Applied Mathematical Sciences, Vol. 3, 2009, no. 35, 1715 - 1726

New Method to Estimate Missing Data by Using the Asymmetrical Winsorized Mean in a Time Series

Ahmad Mahir R. and A. M. H. Al-Khazaleh[1]

School of Mathematical Sciences
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM Selangor D. E., Malaysia

Abstract

In this paper we consider the problem of missing data in a time series analysis. We propose asymmetrical $r \neq s$ winsorized mean to handle the problem of missing data. Beside that we suggested the Neyman allocation method to choose the values of r and s in asymmetric winsorized mean. We used the a absolute mean error and mean square error to compare the result of estimation missing data with other methods, such as trends, average of the whole data, naive forecast and average bound of the holes and simultaneous filling in the missing data. An example had been presented.

Keywords: Missing Data, Winsorized mean, Neyman Allocation

1 Introduction

Stratified sampling is a methodology in which the elements of a heterogeneous population are classified into mutually exclusive and exhaustive subgroups (strata) based on one or more important characteristics [2, 3, 4]; One of the main objectives of stratified sampling is to reduce the variance of the estimator and to get more statistical precision than with the simple random sampling [3].The determination of subgroups (stratum) boundaries using Neyman allocation of sample element among strata. The missing values problem is an old one for analysis tasks. The waste of the data which can result from casewise

[1]ahmed_2005kh@yahoo.com

deletion of missing values, obliges to propose alternatives approaches. A current paper is to try substitute these values by using asymmetric winsorized mean. A problem frequently encounters in data collection is missing observations or observations may be virtually impossible to obtain, either because of time or cost constrains. In order to replace those observations, there are several different options available to the researchers. Firstly, replace with the mean of the series. Secondly replace with the naive forecast. Also replace with a simple trend forecast. Finally replace with an average of the last two known observations that bound the missing observations. This paper is organized as follow: in the second section the determination of the subgroups (stratum) boundaries by using Neyman allocation. In the third section we give a brief of description of the winsorized mean and how to determine (r,s) in the asymmetrical winsorized mean and discuses some primary properties.

2 Neyman allocation

We used the Neyman allocation to allocate the sample size among (subgroups) strata. when the population is highly skewed the Neyman allocation should be used [5].

The suffix h denotes the stratum and i the unit within the stratum. The following notation will be used throughout the paper.

N_h Total number of units

n_h Number of units in sample

x_{hi} Value obtained for the ith unit

$W_h = \frac{N_h}{N}$ Stratum weight

$\overline{x}_h = \frac{\sum_{i=0}^{N_h} X_{hi}}{N_{hi}}$ True mean

$\overline{x}_h = \frac{\sum_{i=0}^{n_h} x_{hi}}{n_{hi}}$ sample mean

when a population of N units are being stratified into L strata and the samples from each stratum are selected with simple random sampling an unbiased estimate of population mean for the estimation variable x, is given by

$$\bar{x}_{st} = \sum_{h=0}^{l} W_h \overline{\mu}_h \tag{1}$$

where in the $hth(h = 1, 2, ..., L)$ stratum $\bar{x}_h$ is the simple mean based the sample of on $n_h(\sum_{h=1}^{L} n_h = n)$ units and W_h is the proportion of population units falling in that stratum. If the finite population corrections are neglected in each stratum, the estimate has the variance

$$V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \sigma_{hy}^2 \Big/ n_h \tag{2}$$

where σ_{hx}^2 is the variance of x in the hth stratum. For the Neyman allocation method the number of units in sample as in equation 3.

$$n_h = n \frac{W_h \sigma_h}{\sum_{h=1}^{L} W_h \sigma_h} \quad h = 1, 2, ..., L \tag{3}$$

The variance depend on Neyman allocation is reduced to

$$V_{Ney}(\bar{x}_{st}) = \frac{1}{n} (\sum_{h=1}^{L} W_h \sigma_h)^2 \tag{4}$$

Let x_0, x_L be the smallest and largest values of x in the population and by differentiating with respect to the stratum boundaries. We get the minimum variance $V_{Ney}(\bar{x}_{st})$ with respect to x_h since x_h appears in the sum only in the terms $W_h \sigma_h$ and $W_{h+1} \sigma_{h+1}$.
Hence we have the formula for finding the optimum stratum.

$$\frac{(x_h - \mu_h)^2 + \sigma_h^2}{\sigma_h} = \frac{(x_h - \mu_{h+1})^2 + \sigma_{h+1}^2}{\sigma_{h+1}} \tag{5}$$

where $h = 1, 2, 3, ..., L - 1$

Let the distribution of x be continuous with the density function, $f(x)$,$a < x < b$ in order to make L (subgroups) strata, the range of x is to be cut up at points $x_1 < x_2 < x_3 < ... < x_{L-1}$. The relative frequency W_h estimated mean μ_h and the variance of the estimated mean σ_h^2 of the hth stratum are given by

$$W_h = \int_{x_{h-1}}^{x_h} f(x)dx \tag{6}$$

$$\mu_h = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x f(x)dx \tag{7}$$

$$\sigma_h^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x^2 f(x)dx - \mu_h^2 \qquad (8)$$

These equations are difficult to solve, since μ_h and σ_h depend on x_h. We must use the iterative method to solve them by using computer program C++. We divide the whole data into subgroups by putting (x_0, x_1) group one, (x_1, x_2) group two and so on.

3 Some primary properties of Winsorized mean

Winsorized mean is a winsorized statistical measurement of central tendency, like the mean and median and even more similar to the truncated mean [6, 9]. It involves the calculation of the mean after replacing given parts of a probability distribution or sample at the high and low end with the most extreme remaining values. The winsorized mean illustrated as the following

$$\bar{x}_w = \frac{1}{n}((r+1)x_{r+1} + \sum_{i=r+2}^{n-s-1} x_i + (s+1)x_{n-t}) \qquad (9)$$

The sum of square deviation defined by

$$s_w^2 = \frac{1}{n}((r+1)(x_{r+1} - \bar{x}_w)^2 + \sum_{i=r+2}^{n-s-1} (x_i - \bar{x}_w)^2 + (s+1)(x_{n-s} - \bar{x}_w)^2) \qquad (10)$$

where $\bar{x}_w$ is the wisorized mean.
And the $t-$statistic is defined by

$$t_w = \frac{\bar{x}_w - \mu_0}{SE(\bar{x}_w)} \qquad (11)$$

where μ_0 the location under the null hypothesis and the standard error of this method is

$$SE_w = \frac{(n-1)S_w}{(n-r-s-1)(\sqrt{n(n-1)}} \qquad (12)$$

When the data from the symmetric distribution, the distribution of the t_w is approximately from the student distribution with n-r-s-1 degree of freedom. The confidence interval $100(1-\frac{\alpha}{2})$ can be calculated for the location parameter has upper and lower limit by

$$\bar{x}_w \pm t_{1-\frac{\alpha}{2}, n-r-s-1} SE(\bar{x}_w) \qquad (13)$$

4 Selection the (r,s) in the asymmetrical winsorized mean

The next step after the determination of the boundaries of the strata (subgroups),we choose the appropriate number of groups with respect to the amount of the whole data,then we account the number of observation in the first group and in the last group. The amount of observation in the first group is r and the amount of observation in the last group is s.

5 Description of the data set

The rainfall data was collected from the 1^{st} January 1969 to 31^{st} December 1997 in the Perak station in Malaysia. In this research the data on rainfall amount were collected and recorded daily. The monthly average was calculated by finding the sum of all amount of rainfall in that particular month and divide it by the number of days in that month for each year.

6 Methodology to determine the boundaries in Exponential distribution

Let a time series of n observation $x_1, x_2, ..., x_n$ are independent identical distributed exponential random variable and the Exponential distribution defined as the following

$$f(x) = \lambda e^{-\lambda x}, \quad \lambda > 0 \qquad , x > 0 \tag{14}$$

where λ is the parameter of Exponential function
Let the distribution of x be continuous with the density function, $f(x)$, $0 < x < \infty$ in order to make L(subgroups) strata, we truncated the domain$0 < x < \infty$ to b where b is the largest value of x that is the domain become $0 < x < b$. The relative frequency W_h of the hth stratum are given by for the exponential distribution [1]

$$W_h = \int_{x_{h-1}}^{x_h} f(x)dx = \int_{x_{h-1}}^{x_h} \lambda e^{-\lambda x} dx \tag{15}$$

$$W_h = e^{-\lambda x_{h-1}} - e^{-\lambda x_h} \tag{16}$$

We got the mean μ_h by using the following formula

$$\mu_h = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x f(x) dx = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x \lambda e^{-\lambda x} dx \qquad (17)$$

$$\mu_h = \frac{e^{-\lambda x_{h-1}}(x_{h-1} + \frac{1}{\lambda}) - e^{-\lambda x_h}(x_h + \frac{1}{\lambda})}{e^{-\lambda x_{h-1}} - e^{-\lambda x_h}} \qquad (18)$$

In addition, we got the mean variance as the following

$$\sigma_h^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x^2 f(x) dx - \mu_h^2 = \frac{\lambda}{W_h} \int_{x_{h-1}}^{x_h} x^{\lambda+1} dx - \mu_h^2 \qquad (19)$$

$$\sigma_h^2 = \frac{e^{-\lambda x_{h-1}}(x_{h-1}^2 + \frac{2}{\lambda} x_{h-1} + \frac{2}{\lambda^2}) - e^{-\lambda x_h}(x_h^2 + \frac{2}{\lambda} x_h + \frac{2}{\lambda^2})}{e^{-\lambda x_{h-1}} - e^{-\lambda x_h}} - (\mu_h)^2 \qquad (20)$$

Since the data order statistics from smallest to the largest it follows the exponential distribution the smallest value is zero and the largest value is 18.5. So we truncated the domain of the exponential distribution at the largest value. The number of subgroups or strata depends on the size of the population.
If we want to divide the whole data into three groups, we need to calculate two boundaries x_1 and x_2 since $x_0 = 0$ and $x_3 = 18.5$.

Since the boundary x_1 depend on the μ_1 and μ_2 and σ_1^2 And σ_2^2, also x_2 depend on μ_2 and μ_3 and σ_2^2 and σ_3^2, we calculated these values as the following respectively.

$$\mu_1 = \frac{(\frac{1}{\lambda}) - e^{-\lambda x_1}(x_1 + \frac{1}{\lambda})}{1 - e^{-\lambda x_1}}$$

$$\mu_2 = \frac{e^{-\lambda x_1}(x_1 + \frac{1}{\lambda}) - e^{-\lambda x_2}(x_2 + \frac{1}{\lambda})}{e^{-\lambda x_1} - e^{-\lambda x_2}}$$

$$\mu_3 = \frac{e^{-\lambda x_2}(x_2 + \frac{1}{\lambda}) - e^{-\lambda(18.5)}(18.5 + \frac{1}{\lambda})}{e^{-\lambda x_2} - e^{-\lambda(18.5)}}$$

$$\sigma_1^2 = \frac{(\frac{2}{\lambda^2}) - e^{-\lambda x_1}(x_1^2 + \frac{2}{\lambda} x_1 + \frac{2}{\lambda^2})}{1 - e^{-\lambda x_1}} - (\mu_1)^2$$

$$\sigma_2^2 = \frac{e^{-\lambda x_1}(x_1^2 + \frac{2}{\lambda} x_1 + \frac{2}{\lambda^2}) - e^{-\lambda x_2}(x_2^2 + \frac{2}{\lambda} x_2 + \frac{2}{\lambda^2})}{e^{-\lambda x_1} - e^{-\lambda x_2}} - (\mu_2)^2$$

$$\sigma_3^2 = \frac{e^{-\lambda x_2}(x_2^2 + \frac{2}{\lambda} x_2 + \frac{2}{\lambda^2}) - e^{-\lambda(18.5) x_1}((18.5)^2 + \frac{2}{\lambda}(18.5) + \frac{2}{\lambda^2})}{e^{-\lambda x_2} - e^{-\lambda(18.5)}} - (\mu_3)^2$$

If we want to divide the whole data into four groups, we need to calculate three boundaries x_1 , x_2 and x_3 since $x_0 = 0$ and $x_4 = 18.5$ Since the boundary x_1 depend on the μ_1 and μ_2 and σ_1^2 And σ_2^2, x_2 depend on μ_2 and μ_3 and σ_2^2 and σ_3^2 and x_3 depend on μ_3 and μ_4 and σ_3^2 and σ_4^2 we calculated these values as the following:

$$\mu_1 = \frac{(\frac{1}{\lambda}) - e^{-\lambda x_1}(x_1 + \frac{1}{\lambda})}{1 - e^{-\lambda x_1}}$$

$$\mu_2 = \frac{e^{-\lambda x_1}(x_1 + \frac{1}{\lambda}) - e^{-\lambda x_2}(x_2 + \frac{1}{\lambda})}{e^{-\lambda x_1} - e^{-\lambda x_2}}$$

$$\mu_3 = \frac{e^{-\lambda x_2}(x_2 + \frac{1}{\lambda}) - e^{-\lambda x_3}(x_3 + \frac{1}{\lambda})}{e^{-\lambda x_2} - e^{-\lambda x_3}}$$

$$\mu_4 = \frac{e^{-\lambda x_3}(x_3 + \frac{1}{\lambda}) - e^{-\lambda(18.5)}(18.5 + \frac{1}{\lambda})}{e^{-\lambda x_3} - e^{-\lambda(18.5)}}$$

$$\sigma_1^2 = \frac{(\frac{2}{\lambda^2}) - e^{-\lambda x_1}(x_1^2 + \frac{2}{\lambda}x_1 + \frac{2}{\lambda^2})}{1 - e^{-\lambda x_1}} - (\mu_1)^2$$

$$\sigma_2^2 = \frac{e^{-\lambda x_1}(x_1^2 + \frac{2}{\lambda}x_1 + \frac{2}{\lambda^2}) - e^{-\lambda x_2}(x_2^2 + \frac{2}{\lambda}x_2 + \frac{2}{\lambda^2})}{e^{-\lambda x_1} - e^{-\lambda x_2}} - (\mu_2)^2$$

$$\sigma_3^2 = \frac{e^{-\lambda x_2}(x_2^2 + \frac{2}{\lambda}x_2 + \frac{2}{\lambda^2}) - e^{-\lambda x_3 x_1}(x_3^2 + \frac{2}{\lambda}x_3 + \frac{2}{\lambda^2})}{e^{-\lambda x_2} - e^{-\lambda x_3}} - (\mu_3)^2$$

$$\sigma_4^2 = \frac{e^{-\lambda x_3}(x_3^2 + \frac{2}{\lambda}x_3 + \frac{2}{\lambda^2}) - e^{-\lambda(18.5)x_1}((18.5)^2 + \frac{2}{\lambda}(18.5) + \frac{2}{\lambda^2})}{e^{-\lambda x_2} - e^{-\lambda(18.5)}} - (\mu_4)^2$$

We found the boundaries of the strata (subgroups) for the three, four, five, six, seven and eight subgroups in the same way by using the C++ program. For example if we divided the whole data into three groups the first group contain all the observation between 0 and 3.2668 ,the second group all observation between 3.2668 and 8.17 and the third group all the observations between 8.17 and 18.5. The values of the boundaries group are obtain in the table 1.

7 Summary and Discussion

To determine the number of observations in the groups, we count the number of observation in each group. For example if we divide the whole data into three groups with 10% missing data,the number of observations in the first group interval[0,3.266796] is 81, the number of observation in the second group interval [3.266796,8.17] is 144 and the number of observation in the third group interval [8.17,18.5] is 67. Hence, we can calculate the value of winsorized mean in this case by putting $r = 81$ and $s = 67$ which is 5.436461 by using Equation 9. In addition the average for the data with 10% is 5.778137. Similarly we can find the winsorized mean for a different datasets. The number of the observation in all groups , the values of the wisorized mean and the average for 10%,15%, 20% and 25% missing data are shown in tables 2, 3, 4, 5, 6, 7 and 8.

We compared the result of estimation missing data in winsorized method with the following estimation missing data methods [6, 7]: Firstly, replace with the mean of the series. This mean can be calculated over the entire range of the sample. Secondly, replace with the naive forecast. Naive model is the simplest form of a Univariate forecast model, this model uses the current time period's value for the next time period, that is $\hat{Y}_{t+1} = Y_t$. Also, replace with a simple trend forecast. This is accomplished by estimating the regression equation of the form $Y_t = a + b$t (where t is the time) for the periods prior to the missing value. Then use the equation to fit the time periods missing. Finally, replace with an average of the last two known observations that bound the missing observations.

The accuracy of estimating the missing data with a winsorized mean depends on how close the estimating values to the actual values. In practice, we define the difference between the actual and the estimating values as an error. If the estimating is doing a good job the error will be relatively small. This means that the error for each time period is purely random fluctuation around original value. So we should get a value equal to or near 0. We tested winsorized mean approach to estimating missing data points on time series data sets with respect to the other methods.

In this paper the following estimator of errors will be used as a measure of accuracy:

1- The mean absolute errors

$$MAE = \frac{\sum_{t=1}^{n} |e_t|}{n} \tag{21}$$

2- The mean square errors

$$MSE = \frac{\sum_{t=1}^{n} e_t^2}{n} \tag{22}$$

To evaluate the accuracy of the estimation of the missing data, we used mean absolute error (MAE) Eq.(21) and mean square error (MSE) Eq.(22).

Tables (8)and (9) show the amount of the MAE and MSE respectively, for the estimation missing data by using all methods, its clear that the amount of the MAE and MSE for the winsorized method for the eight, seven, six, five, four and three groups are less than from the estimation missing data methods: Trend, Average of the entire data, Naive model and the average of bounded the missing values for all percentage 10%, 15%, 20% and 25% missing data.

It has been shown that asymmetrical wisorized mean can be used to effectively estimate missing values in a time series. In fact these estimates were consistently better than the other methods.

References

[1] AL-khazaleh A. M. H. (2007). *Using Taylor Expansion in Finding the Optimum Stratum Boundaries Depending on Neyman Allocation.*, Master Thesis, Al-albayt University, Jordan.

[2] Cochran,W.G. (1977). *Sampling Techniques.*, third ed. Wiley, USA.

[3] Cyert, R.M., Davidson, H.J., (1962). *Statistical Sampling for Accounting Information* . Prentice-Hall, Englewood Cliffs, NJ, pp. 116127.

[4] Hess, I., Sethi, V.K. and and Balakrishnan, T.R., (1966). *Stratification: a practical investigation..* J. Amer. Statist. Assoc. 61 (313), 7490.

[5] Mahir, AL-kahazaleh and AL-kassab (2007). *Approximation Method in Finding Optimum Stratum Depending on Neyman Allocation Applied on Beta Distribution.* Applied Mathematics for Science and Engineering. Vol.8.

[6] Mingxin W. and Yijun Z. and (2009). *Trimmed and winsorised means based on a scaled deviation..* J. of stat. planning and inference. 139 ,350365.

[7] Pankratz A. (1983).*Forecasting with Univariate Box - Jenkins Models.* John Wiley Sons, Inc.

[8] Patricia E. G. (1994), *Introduction to Time Series Modeling and Forecasting in Business and Economics.*McGraw-Hill, Inc.

[9] Vic B. and Toby L. (1994), *Outliers in Statistical Data.* third ed. Wiley, USA.

Received: November, 2008

Table 1: Boundaries stratum for the groups from three until eight

B. G	X0	X1	X2	X3	X4	X5	X6	X7	X8
3groups	0	3.2668	8.17	18.5					
4groups	0	2.3528	5.439	9.9396	18.5				
5groups	0	1.839	4.0955	7.0168	11.172	18.5			
6groups	0	1.50966	3.2889	5.4556	8.227898	12.0848	18.5		
7groups	0	1.2803	2.7493	4.472	6.5556	9.1925	12.79	18.5 8	
8groups	0	1.1115	2.3623	3.7925	5.46215	7.4682	9.9821	13.3531	18.5

Table 2: the number of observations for eight groups , winsorized mean and average of whole data with respect to the 10%, 15% 20% and 25% missing data

Missing data	G1	G2	G3	G4	G5	G6	G7	G8	W mean	μ
10%	22	28	50	60	51	37	30	14	5.733483	5.77814
15%	21	23	46	59	50	35	27	14	5.754055	5.801354
20%	16	25	48	55	41	29	32	13	5.80288	5.860024
25%	16	22	45	55	39	25	27	13	5.751242	5.797742

Table 3: the number of observation in seven groups , winsorized mean and average of whole data with respect to the 10%, 15% 20% and 25% missing data

Missing data	G1	G2	G3	G4	G5	G6	G7	W mean	μ
10%	27	32	66	72	43	34	18	5.717728	5.778137
15%	25	30	63	66	42	31	18	5.730934	5.801354
20%	21	31	62	58	36	34	17	5.780761	5.860024
25%	19	27	62	55	31	31	17	5.72697	5.797742

Table 4: the number of observation in six groups , winsorized mean and average of whole data with respect to the 10%, 15% 20% and 25% missing data

Missing data	G1	G2	G3	G4	G5	G6	W mean	μ
10%	30	51	77	69	42	23	5.687767	5.778137
15%	28	44	77	64	40	22	5.708545	5.801354
20%	26	48	70	55	38	22	5.75166	5.860024
25%	22	45	71	51	32	21	5.695153	5.797742

Table 5: the number of observation in five groups , winsorized mean and average of whole data with respect to the 10%, 15% 20% and 25% missing data

Missing data	G1	G2	G3	G4	G5	W mean	μ
10%	41	70	95	54	32	5.647049	5.778137
15%	36	70	87	53	29	5.660979	5.801354
20%	33	68	80	47	31	5.691803	5.860024
25%	28	70	72	43	29	5.630011	5.797742

Table 6: the number of observation in four groups , winsorized mean and average of whole data with respect of the 10%, 15% 20% and 25% missing data

Missing data	G1	G2	G3	G4	W mean	μ
10%	50	108	90	44	5.529441	5.778137
15%	44	74	116	41	5.55308	5.801354
20%	41	73	100	45	5.551114	5.860024
25%	38	100	64	40	5.492704	5.797742

Table 7: the number of observation in three groups , winsorized mean and average of whole data with respect to the 10%, 15% 20% and 25% missing data

Missing data	G1	G2	G3	W mean	μ
10%	81	144	67	5.436461	5.778137
15%	72	139	64	5.447113	5.801354
20%	75	133	51	5.654724	5.860024
25%	67	120	55	5.37043	5.797742

Table 8: The amount of absolute mean error for all method and suggested method(winsorized mean to a different groups)

MAE	trend	average	bound	naive	$x^w 8$	$x^w 7$	$x^w 6$	$x^w 5$	$x^w 4$	$x^w 3$
10%missing	2.499	1.971	2.448	2.692	1.862	1.868	1.872	1.882	1.888	1.926
15%missing	3.843	2.516	2.792	3.204	2.499	2.491	2.483	2.467	2.430	2.398
20%missing	5.144	2.411	3.099	3.959	2.384	2.373	2.360	2.332	2.270	2.315
25%missing	4.734	2.028	2.522	3.383	2.021	2.017	2.012	2.004	1.990	1.980

Table 9: The amount of mean square error for all method and suggested method(winsorized mean for a different groups)

MAE	trend	average	bound	naive	$x^w 8$	$x^w 7$	$x^w 6$	$x^w 5$	$x^w 4$	$x^w 3$
10%missing	8.890	5.308	10.634	5.672	5.215	5.247	5.265	5.317	5.347	5.533
15%missing	23.730	9.250	12.781	17.595	9.163	9.122	9.082	9.003	8.841	8.703
20%missing	39.961	8.819	14.934	24.610	8.662	8.602	8.526	8.374	8.045	8.283
25%missing	38.652	6.809	11.597	19.927	6.752	6.724	6.688	6.623	6.512	6.445