Review of Economics & Finance
Submitted on 18/04/2018
Article ID: 1923-7529-2018-04-17-08
David A. Swanson, Jeff Tayman, and T.M. Bryan

A Note on Rescaling the Arithmetic Mean for Right-skewed Positive Distributions

David A. Swanson
Department of Sociology, University of California Riverside
900 University Ave, Riverside, CA 92521, U.S.A.
E-mail: dswanson@ucr.edu

Jeff Tayman
Department of Economics
University of California San Diego
9500 Gilman Dr, La Jolla, CA 92093, U.S.A.
E-mail: jtayman@ucsd.edu

T.M. Bryan
McKibben Demographic Research
PO Box 2921, Rock Hill, South Carolina 29732, U.S.A.
E-mail: j.mckibben@mckibbendemographics.com

***A*bstract:** When the arithmetic mean (mean) is used as a measure of location for a set of right-skewed positive observations, it is subject to being pulled upward. This upward movement tends to move the mean away from the bulk of the observations, making it less representative of them. One way to deal with this loss of representativeness is to transform the data. A Box-Cox power transformation can make a right-skewed distribution more symmetrical and then a measure of location for the original observations is found by applying an inverse transformation to the center of the transformed data. This approach was used in a series of papers dealing with the Mean Absolute Percent Error (MAPE) as a measure of forecast and estimation error. In this paper, we show that the Box-Cox power transformation can be used more generally with any mean computed for a set of right-skewed positive observations to develop R-MEAN (Rescaled-Mean). We provide a set of examples to illustrate this approach and show its use in an actual application.

Keywords: Asymmetric distribution; Box-Cox Power Transformation; Outlier; R-MEAN

JEL Classification: B41, C13, C18

1. Background

Any summary measure of location should location should meet five highly desirable criteria for summary measures of error: (1) validity, (2) reliability, (3) ease of interpretation, (4) clarity of presentation, and (5) support of statistical evaluation (National Research Council 1980). The mean, median, and mode can all meet these criteria, but under certain conditions the mean does not meet the validity criterion. The mean fails to meet a normative standard of validity if the data are highly skewed because it may be pulled away from the bulk of the observations. Because the mean is neither a resistant or robust summary measure, a few outliers can dominate it (Hoaglin, Mosteller, and Tukey 1983: 28; Huber, 1964; Huber, 1981; Tukey 1970). One way to deal with this loss of representativeness is to transform the data. Taylor (1985) shows that a Box-Cox power transformation can make a right-skewed distribution more symmetrical and a valid measure of location for the original observations can be found by applying an inverse transformation to the mean of the transformed data. This approach was used in a series of papers dealing with the Mean Absolute Percent Error (MAPE) as a measure of forecast and estimation error (Coleman and Swanson 2007; Swanson, Tayman and Bryan 2011; Swanson, Tayman and Barr 1999, 2000). However, it has not been applied to a general form of the mean when it is used for a set of right-skewed positive observations. To rectify this shortcoming, we show in this paper that the Box-Cox power transformation can be used more generally with any mean computed for a set of right-skewed positive observations.

Our approach involves rescaling the mean to a measure we term "R-MEAN" (Rescaled-Mean) using a Box-Cox power transformation. Rescaling is designed to address the impact of outlying observations, while still preserving the valuable statistical properties of the mean. Our approach using a transformation is not the only way to deal with outliers. One could use a trimmed mean, a Winsorized mean, an M-estimator, or the median. These measures provide summary measures of location designed to represent the bulk of the observations and for which there are various trade-offs relative to using transformations (Hoaglin, Mosteller and Tukey 1983: 28; Huber 1964, 1981; Taylor 1985; Tukey 1970; Wilcox 2012).

2. R-MEAN

To change the shape of a distribution efficiently and objectively and to achieve parity for the observations, Swanson, Tayman, and Barr (2000) use a standardized technique designed to generate a single, nonlinear function. This technique modifies the power transformation developed by Box and Cox (1964)[1], defined as:

$$y(\lambda) = \left(x^{\lambda} - \lambda\right)/\lambda, \text{ when } \lambda \neq 0 \qquad (1)$$

or

$$y(\lambda) = \ln(x), \text{ when } \lambda = 0 \qquad (2)$$

where x is the original observation, y is the transformed observation, and λ is the power transformation constant.

[1] Swanson, Tayman, and Barr (2000) used λ in the numerator. Box and Cox (1964) used 1.0 in their original development to assure continuity in λ when λ=0. The difference is immaterial.

One determines Lambda (λ) by finding the λ value that maximizes the function:

$$ml(\lambda) = -(n/2)\times \ln\left[(1/n)\sum(y_i - \bar{y})^2\right] + (\lambda - 1)\times \sum \ln(x_i) \qquad (3)$$

where n is the sample size; y is the transformed observation; $\bar{y}$ is the mean of the transformed observations; and x is the original observation.

According to Box and Cox (1964), $ml(\lambda)$ at a local maximum provides the power transformation (λ) for x that optimizes the *probability* that the transformed distribution will be symmetrical. In other words, finding λ does not guarantee symmetry, but it represents the transformation power most likely to yield a symmetrical distribution. We can find the maximum value of ml(λ) by solving its function for different values of λ between the range of –2 and 2 and identifying the largest resulting Box-Cox value (Draper and Smith, 1981: 225).

To address the effect of a skewed distribution on the mean, we transform the original distribution using a Box-Cox power transformation to create a measure we term "MEAN-T (Mean Transformed) as the mean for this transformed distribution. The transformed distribution considers the entire data series, but assigns a proportionate amount of influence to each case through normalization, thereby reducing the otherwise disproportionate effect of outliers on a summary measure of error. Transformation, however, may move the observations into a unit of measurement that is difficult to interpret, so it is desirable to have a simple procedure for re-expressing MEAN-T back into the original scale of mean. To do this, we use the inverse approach developed and tested by Coleman and Swanson (2007)[2, 3] :

$$\text{R-MEAN} = [(\lambda)(\text{MEAN-T} + 1)]^{1/\lambda} .$$

3. Is R-MEAN Needed?

Swanson, Tayman, and Barr (2000) suggested using a statistical skewness test developed by D'Agostino, Belanger, and D'Agostino Jr. (1990) to make a determination in regard to transformation of positive distribution. The null hypothesis tested is that the skewness value = 0, using the 0.10 level of significance. We recommend this significance level rather than more stringent ones (e.g., 0.05 and 0.01) because there is a greater cost in terms of a downwardly biased measure of accuracy in not transforming a potentially skewed distribution. When the skewness test indicates a potentially useful transformation of a skewed distribution to a symmetrical distribution, the transformation is assumed to be successful when the average of the new distribution is representative of the bulk of the observations and uses all of them. In this situation, the observations

[2] A potential shortcoming of the Box-Cox transformation is it is not globally monotonic. Individual values may have differential influence on the function. Values near the mean of the transformed distribution have little effect, while extreme outliers may actually *reduce* the MAPE-T. Because the Box-Cox transformation has no associated influence function, it is difficult to determine if and when the Box-Cox will perform this way (Coleman and Swanson 2007).

[3] Coleman and Swanson (2007) find this closed-form expression to be a member of the family of power mean-based accuracy measures. This enables it to be placed in relation to other members of this family, which includes the Harmonic Mean and the Geometric Mean. It also can serve as an estimator of the median.

receive nearly equal weights, closer to *1/n*, while the resulting mean remains intuitively interpretable and clear in its presentation.

4. Examples

Our examples are based on "Anscombe's Quartet" (Anscombe 1973), which is comprised of four sets of positive data (all with n = 11) used to show the importance of graphing when considering a regression analysis. We have taken the original dependent variable observations from each of the four sets and added 14 observations to each of the then in order to create data with a wider range of right skewness compared to Anscombe's original data to examine of efficacy of R-MEAN (see Table 1).

As can be seen in Table 1, the skewness for data set I is 0.697, while for sets II, III, and IV, the skewness is 0.050, 1.543, and 4.680, respectively. Using The NCSS statistical package (release 8) and a critical p-value of 0.10, the D'Agostino tests for skewness found that the assumption of normality could not be rejected for Data Sets I (p =0.127) and II (p = 0.909). However, it is rejected for Data Sets III (p = 0.023) and IV (p < 0.001). These results indicate that the Box-Cox power transformation is needed for Data Sets III and IV, but not for Data Sets I and II. As such, we compute R-MEAN for Data Sets III and IV. For heuristic purposes we also compute R-MEAN for Data Set II, where the D'Agostino test indicates it is not needed.

Table 1. Anscombe's quartet dependent variables[a]

Observation	Data Set			
	I	II	III	IV
1	8.04	9.14	7.46	6.58
2	6.95	8.14	6.77	5.76
3	7.58	8.74	12.74	7.71
4	8.81	8.77	7.11	8.84
5	8.33	9.26	7.81	8.47
6	9.96	8.10	8.84	7.04
7	7.24	6.13	6.08	5.25
8	4.26	3.10	5.39	12.50
9	10.84	9.13	8.15	5.56
10	4.82	7.26	6.42	7.91
11	5.68	4.74	5.73	6.89
12	5.20	5.20	5.20	5.20
13	5.90	5.90	5.90	5.90
14	8.00	8.00	8.00	8.00
15	6.92	6.92	6.92	6.92
16	3.20	3.20	3.20	3.20
17	5.62	5.62	5.62	5.62
18	6.01	6.01	6.01	6.01
19	6.05	6.05	6.05	6.05
20	5.98	5.98	5.98	5.98
21	4.97	4.97	4.97	4.97
22	5.23	5.23	5.23	5.23
23	5.84	5.84	5.84	5.84
24	5.00	5.01	9.20	15.00
25	4.60	4.80	13.40	65.00
Mean	6.441	6.450	6.961	9.257
Median	5.980	6.010	6.080	6.050
Skewness	0.697	0.050	1.543	4.680
D' Agostino p-value	0.127	0.909	0.023	<0.001
R-Mean	n/a	6.397	6.581	6.550

[a] Observations 1 through 11 from Anscombe (1973); Observations 12 through 25 from the authors.

To generate R-MEAN, we use an excel macro developed by Charles Barr (Swanson, Tayman and Barr 2000) that comes with instructions for use (available from the authors on request). For Data Set III, we find that R-MEAN is equal to 6.581, which is somewhat less than the mean of 6.961 in Data Set III (see Table 1). This result shows the slight upward bias of the mean in a data set that is modestly right-skewed (skew = 1.543).

An even more pronounced difference is found between the mean and R-MEAN in Data Set IV. Here, we find that R-MEAN is equal to 6.550, which is considerably less than the mean of 9.275. These results show a considerable upward bias of the mean in a data set that is substantially right-skewed (skew = 4.680).

Data Set II is used to illustrate the results when a transformation is not likely to be needed. Here we find the R-MEAN (6.397) is very close to the mean (6.450),which would be expected given that the null hypothesis of normality was not rejected (p = 0.909) and the right-skewness is virtually zero (skew = 0.050). In this situation, a Box-Cox power transformation is not needed.

5. Application

Our application uses county unemployment rates (not seasonally adjusted) as of October 2017 for Arizona and New Mexico (Bureau of Labor Statistics 2017). As can be seen in Table 2 on the next page, the unemployment rates in Arizona's counties are more right skewed than those in New Mexico with skews of 2.432 and 0.611, respectively. The D'Agostino tests for skewness found that the assumption of normality is rejected for Arizona (p = < 0.001) and the mean (6.287) is considerably larger than the median (4.800). The normality test is not rejected for New Mexico (p = 0.129) and the mean (6.242) and the median (6.100) are close in value.

These results indicate that the R-MEAN transformation is needed for Arizona, but not for New Mexico. As such, we will compute R-MEAN for Arizona's counties, but not for New Mexico. We find that the R-MEAN (5.212) is noticeably less than the mean (6.287) and somewhat higher than the median (4.800).

Again, these results show the efficacy of using the Box-Cox power transformation as a strategy for rescaling the mean in situations with right-skewed positive values.

Table 2. County unemployment rates, Arizona and New Mexico, October 2017[a]

Arizona (15 observations)		New Mexico (33 observations)	
Apache County	9.4	Bernalillo County	5.4
Cochise County	4.8	Catron County	7.0
Coconino County	4.3	Chaves County	6.1
Gila County	5.2	Cibola County	7.7
Graham County	5.0	Colfax County	6.3
Greenlee County	4.7	Curry County	4.9
La Paz County	4.7	Debaca County	4.7
Maricopa County	3.7	Dona Ana County	6.2
Mohave County	5.1	Eddy County	4.9
Navajo County	6.2	Grant County	6.1
Pima County	4.0	Guadalupe County	6.0
Pinal County	4.3	Harding County	7.5
Santa Cruz County	11.0	Hidalgo County	4.8
Yavapai County	3.9	Lea County	6.0
Yuma County	18.0	Lincoln County	5.6
		Los Alamos County	3.7
		Luna County	10.3
		McKinley County	8.5
		Mora County	7.2
		Otero County	6.0
		Quay County	6.1
		Rio Arriba County	6.1
		Roosevelt County	5.0
		San Juan County	6.5
		San Miguel County	7.0
		Sandoval County	6.1
		Santa Fe County	5.1
		Sierra County	7.0
		Socorro County	6.3
		Taos County	7.7
		Torrance County	8.0
		Union County	3.7
		Valencia County	6.5
Mean	6.287		6.242
Median	4.800		6.100
Skewness	2.432		0.611
D' Agostino p-value	<0.001		0.129
R-Mean	5.212		n/a

[a] Data from Bureau of Labor Statistics (2017)

6. Summary and Conclusion

The arithmetic mean has a long history of use and is instrumental in the development of the method of "least squares" (Stigler 1986: 61). As noted by Cobb and Moore (1997: 801) in the field of statistics "… data are not just numbers, they are numbers with a context;" and the context for the arithmetic mean is that it is used to represent certain relationships in the data (Russell and Mokros 1996: 362). For the mean, these "certain relationships" are denoted by the fact that the distance between two numbers is defined to be the square of their difference and the sum of the squared differences between each observation and the mean is smaller than the sum of squares of the differences between each observation and any other number. Although the definition used by the arithmetic mean is different than the definitions of distance used by the median and the mode, it shares with them the idea that it is representative of the observations. This is the normative standard of validity for each of these three measures of location. However, when the arithmetic mean is used as a measure of location for right-skewed positive observations, it is subject to being pulled upward, making it move it away from the bulk of the observations. This makes the arithmetic mean less representative of the observations, which can lead to its failure to meet a normative standard of validity.

In this paper, we use a skewness test to determine if the arithmetic mean of a positive distribution represents the data from which it is computed. If it is not, we offer an alternative measure of central tendency that uses a Box-Cox power transformation to mitigate the impact of outlying observations and compute the mean from that transformed distribution (MEAN-T). Transformation may move the observations into a unit of measurement that is difficult to interpret, so we use an inverse approach for expressing MEAN-T back into the original scale of mean (R-MEAN). While other statistics such as the median, M-estimators, trimmed mean have been developed to mitigate the impact of outlying observation when measuring central tendency, R-MEAN has a number of advantages. It is easy to calculate, is readily understandable, uses all the observations in a distribution, and preserves the important statistical properties of the mean.

We first illustrated this approach using the set of four distributions from Anscombe's Quartet augmented with additional observations to create a wider range of skewness alternatives. We found the skewness test appropriately identified the distributions requiring transformation. In these distributions, the R-MEAN was substantially lower than the means from the untransformed data, indicating the ability of the transformation and re-expression process to create a mean that mitigates the impacts of outlying observations. We also found that the R-MEAN was very close to the mean in the distribution that did not require a transformation as would be expected if the original distribution was symmetrical. Similar results were found when analyzing county unemployment rates for Arizona and New Mexico (2017), with the former having a right-skewed distribution and the latter having a symmetrical distribution.

References

[1] Anscombe F. (1973). "Graphs in statistical analysis", *The American Statistician*, 27(1): 17-22.

[2] Box G., Cox D. (1964). "An analysis of transformations", *Journal of the Royal Statistical Society* (Series B), 26(2): 211-252.

[3] Bureau of Labor Statistics (2017). "Local area unemployment statistics map", [Online] Retrieved from https://data.bls.gov/map/MapToolServlet?survey=la .

[4] Cobb G., Moore D. (1997). "Mathematics, statistics, and teaching", *The American Mathematical Monthly*, 104(9): 801-823.

[5] Coleman C., Swanson D. (2007). "On MAPE-R as a measure of cross-sectional estimation and forecast accuracy", *Journal of Economic and Social Measurement*, 32(4): 219-233.

[6] D'Agostino R., Belanger A., D'Agostino R. Jr. (1990). "A suggestion for using powerful and informative tests of normality", *The American Statistician*, 44(3): 316-321.

[7] Draper N., Smith H. (1981). *Applied regression analysis*, 2nd Edition. New York: Wiley.

[8] Hoaglin D., Mosteller F., Tukey J. (1983). "*Introduction to more refined estimator*", In: Hoaglin D., Mosteller F., Tukey J. (Eds.) *Understanding Robust and Exploratory Data Analysis.* New York: Wiley, Chapter 9:283-296.

[9] Huber P. (1964). "Robust estimation of a location parameter", *Annals of Mathematical Statistics*, 35(1): 73-101.

[10] Huber P. (1981). *Robust statistics*. New York: Wiley.

[11] National Research Council (1980). *Estimating population and income for small places*. National Academy Press, Washington, D.C.

[12] Russell S., Mokros J. (1996). "Research into practice: What do children understand about average? ", *Teaching Children Mathematics*, 2(6): 360-364.

[13] Stigler S. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge: Belknap.

[14] Swanson D., Tayman J., Barr C. (2000). "A note on the measurement of accuracy for subnational demographic estimates", *Demography*, 37(2): 193-201.

[15] Swanson D., Tayman J., Bryan T. (2011). "MAPE-R: A rescaled measure of accuracy for cross-sectional, subnational forecasts", *Journal of Population Research*, 28(2-3): 225-243.

[16] Taylor J. (1985). "Measures of location of skew distributions obtained through Box-Cox transformations", *Journal of the American Statistical Association*, 80(390): 427-432.

[17] Tayman J., Swanson, D., Barr C. (1999). "In search of the ideal measure of accuracy for subnational demographic forecasts", *Population Research and Policy Review*, 18(5): 387-409.

[18] Tukey J. (1970). *Exploratory data analysis Vol. I*, Addison-Wesley, Reading.

[19] Wilcox R. (2012). *Introduction to robust estimation and hypothesis testing*. San Diego: Academic Press.