SHORT REPORT
doi: 10.2306/scienceasia1513-1874.2009.35.310

ScienceAsia 35 (2009): **310–313**

Confidence intervals for the difference between two means with missing data following a preliminary test

Pawat Paksaranuwat, Sa-aat Niwitpong*

Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand

*Corresponding author, e-mail: snw@kmutnb.ac.th

Received 2 Feb 2009
Accepted 21 Jun 2009

ABSTRACT: Unit nonresponse and item nonresponse in sample surveys are a typical problem of nonresponse which can be handled by weighting adjustment and imputation methods, respectively. The objective of this study is to compare the efficiency of confidence intervals for the difference between two means when the distributions are non-normal distributed and item nonresponse occurs in the sample. The confidence intervals considered are Welch-Satterthwaite confidence interval and the adaptive interval that incorporates a preliminary test of symmetry for the underlying distributions. The adaptive confidence intervals use the Welch-Satterthwaite confidence interval if the preliminary test fails to reject symmetry for the distributions. Otherwise, the Welch-Satterthwaite confidence interval is applied to the log-transformed data, and then the interval is transformed back. Simulation studies show that the adaptive interval that incorporates the test of symmetry performs better than the Welch-Satterthwaite confidence interval when we imputed values for the missing data in two random samples based on the random hot deck imputation method.

KEYWORDS: coverage probability, imputation, random hot deck method, Welch-Satterthwaite confidence interval

INTRODUCTION

The problem of calculating confidence intervals for the difference between the means of two independent normal distributions is an important research topic in statistics. The common way is to use the Welch-Satterthwaite confidence interval when the population variances are known to be unequal[1]. Miao and Chiou[2] compared three confidence intervals for the difference between two means when both normality and equal variances assumptions may be violated. The confidence intervals considered were the Welch-Satterthwaite interval, the adaptive interval that incorporates a preliminary test (pre-test) of symmetry for the underlying distributions, and the adaptive interval that incorporates the Shapiro-Wilk test for normality as a pre-test. The adaptive confidence intervals use the Welch-Satterthwaite interval if the pre-test fails to reject symmetry (or normality) for both distributions. Otherwise, the Welch-Satterthwaite interval is applied to the log-transformed data and the interval is transformed back. Their study showed that the adaptive interval with a pre-test of symmetry has best coverage among the three intervals considered. The aim of this paper is to generalize Miao and Chiou[2]'s confidence intervals to the missing data case.

Incomplete or missing data in sample surveys generally occurs in two ways: unit nonresponse and item nonresponse[3]. Unit nonresponse occurs if a unit is selected for the sample, but no response is obtained for the unit. Weighting adjustment is often used to handle unit nonresponse. Item nonresponse sometimes occurs for certain questions; either the questions that should be answered are not answered or the answers are deleted during editing. Item nonresponse is usually handled by some form of imputation to fill in missing item values. Brick and Kalton[4] list the main advantages of imputation over other methods for handling missing data. First, imputation permits the creation of a general purpose complete public-use data file with or without identification flags on the imputed values that can be used for standard analyses, such as the calculation of item means (or totals), distribution functions, and quantiles. Secondly, analyses based on the imputed data file are internally consistent. Thirdly, imputation retains all the reported data in multivariate analyses.

As there are a number of imputation methods, it is not immediately clear which method should be chosen, especially when an imputation method may be best in one respect but not in others[5]. Qin et al[6] proposed the random hot deck imputation method to impute the missing values for confidence intervals for the differences between two datasets with missing data but they did not consider the effect when both normality and equal variances assumptions may be violated. This paper studies the confidence intervals of the difference between two means with missing data when both normality and equal variances assumptions may be violated. We use the random hot deck imputation method to impute the missing values as in Ref. 6. We consider two confidence intervals: the Welch-Satterthwaite interval and the adaptive interval with pre-test of symmetry.

CONFIDENCE INTERVALS FOR THE DIFFERENCE BETWEEN TWO MEANS WITH MISSING DATA

Let x_i with $i = 1, \dots, n_x$ and y_j with $j = 1, \dots, n_y$ be random samples from two distributions (not necessary normal) with means μ_x, μ_y and standard deviations σ_x, σ_y, respectively. Let $\bar{x}$, $\bar{y}$, s_x^2 and s_y^2 be the sample means and variances for x and y, respectively. We are interested in the $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$ when there are missing data in both x_i and y_j.

Random hot deck imputation method

Consider the following simple random samples of incomplete data $\{x_i, \delta_{xi}\}$ and $\{y_j, \delta_{yj}\}$ associated with populations (x, δ_x) and (y, δ_y) where $\delta_{zk} = 0$ if z_k is missing, and $\delta_{zk} = 1$ otherwise, in which z is x or y and k is i or j. Generally, missing data can be classified as being non-ignorable or ignorable[7] depending, respectively, on whether the probability of missing a datum is dependent upon its value or not. There are three forms of ignorable missing data. The first is associated with sampling. In most situations it is neither efficient nor possible to obtain data from a whole population. Probability sampling is widely used to obtain a representative population sample[7]. The second form of ignorable missing data is missing at random[7]. It occurs where the pattern of missingness for a particular variable may vary for subsets. A third form of ignorable missing data is missing completely at random (MCAR)[7], where the missingness occurs at random across the whole data set[7]. Throughout this paper, we assume that the data is MCAR, i.e. $P(\delta_x = 1|x) = p_x$ and $P(\delta_y = 1|y) = p_y$ where p_x and p_y are constants. We also assume that (x, δ_x) and (y, δ_y) are independent. Let $r_x = \sum_{i=1}^{n_x} \delta_{xi}$, $r_y = \sum_{j=1}^{n_y} \delta_{yj}$, $m_x = n_x - r_x$, and $m_y = n_y - r_y$. We denote the sets of respondents with respect to x and y by s_{rx} and s_{ry}, respectively, and the sets of non-respondents with respect to x and y by s_{mx} and s_{my}. Let x_i^* and y_j^* be the imputed values for the missing data with respect to x and y, respectively. Random hot deck imputation selects a simple random sample of size m_x with replacement from s_{rx} and then uses the associated x-values as donors, i.e., $x_i^* = x_j$ for some $j \in s_{rx}$, and similarly for y_j^*. Let $z_{\mathrm{I},k} = \delta_{zk} z_k + (1 - \delta_{zk}) z_k^*$ which represent 'complete' data after imputation[6].

The Welch-Satterthwaite confidence interval with missing data

Let the estimators of μ_x and μ_y after imputation by random hot deck imputation method be defined as

$$\bar{z}_{\mathrm{I}} = \frac{1}{n_z} \sum_{k=1}^{n_z} z_{\mathrm{I},k}. \tag{1}$$

Qin et al[6] showed that

$$\sqrt{n_z}(\bar{z}_{\mathrm{I}} - \mu_z) \xrightarrow{d} N(0, (1 - p_z + p_z^{-1})\sigma_z^2), \tag{2}$$

Let t_ν^* be the $(1-\alpha/2)$ quantile of the t distribution with ν degrees of freedom. The Welch-Satterthwaite interval is defined by

$$I_{\mathrm{WS}} = (\bar{x}_{\mathrm{I}} - \bar{y}_{\mathrm{I}}) \pm t_\nu^* \sqrt{w_x + w_y} \tag{3}$$

where

$$\nu = \frac{(w_x + w_y)^2}{w_x^2/(n_x - 1) + w_y^2/(n_y - 1)},$$
$$w_z = \frac{(1 - p_z + p_z^{-1}) s_{z_{\mathrm{I}}}^2}{n_z},$$

and $s_{z_{\mathrm{I}}}^2$ is the sample variance for z_{I}.

Pre-test of symmetry used in the adaptive interval

Let $\{x_i\}$ for $i = 1, \dots, n$ be a random sample from some distribution. Following Miao and Chiou[2] the null hypothesis and alternative hypothesis of the pre-test are

H_0 : the underlying distribution is symmetric,

H_{a} : the underlying distribution is not symmetric.

The test statistic is $T = (\bar{x} - M)/J$ where $\bar{x}$ and M are the sample mean and median, and

$$J = \sqrt{\frac{\pi}{2}} \frac{1}{n} \sum_{i=1}^{n} |x_i - M|. \tag{4}$$

Note that J is a robust estimate of standard deviation. The test calls to reject the null hypothesis at the α' level of significance if $|T| \geqslant z_{\alpha'/2} \sqrt{0.5708/n}$ where $z_{\alpha'/2}$ is the upper $\alpha'/2$ percentile of the standard normal distribution.

Confidence interval when the samples are not symmetric

After imputation, if the pre-test concludes that both underlying distributions are not symmetric, we apply the Welch-Satterwaite interval I_{WS} to the log-transformed data. Then the delta method is applied to adjust the interval back to the original scale. Following Miao and Chiou[2], first we transform the data $z_{\mathrm{I},k}$ to $\log(z_{\mathrm{I},k}+c_z)$ where c_z are constants chosen to ensure that $z_{\mathrm{I},k}+c_z>0$. We then apply the Welch-Satterthwaite interval to the log-transformed data. Let $[L_{\log}, U_{\log}]$ be the Welch-Satterthwaite confidence interval obtained from $\log(z_{\mathrm{I},1}+c_z),\dots,\log(z_{\mathrm{I},n_z}+c_z)$. The first-order Taylor expansion for $\log(z_{\mathrm{I}}+c_z)$ is[2] $\log(z_{\mathrm{I}}+c_z)=\log(\mu_z+c_z)+(z_{\mathrm{I}}-\mu_z)/(\mu_z+c_z)+H$, where H is the remainder. Consequently[2], $E[\log(z_{\mathrm{I}}+c_z)]\approx\log(\mu_z+c_z)$. Let $\gamma_{\log}$ be the probability that $E[\log(x_{\mathrm{I}}+c_x)]-E[\log(y_{\mathrm{I}}+c_y)]$ is in the interval $[L_{\log}, U_{\log}]$ and let γ be the probability that $\mu_x-\mu_y$ is in the interval $[\bar{y}_{\mathrm{I}}(\mathrm{e}^{L_{\log}}-1)+L, \bar{y}_{\mathrm{I}}(\mathrm{e}^{U_{\log}}-1)+U]$ where $L=(c_y\mathrm{e}^{L_{\log}}-c_x)$ and $U=(c_y\mathrm{e}^{U_{\log}}-c_x)$. Following exactly the same steps as in Ref. 2 it can be shown that $\gamma_{\log}\approx\gamma$ and hence that the confidence interval for $\mu_x-\mu_y$ when both distributions are not symmetric is

$$I_{\log}=[\bar{y}_{\mathrm{I}}(\mathrm{e}^{L_{\log}}-1)+L,\ \bar{y}_{\mathrm{I}}(\mathrm{e}^{U_{\log}}-1)+U] \quad (5)$$

where $L=(c_y\mathrm{e}^{L_{\log}}-c_x)$ and $U=(c_y\mathrm{e}^{U_{\log}}-c_x)$.

The adaptive intervals with missing data

In the adaptive procedure we use, the adaptive confidence interval for $\mu_x-\mu_y$ incorporating the pre-test of symmetry is defined by[2]

$$I_{\mathrm{a}}=\begin{cases} I_{\log}, & \text{pre-test rejects symmetry} \\ & \text{for both imputed data sets,} \\ I_{\mathrm{WS}}, & \text{otherwise.}\end{cases} \quad (6)$$

Example: Suppose x and y are the performance values of a product from two manufacturers which are monitored by machines. We used the R program to generate sample data (x and y, sample size $n_x=n_y=20$) from a normal distribution with zero mean and unit variance. Some observations of x and y were removed to simulate missing data from machine failure or human error. The random hot deck method was used to complete the data.

The sample means and sample standard deviations are $\bar{x}_{\mathrm{I}}=0.9834$, $\bar{y}_{\mathrm{I}}=0.2297$, $s_{x_{\mathrm{I}}}=0.7281$, and $s_{y_{\mathrm{I}}}=0.2370$. We use $p_x=0.90$ and $p_y=0.85$. Because the pre-test rejects symmetry for both imputed data sets, a 95% confidence interval for the difference between μ_x and μ_y from (5) is [0.2719,0.8683].

COVERAGE PROBABILITY OF CONFIDENCE INTERVALS

Coverage probability is an important factor in judging the performance of a confidence interval. Generally, we prefer a confidence interval which has a coverage probability close to the nominal level. This section provides simulation studies for the coverage probabilities of the two confidence intervals proposed in previous section. The nominal level of the confidence interval is 95%. For adaptive confidence intervals, the level of the preliminary test is set at 10%. The symmetric distributions we consider are the normal distribution with zero mean and unit variance, t_3 (which is heavy tailed), and the uniform distribution from 0 to 1 (which is short tailed). The non-symmetric distributions we look at are the chi-squared distribution with 8 degrees of freedom (χ^2_8), which is only slightly skewed, and the lognormal distribution (with zero mean and unit variance) and exponential distribution (with parameter equal to 3) which are heavily skewed. The following two cases of response probabilities were used under the MCAR assumption

Table 1 Coverage probability of confidence interval between two means with missing data when $n_x=n_y=20$.

		σ_y/σ_x				
		0.2	0.25	1/3	0.5	1
Normal	I_{WS}	0.9375	0.9348	0.9385	0.9373	0.9449
	I_{a}	0.9410	0.9385	0.9412	0.9401	0.9479
	I_{WS}	0.9488	0.9464	0.9478	0.9431	0.9476
	I_{a}	0.9498	0.9473	0.9484	0.9442	0.9487
t_3	I_{WS}	0.9449	0.9473	0.9472	0.9502	0.9528
	I_{a}	0.9492	0.9505	0.9515	0.9527	0.9549
	I_{WS}	0.9504	0.9535	0.9536	0.9532	0.9563
	I_{a}	0.9522	0.9556	0.9558	0.9548	0.9590
Uniform	I_{WS}	0.9300	0.9302	0.9228	0.9339	0.9403
	I_{a}	0.9358	0.9361	0.9295	0.9383	0.9418
	I_{WS}	0.9400	0.9494	0.9437	0.9424	0.9475
	I_{a}	0.9425	0.9512	0.9464	0.9451	0.9483
χ^2_8	I_{WS}	0.9360	0.9398	0.9394	0.9326	0.9440
	I_{a}	0.9411	0.9441	0.9449	0.9381	0.9426
	I_{WS}	0.9422	0.9485	0.9440	0.9455	0.9456
	I_{a}	0.9444	0.9505	0.9465	0.9483	0.9448
Lognormal	I_{WS}	0.8518	0.8587	0.8652	0.9079	0.9615
	I_{a}	0.9022	0.9073	0.9070	0.9299	0.9445
	I_{WS}	0.8656	0.8793	0.8836	0.9137	0.9619
	I_{a}	0.9158	0.9246	0.9261	0.9425	0.9488
Expo	I_{WS}	0.8941	0.8999	0.9040	0.9212	0.9494
	I_{a}	0.9184	0.9233	0.9253	0.9344	0.9396
	I_{WS}	0.9157	0.9202	0.9192	0.9303	0.9550
	I_{a}	0.9379	0.9359	0.9375	0.9404	0.9474

For each distribution, the first two rows are for p_x=0.6, p_y=0.7, the third to fourth rows are for p_x=0.8, p_y=0.9.

Table 2 Coverage probability of confidence interval between two means with missing data when $n_x = n_y = 40$.

		σ_y/σ_x				
		0.2	0.25	1/3	0.5	1
Normal	I_{WS}	0.9473	0.9446	0.9447	0.9455	0.9457
	I_{a}	0.9494	0.9475	0.9477	0.9482	0.9485
	I_{WS}	0.9499	0.9525	0.9495	0.9511	0.9506
	I_{a}	0.9512	0.9533	0.9502	0.9522	0.9519
t_3	I_{WS}	0.9531	0.9554	0.9560	0.9531	0.9552
	I_{a}	0.9579	0.9603	0.9617	0.9577	0.9606
	I_{WS}	0.9589	0.9579	0.9541	0.9565	0.9530
	I_{a}	0.9618	0.9607	0.9570	0.9608	0.9564
Uniform	I_{WS}	0.9436	0.9418	0.9409	0.9455	0.9489
	I_{a}	0.9527	0.9499	0.9518	0.9539	0.9589
	I_{WS}	0.9483	0.9499	0.9480	0.9480	0.9492
	I_{a}	0.9531	0.9537	0.9528	0.9537	0.9542
χ^2_8	I_{WS}	0.9398	0.9384	0.9417	0.9465	0.9508
	I_{a}	0.9495	0.9472	0.9494	0.9522	0.9494
	I_{WS}	0.9469	0.9446	0.9443	0.9453	0.9538
	I_{a}	0.9538	0.9510	0.9488	0.9495	0.9509
Lognormal	I_{WS}	0.8801	0.8900	0.8982	0.9197	0.9612
	I_{a}	0.9526	0.9511	0.9554	0.9615	0.9449
	I_{WS}	0.8958	0.8987	0.9018	0.9283	0.9611
	I_{a}	0.9633	0.9629	0.9629	0.9702	0.9505
Expo	I_{WS}	0.9161	0.9218	0.9273	0.9356	0.9498
	I_{a}	0.9553	0.9581	0.9565	0.9561	0.9446
	I_{WS}	0.9288	0.9293	0.9322	0.9379	0.9532
	I_{a}	0.9635	0.9624	0.9629	0.9617	0.9501

Table 3 Coverage probability of confidence interval between two means with missing data when $n_x = n_y = 100$.

		σ_y/σ_x				
		0.2	0.25	1/3	0.5	1
Normal	I_{WS}	0.9475	0.9480	0.9512	0.9481	0.9476
	I_{a}	0.9499	0.9507	0.9544	0.9504	0.9503
	I_{WS}	0.9529	0.9541	0.9519	0.9533	0.9511
	I_{a}	0.9535	0.9549	0.9532	0.9541	0.9520
t_3	I_{WS}	0.9573	0.9520	0.9549	0.9566	0.9551
	I_{a}	0.9644	0.9607	0.9631	0.9631	0.9628
	I_{WS}	0.9547	0.9569	0.9566	0.9565	0.9534
	I_{a}	0.9597	0.9621	0.9608	0.9615	0.9602
Uniform	I_{WS}	0.9489	0.9519	0.9477	0.9496	0.9464
	I_{a}	0.9570	0.9588	0.9565	0.9592	0.9558
	I_{WS}	0.9516	0.9514	0.9524	0.9506	0.9556
	I_{a}	0.9570	0.9564	0.9572	0.9561	0.9602
χ^2_8	I_{WS}	0.9480	0.9452	0.9452	0.9486	0.9516
	I_{a}	0.9669	0.9626	0.9626	0.9626	0.9494
	I_{WS}	0.9485	0.9508	0.9523	0.9503	0.9491
	I_{a}	0.9664	0.9684	0.9670	0.9628	0.9480
Lognormal	I_{WS}	0.9085	0.9115	0.9243	0.9345	0.9603
	I_{a}	0.9737	0.9745	0.9782	0.9755	0.9498
	I_{WS}	0.9199	0.9227	0.9219	0.9358	0.9571
	I_{a}	0.9742	0.9770	0.9782	0.9784	0.9499
Expo	I_{WS}	0.9361	0.9341	0.9364	0.9424	0.9471
	I_{a}	0.9775	0.9740	0.9759	0.9722	0.9467
	I_{WS}	0.9394	0.9396	0.9368	0.9450	0.9493
	I_{a}	0.9762	0.9752	0.9774	0.9749	0.9487

(in which the response rates are denoted as p_x and p_y for populations x and y respectively): Case 1. $p_x = 0.6$ and $p_y = 0.7$, Case 2. $p_x = 0.8$ and $p_y = 0.9$. Sample sizes $n_x = n_y = 20, 40$ and 100 are considered. The ratio of the standard deviations (σ_y/σ_x) ranges from 0.2 to 1. The results, based on 10 000 simulations, are computed using the R program (www.r-project.org).

Tables 1–3 show that when two distributions are either symmetric or only slightly skewed, both intervals have coverage probabilities close to nominal level (0.95). However, when both distributions are skewed, I_{WS} is not acceptable as its coverage may drop below 90% in some situations, i.e the data is from Lognormal and Exponential distributions, but adaptive interval has coverage probabilities higher more than Welch-Satterthwaite interval. This result agrees with Miao and Chiou[2] studied for complete data. Further research is to find a new method for constructing the confidence interval for the difference between two means when missing data are from heavily skewed Lognormal and Exponential distributions.

Acknowledgements: The authors are grateful to the referees, the editors, and Dr. Gareth Clayton for the useful comments which improved the presentation of the paper.

REFERENCES

1. Welch BL (1938) The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350–62.
2. Miao W, Chiou P (2008) Confidence intervals for the difference between two means. *Comput Stat Data Anal* **52**, 2238–48.
3. Kalton G, Kasprzyk D (1982) Imputing for missing survey responses. In: Proceedings of the Section on Survey Research Methods, American Statistical Association, pp 22–31.
4. Brick JM, Kalton G (1996) Handling missing data in survey research. *Stat Meth Med Res* **5**, 215–38.
5. Chaimongkol W, Suwattee P (2004) Weighted nearest neighbor and regression imputation. In: Proceedings of the 9th Asia-Pacific Decision Sciences Institute Conference, Seoul.
6. Qin Y, Zhang S (2008) Empirical likelihood confidence intervals for differences between two datasets with missing data. *Pattern Recogn Lett* **29**, 803–12.
7. Mohamed N, Yahaya S, Ramli A, Abdukkah B (2008) Estimation of missing values in air pollution data using single imputation techniques. *Sci Asia* **34**, 341–5.