DOI: 10.6196/TAER.2018.24.1.1

Distributional Distortion from Missing Data Replacement in Consumer Survey*

Jane Lu Hsu**, Xiu-Ying Li***

Missing data can cause biasedness in estimation and interpretations of analytical results. This study utilizes consumer survey data to examine skewness and kurtosis distortion of missing data replacements using six different approaches to provide further insights into the features of incomplete observation restorations in consumer survey data. Results of this study indicate that replacing incomplete observations with mean or median values of complete observations can distort the distributional properties of the variables. Based on the results of two-sample Kolmogorov-Smirnov tests, replacing missing data with expected values of stepwise regressions, of multiple regressions, and of cluster means are considered preferred procedures.

Keywords: *skewness and kurtosis, missing data replacement, consumer survey*

---

* This project was funded by the Ministry of Science and Technology (102-2410-H-005-030-). Suggestions from anonymous referees are greatly appreciated.

** Professor and Director, Department of Marketing, National Chung Hsing University (Corresponding Author).

*** Former student in Executive Master of Business Administration at National Chung Hsing University. Currently working as Administrative Officer, Institutional Research Center, Office of Research and Development, National Chung Hsing University.

I. Introduction

Missing data have been a phenomenon in consumer surveys. Missing data are caused by ignorance of surveyors, unwillingness to answer, incompleteness or incapability in response to certain questions in questionnaires. Missing data can cause biasedness in estimation and interpretations of analytical results. Kao and Liu (2000) mention that the decision-making individuals sometimes are unable to provide certain data, and then the probability distributions of imprecise data need to be reconstructed. Koslowsky (2002) describes missing data as questions without answers or variables without cases.

In cases that complete observations are used in analyses while the incomplete observations are salvaged to extract limited information, Griliches (1986) refers this as the 'ignorable case.' Afifi and Elashoff (1967) state that the missing data can be considered as functions of additional parameters. Further research regards missing data as random variables and can be replaced by real values under certain assumptions (Dagenais, 1973; Gourieroux & Monfort, 1981). Greene (2000) suggests two schemes to replace missing data using the zero-order or the first-order regressions. The zero-order regression refers to replacing the missing values by the means of the complete observations. The first-order regression is to use predicted values of the variables resulting from a regression model to restore the missing observations.

Sargan and Drettakis (1974) use simultaneous equation models and treat missing data as parameters. Using an iterative procedure, the predicted values of missing data are re-estimated with the parameters of the whole sample period. Greene (2000) mentions that properties of predicted values of variables used in replacing missing observations are not thoroughly discussed in literature.

Vriens and Melton (2002) suggest several methods to replace missing data, including replacing the missing data by the mean values, using the grouping techniques and replacing by the nearest complete record (random hot-deck imputation), replacing by the predicted values based on regressions using other variables with complete observations, and utilizing model-based multiple imputation for the missing values. Little and Rubin (1987) describe the method of the Bayesian approach to restore nonignorable missing data. The advantage of utilizing the Bayesian approach is that uncertainty about the value of an unknown parameter can be expressed using a probability distribution (Judge, Griffiths, Hill, Lütkepohl, & Lee, 1985). The influence of the missing data can be assessed by a probability of the predictive distribution of the complete-data statistics (Little & Rubin, 1987).

This study utilizes consumer survey data to examine skewness and kurtosis distortion of missing data replacements using six different approaches to provide further insights into the features of incomplete observation restorations in consumer survey data. The missing data replacement approaches utilized in this study include methods of replacing by the mean values and by the median values of the compete observations, using predicted values of stepwise regressions and multiple regressions to replace missing observations, using cluster means to replace missing observations within clusters, and utilizing the Bayesian approach to estimate the posterior distributions.

II. Methodology

Missing data can be classified into several categories. When data is missing completely at random (MCAR), it is independent of observed or unobserved variables. MCAR is theoretically applicable but rarely found in empirical studies (Zhong, Hu, & Penn, 2018). Another type of missing data is missing at random

(MAR), which assumes probability of missing data is related to observed data but not to unobserved data (Zhong et al., 2018). MAR assumption holds for this study. The other type of missing data is missing not at random (MNAR) which implies missing data is related to unobserved value in variables. Verification of MNAR is impractical unless unobserved values can be obtained (Zhong et al., 2018).

Zurada (2012) explains different approaches for missing data replacement. Eliminating samples with missing data or ignorance (treating missing data as "do not care" values) is considered applicable approaches in practice. For replacement of missing data, values of observed data can be applied, including mean, mode, median, predicted values, or using imputation methods.

Gourieroux and Monfort (1981) apply linear models in missing data replacement. Donner (1982) discusses procedures used in dealing with missing data in multiple regression analysis.

2.1 Methods used in missing data replacement

Replacing missing data by the mean values: The dataset can be partitioned into subsets of complete and incomplete observations for the respective variables. The mean value for each variable based on available observations is calculated to substitute for the missing observations.

Replacing missing data by the median values: This method also uses the complete observations to generate the distributions. The median values are used to substitute for the missing observations.

Replacing missing data by the predicted values of stepwise regressions: Although stepwise regressions are not recommended for finalizing regression models, the technique can be used for replacing missing values. The explanatory variables included in the stepwise regression models are possible influential

variables like trust, risk (in different forms), personal confidence, whether an incident is cleared up (from television, news, or government announcement), demographic variables, etc. Each stepwise procedure needs to be converged before the final results are obtained.

Replacing missing values using predicted values of multiple regressions: Vriens and Melton (2002) refer this method to 'model-based mean imputations' and state that this method makes better use of the data structure. This method includes two steps. First, a regression is performed on the variable with incomplete observations with other variables as predictors. Second, the predicted value of each missing observation is imputed as the replacement. This procedure is repeated for all the variables with incomplete observations. The explanatory variables included in the regression models are identical, including trust, risk factor, personal confidence, word-of-mouth, reputation of the food item after incident, and demographic variables.

Replacing missing data by the cluster means: Complete observations are used in the clustering procedure to form groups with similar attributes. The most commonly used technique for grouping individuals or households with similar characteristics is the cluster analysis (Hair, Anderson, Tatham, & Black, 1992). The essential criterion needed for cluster analysis is to classify experimental units into classes or groups, so that the units within a class or group are similar to one another while units in distinct classes or groups are different (Johnson, 1998). Two basic methods, hierarchical and the nonhierarchical methods, can be used to search and define clusters (Hand, 1981). The hierarchical method separates units into different groups in a nested sequence of clustering. The major defect of hierarchical method is that subsequent steps can never repair mistakes made in earlier steps (Kaufman & Rousseeuw, 1990). The nonhierarchical method, referred to as the *K*-means clustering method, is a popular practical technique in the clustering analysis (Afifi & Clark, 1990) and is applied in this study. Observations with incomplete data are grouped using the variables of

complete observations or weaved among observations with complete data. The mean values of clusters are used to substitute the missing values of observations within clusters.

Replacing missing data using the Bayesian approach: Assuming the prior probabilities of variable x are known, the posterior probabilities of observations with missing data can be calculated using the Bayes' theorem:

$$p(\varphi|x)=\frac{q_{\varphi}f_{\varphi}(x)}{f(x)} \tag{1}$$

where $f(x)$ is the unconditional density distribution of x, q_{φ} is the prior probability of observations having the values represented by parameter φ, $f_{\varphi}(x)$ is the value-specific density distribution of x, and $p(\varphi|x)$ is the posterior probability of the observation having specific value. The Bayesian approach is used to generate the values of missing data with the highest posterior possibilities. A nonparametric method, kernel density estimates with normal kernels, applying with unequal bandwidths in smoothing parameters is used. Vriens and Melton (2002) outline Bayesian approach in missing data replacement.

2.2 Comparisons of distributional properties with missing data replacements

General distributional properties of variables before and after missing data replacements are presented in this study for comparisons. Missing data restorations are not to reshape the distributions of the variables to the normal distributions. The rationale of missing data replacements is to retain the original distributional properties of the variables so the subsequent analyses are not severely affected by data restorations.

Distributional properties, skewness and kurtosis, are examined in addition to mean and variance. Zurada (2012) examined distributional properties of missing data replacement using mean, standard deviation, and skewness. Kurtosis is added in this study to reflect distributional distortion from missing data replacement due to its applicability in statistical tests. The skewness of a variable describes the asymmetry of the probability distribution around the sample mean. Negative skewness indicates that the data are spread out more to the left of the mean, and positive skewness specifies that the data are spread out more to the right of the mean. The skewness of a distribution is the following.

$$\text{skewness}(x) = \frac{E\left[(x-\mu)^3\right]}{\sigma^3} \tag{2}$$

The kurtosis of a variable measures the peakedness of the probability distribution. Normal distributions have kurtosis of three. More peaked distributions have kurtosis larger than three, and less peaked distributions have kurtosis less than three. The kurtosis of a distribution is the following.

$$\text{kurtosis}(x) = \frac{E\left[(x-\mu)^4\right]}{\sigma^4} \tag{3}$$

III. Data

An in-person consumer survey was administered in Taipei, Taichung, Kaohsiung metropolitan areas from May 6th to 20th, 2014 to examine consumer awareness of various categories of food safety risks (Endnotes 1). In person consumer survey was administered following age and gender distribution between the ages of 18 and 59 using quota sampling method. The age and gender distribution of the population in three metropolitan areas are listed in Table A.1

in the Appendix. The date was extracted from the database in Department of Household Registration Affairs, Ministry of Interior. The sampling distribution is listed in Table A.2 in the Appendix. Total valid samples are 510. Gender and age distribution of the samples slightly deviated from the population due to the sampling method.

Demographical distribution of samples is listed in Table A.3 in the Appendix. Approximately half of the respondents are male, and close to 60% of respondents are married. Average age is 38.89 years old. Slightly more than half of respondents are with college/university degrees, and less than a quarter of respondents are with graduate degrees. About 28% of respondents are working in business sector, 14% of them are students, and less than 10% of them are housewives.

Food and Drug Administration, Ministry of Health and Welfare, has announced categorization of food safety risk levels on November 20th, 2013, in Taiwan. The first category of food safety risk level refers to "immediate danger upon intake," the second category refers to "poses no immediate danger but does not meet the food-sanitation standards as regulated by domestic laws and regulations," the third category refers to "exaggerated labeling with fake ingredients or adulteration," and the fourth category refers to "incomplete or nonfactual labeling." The aim of risk categorization in food safety issues is to educate consumers about various risk levels involved with food safety incidents (Table 1).

The dataset used in this study are real consumer survey data to reduce the limitations of applicability of the results. Respondents needed to answer all of the relevant questions regarding various levels of food risk incidents, risk perceptions, and willingness to pay for a food item worth NT$300 based on different categories of food safety risks.

Table 1. Categorization of food safety risk levels

	Description
Level I	Immediate danger upon intake
Level II	Poses no immediate danger but does not meet the food-sanitation standards as regulated by domestic laws and regulations
Level III	Exaggerated labeling with fake ingredients or adulteration
Level IV	Incomplete or nonfactual labeling

Source: Food and Drug Administration, Ministry of Health and Welfare (2013).

The considerations are to have the datasets with missing observations replaced in a way that the general statistical properties are not seriously distorted. The closer the properties of the dataset with restorations are to the properties of the originals, the better the replacements are.

Variables used in missing data replacements are willingness-to-pay (WTP) measures of risk-free product worth of NT$300. WTP measures are commonly applied in research to generate data for quantitative analyses. In this study, WTPs are measured using the payment card method. Ranges of WTP were provided for each risk category (referred to level in Tables), and equally divided into 8 smaller ranges for respondents to choose from.

In demographics, 50.78% of the respondents were females and 58.22% of the respondents were married. The average age of respondents was 38.89 years old, living in households of 3.89 persons on average. Slightly more than a quarter of respondents worked in the business sector, and additional 14.31% were students. Less than 10% of respondents were housewives. Average monthly household income was NTD 93,701 approximately.

IV. Results

This study examines distributional distortions of missing data replacements in consumer surveys. In missing data replacements, one-fifth of observations from a large-scale consumer survey were removed from the dataset. In the process of data deletion, observations were coded with sequential numbers and the deleting process was randomized using SAS software to remove one-fifth of the samples, a total of 102 observations.

The first and second methods of missing data replacements use mean and median values, which are commonly applied in applications. Results are listed in Table 2. Mean values remain relatively stable, but the variances of restored samples are reduced. For skewness, replacing the missing data by median values skews the data distributions more than replacing by the mean values. Both methods for missing data replacements cannot maintain the distributional properties at a higher moment. The kurtosis of the distribution becomes much larger with either mean or median value replacement. In sum, both mean and median values distorted distributional properties of observations once used in missing data replacement. These two commonly applied methods in applications need to be used with caution.

The third method in missing data replacement is to use the predicted values of stepwise regressions. The explanatory variables included in the stepwise regression models are trust, risk (in different forms), personal confidence, whether an incident is cleared up (from television, news, or government announcement), demographic variables, etc. Each stepwise regression includes different combination of explanatory variables. The fourth method is multiple regression models as suggested by Gourieroux and Monfort (1981). The explanatory variables included in the regression models are identical, including

trust, risk factor, personal confidence, word-of-mouth, reputation of the food item after incident, and demographic variables. Expected values of regression models are used for missing data replacement. As indicated in Table 2, variances are shrunk using expected values from regressions to replace missing observations. Peakedness of data distribution (kurtosis) is reduced, and data is more likely to be negatively skewed. In sum, predicted values of regression models can maintain unbiasedness of mean values.

The results of missing data replacement using stepwise regressions and using multiple regressions differ, especially variance, due to the fact that stepwise regression procedure is to find the best-fit models using 'data dredging' as described in Ray-Mukherjee et al. (2014) while multiple regression is to predict dependent variable with a set of meaningful independent variables. Different results in variances can be contributed by computational algorithms.

The fifth method is to use a nonhierarchical clustering method to replace missing data that have been randomly removed. The mean values of clusters were utilized to replace missing values. The nonhierarchical clustering method provides the mean values close to the original means. Variances are not largely shrunk, and distributional properties of higher moments (i.e., skewness and kurtosis) are close to those of the original distributions. The sixth method uses the Bayesian approach in order to generate the values of missing data with the highest posterior possibilities. A nonparametric method, kernel density estimates with normal kernels, with unequal bandwidths in smoothing parameters is used. The general statistical properties of the original data and the data with restorations can be referred to results in Table 2. Replacing missing data with Bayesian approach provides distributional properties close to those using cluster means, but are slightly skewed.

Table 2. Distributional properties of missing data replaced by various methods

WTP by Categories	Dist. properties	Original data	Missing data replaced by					
			Mean values	Median values	Expected values of stepwise regressions	Expected values of multiple regressions	Expected values of cluster means	Expected values using Bayesian approach
Level I	Mean	63.94	63.08	59.29	62.05	62.56	64.52	62.78
Level II		61.74	61.08	57.74	60.27	60.82	62.60	60.77
Level III		42.44	41.55	36.06	41.24	41.21	42.41	41.18
Level IV		38.97	38.15	33.35	37.40	37.38	38.73	37.58
Level I	Variance	3,644.55	2,910.19	2,961.19	514.43	380.66	3,394.57	2,919.13
Level II		3,591.24	2,892.60	2,933.12	425.12	248.81	3,423.81	2,902.06
Level III		2,460.48	1,977.68	2,088.56	110.06	144.80	2,189.65	1,982.23
Level IV		2,416.47	1,932.63	2,016.83	212.27	122.13	2,051.70	1,940.11
Level I	Skewness	0.57	0.67	0.85	0.12	0.17	0.59	0.68
Level II		0.70	0.81	0.98	0.31	0.10	0.72	0.82
Level III		1.32	1.56	1.76	-0.07	-0.42	1.36	1.58
Level IV		1.55	1.80	1.99	0.25	-0.10	1.65	1.83
Level I	Kurtosis	2.09	2.64	2.77	3.11	2.90	2.11	2.64
Level II		2.32	2.93	3.07	3.39	3.04	2.29	2.93
Level III		4.09	5.38	5.61	2.13	3.04	4.39	5.41
Level IV		4.80	6.31	6.59	3.45	2.88	5.55	6.35

Source: Analytical results from this study.

In order to examine equality of distributions, two-sample Kolmogorov-Smirnov test is applied in this study. Distribution of original data (before data deletion) and distribution after missing data replacement are compared using asymptotic Kolmogorov-Smirnov values. The null hypothesis is that two distributions are identical, and the alternative hypothesis is that two distributions are statistically different. Results of Kolmogorov-Smirnov tests are shown in Table 3.

Failing to reject null hypotheses of equality in distributions using two-sample Kolmogorov-Smirnov tests indicates two distributions are asymptotically identical. In results, six out of 24 tests failed to reject null hypotheses at 5% significance level, two of each from expected values of stepwise regressions, of multiple regressions, and of cluster means, indicating distributional properties are remained asymptotically in missing data replacement. Rejecting null hypotheses of two-sample Kolmogorov-Smirnov tests indicates two distributions are not asymptotically identical. In this study, the majority of missing data replacement procedures generate asymptotically unidentical distributions. The results reveal that missing data replacement, even with complicated procedures, can distort distributional properties.

Based on the results of two-sample Kolmogorov-Smirnov tests, replacing miss data with expected values of stepwise regressions, of multiple regressions, and of cluster means are considered preferred procedures. The findings in this study need to be inferred with caution due to sampling and distributional properties of datasets.

Table 3. Results of Kolmogorov-Smirnov two-sample test

Procedure	Asymptotic KS value	*p*-value
Missing data replaced by mean values		
Level I	1.6595	0.0081
Level II	1.6595	0.0081
Level III	1.6908	0.0066
Level IV	1.6595	0.0081
Missing data replaced by median values		
Level I	1.6595	0.0081
Level II	1.6595	0.0081
Level III	1.6908	0.0066
Level IV	1.5343	0.0180
Missing data replaced by expected values of stepwise regressions		
Level I	1.2211	0.1013
Level II	1.1585	0.1365
Level III	1.5029	0.0218
Level IV	1.5029	0.0218
Missing data replaced by expected values of multiple regressions		
Level I	1.1272	0.1575
Level II	1.1898	0.1178
Level III	1.4403	0.0316
Level IV	1.5969	0.0122
Missing data replaced by expected values of cluster means		
Level I	1.1585	0.1365
Level II	1.1898	0.1178
Level III	1.5029	0.0218
Level IV	1.6595	0.0081
Missing data replaced by expected values using Bayesian approach		
Level I	1.6595	0.0081
Level II	1.5343	0.0180
Level III	1.5029	0.0218
Level IV	1.6595	0.0081

Source: Analytical results from this study.

V. Conclusion

This study utilized real consumer survey data to examine the skewness and kurtosis in distributional properties using six missing data replacement approaches. The considerations are to have the datasets with missing observations replaced while the general statistical properties remain unchanged.

The results of this study indicate that when the missing observations are accounted for large proportions of the dataset, using the mean or median values of complete observations to replace the missing values can seriously affect the distributional properties of the variables. Based on the results of two-sample Kolmogorov-Smirnov tests, replacing miss data with expected values of stepwise regressions, of multiple regressions, and of cluster means are considered preferred procedures. Furthermore, researchers who need to replace missing data in large-scale consumer surveys may need to use different approaches and examine equality of distributions to determine the suitable method for missing data replacements.

Missing data can be considered as challenges related to the characteristics of the data (Sivarajah, Kamal, Irani, & Weerakkody, 2017). Findings in this study coincide with Davenport (2014) explained in using large-scale datasets that more continuous approach in sampling, analyzing, and acting on data is essential. Recent study by Zhong et al. (2018) also confirms that applying mean value replacement in handling missing data issue is not suggested due to distortion of the variable distribution. Multiple imputation methods are more reliable in missing data replacement.

The contributions of this study are three-folded. The first contribution is to reveal importance of a commonly ignorant issue in handling missing data in

consumer surveys. The second contribution of this study is to demonstrate various approaches in missing data replacement, and provide statistical properties in mean, variance, skewness, and kurtosis for comparisons. The third contribution of this study is the use of two-sample Kolmogorov-Smirnov tests to examine quality of distribution in missing data replacement.

This study is an exploratory examination of statistical properties in missing data replacement. Restrictions of this study are that only six methods are applied. Further studies may use different types of data and may consider other advanced method like neutral network-based analysis for missing data replacement as suggested in Gheyas and Smith (2010).

Endnotes

1. Data used in this study is from Li (2014). This study is a complete different approach using the same data. Original research perspectives can be found in Li (2014).

References

Afifi, A. A., & Clark, V. (1990). *Computer-aided multivariate analysis* (2nd ed.). New York, NY: Van Nostrand Reinhold.

Afifi, A. A., & Elashoff, R. M. (1967). Missing observation in multivariate statistics II: Point estimation in simple linear regression. *Journal of the American Statistical Association*, *62*(317), 10-29.

Dagenais, M. G. (1973). The use of incomplete observations in multiple regression analysis. *Journal of Econometrics*, *1*(4), 317-328.

Davenport, T. H. (2014). How strategists use 'big data' to support internal business decisions, discovery and production. *Strategy and Leadership*, *42*(4), 45-50.

Department of Household Registration Affairs, Ministry of the Interior, *Population by age and sex*. Retrieved March 31, 2014, from https://www.ris.gov.tw/zh_TW/346.

Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *The American Statistician*, *36*(4), 378-381.

Food and Drug Administration, Ministry of Health and Welfare. (2013). *Categorization of food safety risk levels*. Retrieved March 31, 2014, from https://www.mohw.gov.tw/fp-3218-22752-1.html.

Gheyas, I. A., & Smith, L. S. (2010). A neural network-based framework for the reconstruction of incomplete data sets. *Neurocomputing*, *73*(16-18), 3039-3065.

Gourieroux, C., & Monfort, A. (1981). On the problem of missing data in linear models. *The Review of Economic Studies*, *48*(4), 579-586.

Greene, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Griliches, Z. (1986). Economic data issues. In Z. Griliches & M. Intriligator (Eds.), *Handbook of econometrics* (Vol. 3, pp. 1165-1514). Amsterdam, Netherlands: North Holland.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1992). *Multivariate data analysis with readings* (3rd ed.). New York, NY: Macmillam Publishing Company.

Hand, D. J. (1981). *Discrimination and classification.* Chichester, UK: John Wiley & Sons.

Johnson, D. E. (1998). *Applied multivariate methods for data analysis.* California, CA: Brooks/Cole Publishing.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., & Lee, T. C. (1985). *The theory and practice of econometrics* (2nd ed.). New York, NY: John Wiley & Sons.

Kao, C., & Liu, S. T. (2000). Data envelopment analysis with missing data: An application to university libraries in Taiwan. *Journal of the Operational Research Society*, *51*(8), 897-905.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data - An introduction to cluster analysis.* New York , NY: John Wiley & Sons.

Koslowsky, S. (2002). The case of the missing data. *Journal of Database Marketing*, *9*(4), 312-318.

Li, X.-Y. (2014). *Risk perception of food safety incident categorization and its linkage to repeat purchase* (Unpublished master's thesis). Executive Master of Business Administration, National Chung Hsing University, Taiwan. (in Chinese)

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York, NY: John Wiley & Sons.

Ray-Mukherjee, J., Nimon, K., Mukherjee, S., Morris, D. W., Slotow, R., & Hamer, M. (2014). Using commonality analysis in multiple regressions: A tool to decompose regression effect in the face of multicollinearity. *Methods in Ecology and Evolution*, *5*(4), 320-328.

Sargan, J. D., & Drettakis, E. G. (1974). Missing data in an autoregressive model. *International Economic Review*, *15*(1), 39-58.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, *70*, 263-286.

Vriens, M., & Melton, E. (2002). Managing missing data. *Marketing Research*, *14*(3), 12-17.

Zhong, H., Hu, W., & Penn, J. M. (2018). Application of multiple imputation in dealing with missing data in agricultural surveys: The case of BMP adoption. *Journal of Agrucultural Resource Economics*, *43*(1), 78-102.

Zurada, J. (2012). Does removing/replacing missing value improve the models' classification performances? *International Journal of Management and Information Systems*, *16*(3), 215-219.

Appendix A

Table A.1. Distribution of the population in three metropolitan areas (persons and %)

Areas	Gender	Age				Sum
		18-29	30-39	40-49	50-59	
Taipei	Male	191,393	210,227	194,132	194,969	790,721
	%	50.64	46.71	46.24	46.69	47.48
	Female	186,561	239,858	225,745	222,642	874,806
	%	49.36	53.29	53.76	53.31	52.52
	Sum	377,954	450,085	419,877	417,611	1,665,527
	%	100.00	100.00	100.00	100.00	100.00
Taichung	Male	246,628	229,136	205,060	189,523	870,347
	%	51.49	49.18	47.93	48.36	49.32
	Female	232,391	236,813	222,751	202,406	894,361
	%	48.51	50.82	52.07	51.64	50.68
	Sum	479,019	465,949	427,811	391,929	1,764,708
	%	100.00	100.00	100.00	100.00	100.00
Kaohsiung	Male	234,715	236,564	221,173	209,606	902,058
	%	51.68	49.73	49.79	48.61	49.97
	Female	219,473	239,115	223,074	221,612	903,274
	%	48.32	50.27	50.21	51.39	50.03
	Sum	454,188	475,679	444,247	431,218	1,805,332
	%	100.00	100.00	100.00	100.00	100.00
Total	Male	672,736	675,927	620,365	594,098	2,563,126
	%	51.31	48.57	48.02	47.88	48.96
	Female	638,425	715,786	671,570	646,660	2,672,441
	%	48.69	51.43	51.98	52.12	51.04
	Sum	1,311,161	1,391,713	1,291,935	1,240,758	5,235,567
	%	100.00	100.00	100.00	100.00	100.00

Source: Department of Household Registration Affairs, Ministry of Interior.

Table A.2. Sampling distribution(persons and %)

Areas	Gender	Age				Sum
		18-29	30-39	40-49	50-59	
Taipei	Male	18	20	19	20	77
	%	50.00	45.45	46.34	47.62	47.24
	Female	18	24	22	22	86
	%	50.00	54.54	53.66	52.38	52.76
	Sum	36	44	41	42	163
	%	100.00	100.00	100.00	100.00	100.00
Taichung	Male	24	23	20	18	85
	%	52.17	48.94	48.78	48.65	49.71
	Female	22	24	21	19	86
	%	47.83	51.06	51.22	51.35	50.29
	Sum	46	47	41	37	171
	%	100.00	100.00	100.00	100.00	100.00
Kaohsiung	Male	23	23	21	22	89
	%	50.00	50.00	50.00	52.38	50.57
	Female	23	23	21	20	87
	%	50.00	50.00	50.00	47.62	49.43
	Sum	46	46	42	42	176
	%	100.00	100.00	100.00	100.00	100.00
Total	Male	65	66	60	60	251
	%	50.78	48.18	48.39	49.59	49.22
	Female	63	71	64	61	259
	%	49.22	51.82	51.61	50.41	50.78
	Sum	128	137	124	121	510
	%	100.00	100.00	100.00	100.00	100.00

Source: Sampling distribution of survey data utilized in this study.

Table A.3. Demographical distribution of samples(n=510)

		Persons	%
Gender	Male	251	49.22
	Female	259	50.78
Marital status	Married	297	58.22
	Unmarried	209	40.78
	Other	4	1.00
Age	18-29	128	25.10
	30-39	137	26.86
	40-49	124	24.31
	50-59	121	23.73
Education	Elementary	7	1.37
	Junior high school	8	1.57
	Senior high school	106	20.78
	College/University	269	52.75
	Graduate school	120	23.53
Occupation	Military/Public sector/Education	55	10.78
	Agriculture	3	0.59
	Manufacturing	67	13.14
	Business	145	28.43
	Housewife	47	9.22
	Student	73	14.31
	Other	120	23.53
Personal monthly income (NTD)			36,625
Household monthly income (NTD)			93,701

Source: Descriptive statistics of survey data utilized in this study.

消費者調查缺失資料修補所形成之分配形變*

魯眞**、李秀瑩***

缺失資料會造成統計推估及解釋結果時的偏差。本研究以六種不同方式測試修補缺失資料時偏態及峰態的改變，結果顯示當以完整資料的平均數或中位數修補缺失資料時，資料的分配會因此而產生形變。Two-sample Kolmogorov-Smirnov tests 檢定結果顯示在修補缺失資料時，以逐步迴歸（Stepwise regression）、複迴歸（Multiple regression）、及集群分析法（Clustering method）的期望值進行資料修補時，相對較可維持資料分配的特性。

關鍵詞：偏態及峰態、缺失資料修補、消費者調查

* 本研究承科技部計畫補助（102-2410-H-005-030-），謹致謝忱。匿名審查人對本研究提供了非常寶貴的修正建議，特此致謝。

** 中興大學行銷學系教授兼系主任。本文通訊作者。

*** 中興大學研究發展處校務發展中心行政組員，曾為中興大學 EMBA 學生。