of Modern Applied Statistical Methods 15(2) (2016), 207-214.

DOI: https://doi.org/10.22237/jmasm/1478002380

[3] S. Khan, H. Ali, S. Manzoor and Alamgir, A class of transformed efficient ratio estimators of finite population mean, Pakistan Journal of Statistics 31(4) (2015), 353-362.

[4] M. N. Murthy, Product method of estimation, Sankhya: The Indian Journal of Statistics, Series A 26(1) (1964), 294-307.

[5] D. S. Robson, Applications of multivariate polykays to the theory of unbiased ratio-type estimation, Journal of American Statistical Association 52(280) (1957), 511-522.

[6] H. P. Singh and R. Tailor, Use of known correlation coefficient in estimating the finite population mean, Statistics in Transition 6(4) (2003), 555-560.

[7] R. Singh, P. Chauhan, N. Sawan and F. Smarandache, Improvement in estimating the population mean using exponential estimator in simple random sampling, International Journal of Statistics and Economics 3(A09) (2009), 13-18.

[8] D. J. Watson, The estimation of leaf area in field crops, Journal of Agriculture Science 27(3) (1937), 474-483.

DOI: https://doi.org/10.1017/S002185960005173X

[9] Z. Yan and B. Tian, Ratio method to the mean estimation using coefficient of skewness of auxiliary variable, International Conference on Information Computing and Applications, Part II 106 (2010), 103-110.

DOI: https://doi.org/10.1007/978-3-642-16339-5_14

■