ED 463 313 TM 033 745

AUTHOR	Osborne, Jason W.
TITLE	The Effects of Minimum Values on Data Transformations.
PUB DATE	2002-04-00
NOTE	7p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Data Analysis; Research Methodology; Statistical Distributions; *Transformations (Mathematics); *Values

ABSTRACT

Data transformations are commonly used tools in quantitative analysis of data. However, data transformations can be a mixed blessing to researchers, improving the quality of the analysis while at the same time making the interpretation of the results difficult. Few, if any, statistical texts discuss the tremendous influence a distribution's minimum value has on the outcome of a transformation. The goal of this paper is to promote thoughtful and informed use of data transformation. The focus is on three common data transformations: square root, logarithmic, and inverse transformations. All three are curvilinear transformations that change the nature of the variable to a certain extent. Examples illustrate the importance of the minimum value of a distribution should the researcher intend to use data transformation on that variable. (Contains 10 references.) (SLD)

The Effects of Minimum Values On Data Transformations

Jason W. Osborne

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in New Orleans, LA
April 2002

Paper presented at the 2002 meeting of the
American Education Research Association, New Orleans, LA

The Effects Of Minimum Values On Data Transformations.

Jason W. Osborne, Ph.D
North Carolina State University

Data transformations are commonly used tools in quantitative analysis of data. However, data transformations can be a mixed blessing to a researcher, improving the quality of the analysis while at the same time making the interpretation of the results difficult. Further, few (if any) statistical texts discuss the tremendous influence a distribution's minimum value has on the outcome of a transformation. The goal of this paper is to promote thoughtful and informed use of data transformations.

Data transformations are the application of a mathematical modification to a variable. There are a great variety of possible data transformations, from adding constants to multiplying, squaring or raising to a power, converting to logarithmic scales, inverting, taking the square root of the values, and even applying sine wave transformations.

There are a variety of reasons why researchers might want to employ data transformations. First, as many statistical procedures assume or benefit from normality of variables, data transformations can be employed to improve the normality of a variable's distribution. Authors of prominent statistical texts, such as Tabachnick and Fidell (2001, p. 81), argue that researchers should "consider transformation of variables in all situations" unless there is a specific reason not to. Other reasons for utilization of data transformations involve equalizing variance (e.g., Bartlett, 1947), although this is less commonly the reason researchers turn to transformation. Our focus here is explicitly on the former reason, although many points will apply to variance equalizing as well.

Data transformation and normality

If a researcher has a variable that is substantially non-normal, even if analyses utilized do not assume normality, improving normality can often enhance the outcome of analyses by reducing error. In fact, Tabachnick and Fidell (2001) explicitly state that, even when normality is not an issue, transformations can improve analyses. Zimmerman (e.g., 1995, 1998) pointed out that non-parametric tests can suffer as much, or more, than parametric tests when normality assumptions are violated, confirming the importance of normality in all statistical analysis, not just parametric analyses.

There are multiple options for dealing with non-normal data. First, the researcher must make certain that the non-normality is due to a valid reason. Invalid reasons include things such as mistakes in data entry, and missing data values not declared missing. These are simple to remedy. Outliers, scores that are extreme relative to the rest of the sample, are another reason for non-normality. There is great debate in the literature about whether outliers should be removed or not. I am sympathetic to Judd and McClelland's (1989) argument that outlier removal is desirable, honest, and important. However, not all researchers feel that way (Orr, Sackett, and DuBois, 1991).

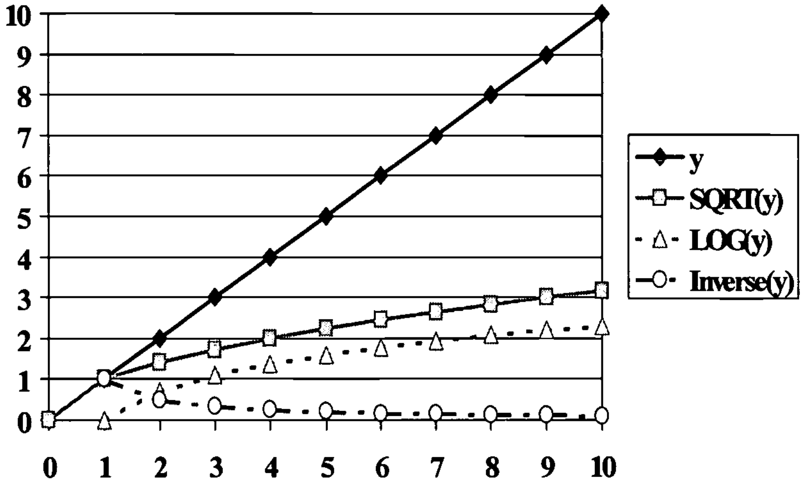
Should outlier removal not be an option, or not produce the desired results, another option is the use of data transformations. It is beyond the scope of this paper to fully discuss all options. Thus, I will focus on three more common data transformations discussed in texts and the literature: square root, logarithmic, and inverse transformations.

How does one tell when a variable is violating the assumption of normality?

There are several ways to tell whether a variable is substantially non-normal. While researchers tend to report favoring "eyeballing the data," or visual inspection (Orr, Sackett, and DuBois, 1991), this can lead to a mistakes or the perception that one is "cooking the data." There are objective methods of assessing normality,

Address correspondence to: Jason W. Osborne, Ph.D, ERLCE, Poe Hall 608, Campus Box 7801, NCSU, Raleigh, NC, 27695-7801. The author may also be contacted at jason_osborne@ncsu.edu.

Figure 1.
The Effect of Transformations on Variables.



from simple examination of skew and kurtosis to examination of P-P plots. Finally, there are inferential methods of comparing distributions to other known distributions, such as the Kolmorogov-Smirinov test, which provides a very sensitive test for deviation from normality. All of these, and more, are available in commonly-used statistical packages. Once a determination of non-normality is made, and obvious routes such as outlier detection have been tried, the researcher is faced with the decision to analyze the data in a non-normal state or to transform.

Theoretical issues surrounding a data transformation: How does one interpret transformed data?

In brief, data transformations should not be undertaken lightly. Data transformations change the fundamental nature of the data, and hence the interpretation of the results. For example, an analysis involving substantively-interpretable variables, such as yearly income, age, or IQ test scores are made tremendously more complicated once transformations are introduced. Many people can easily interpret results regarding these variables, but how many can easily (or correctly) interpret analyses involving the logarithm of IQ, the square root of age, or the inverse of income? Not only are these different variables, many of them are non-linear transformations of the original variables. Again, these are issues that are beyond the scope of this paper to address sufficiently. However, briefly, all three of these transformations are curvilinear transformations that change the nature of the variable you are studying to a certain extent. Once a variable, such as income has been transformed, it is no longer straightforward to interpret that variable, as it is now the square root of income, or the log of income, or the inverse of income. Thus, researchers must be careful when interpreting results based on transformed data.

As presented in Figure 1, as variables are transformed they take a curvilinear relationship to the original variable. Thus, interpretation is now more complicated. Not only does the author need to take into account that there is now a curvilinear relationship between the original variable and the new variable, but likely also a curvilinear relationship between the transformed variable and any other variable in the analysis. Further, the quality of the variable has now changed. If it had been ratio or interval, it is no longer so. If a variable with those qualities were subjected to a square root transformation, where the variable's old values were {0, 1, 2, 3, 4} the new values are now {0, 1, 1.41, 1.73, 2}—the intervals are no longer equal between successive values. This is addressed more explicitly in Table 1, below.

Mathematical issues surrounding a data transformation: Does the minimum value of a distribution influence the efficacy of a transformation?

All three of these transformations are designed to reduce positive skew. Should a researcher have a negatively skewed variable, the procedure is to reflect, or reverse the distribution, apply one of these transformations, and then reflect again to return the distribution to its

Table 1.
Effects of various transformations on variables

Original Y	**0.00**	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**8.00**	**9.00**
SquareRoot(Y)	0.00	1.00	1.41	1.73	2.00	2.24	2.45	2.65	2.83	3.00
gap		1.00	0.41	0.32	0.27	0.24	0.21	0.20	0.18	0.17
Log (Y)	---	0.00	0.69	1.10	1.39	1.61	1.79	1.95	2.08	2.20
gap			0.69	0.41	0.29	0.22	0.18	0.15	0.13	0.12
Inverse(Y)	---	1.00	0.50	0.33	0.25	0.20	0.17	0.14	0.13	0.11
gap			-0.50	-0.17	-0.08	-0.05	-0.03	-0.02	-0.02	-0.01
Original Y	**10.00**	**11.00**	**12.00**	**13.00**	**14.00**	**15.00**	**16.00**	**17.00**	**18.00**	**19.00**
SquareRoot(Y)	3.16	3.32	3.46	3.61	3.74	3.87	4.00	4.12	4.24	4.36
gap	0.16	0.15	0.15	0.14	0.14	0.13	0.13	0.12	0.12	0.12
Log (Y)	2.30	2.40	2.48	2.56	2.64	2.71	2.77	2.83	2.89	2.94
gap	0.11	0.10	0.09	0.08	0.07	0.07	0.06	0.06	0.06	0.05
Inverse(Y)	0.10	0.09	0.08	0.08	0.07	0.07	0.06	0.06	0.06	0.05
gap	-0.01	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00
Original Y	**100.00**	**101.00**	**102.00**	**103.00**	**104.00**	**105.00**	**106.00**	**107.00**	**108.00**	**109.00**
SquareRoot(Y)	10.00	10.05	10.10	10.15	10.20	10.25	10.30	10.34	10.39	10.44
gap		0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Log (Y)	4.61	4.62	4.62	4.63	4.64	4.65	4.66	4.67	4.68	4.69
gap		0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Inverse(Y)	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
gap		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

original order).When a researcher is considering utilizing a data transformation, that researcher must be aware of the mathematical considerations of that transformation. For example, the square root of a negative number is undefined, and one cannot take the log of a negative number or 0, and the inverse of 0 is undefined. Thus, should one have negative or zero values in the distribution, the researcher must first add a constant to the variable to move the distribution to a point where data transformations are possible.

Note that adding a constant to a variable changes only the mean, not the standard deviation or variance, skew, or kurtosis. However, the size of the constant and the place on the number line that the constant moves the distribution to can influence the effect of any subsequent data transformations. The argument posited here is that the researcher should only add a constant in such a way that (a) the distribution is moved to a point on the number line where there are no values that will yield undefined results (i.e., negative numbers for a square root transformation, or negative numbers and zeros for log and inverse transformations), and (b) the minimum value (left anchor) of the distribution should be moved to exactly 0 if the researcher is planning on using a square root transformation, or exactly 1 if the researcher is planning to use log or inverse transformations. This point is one generally not made in discussions of transformations; but is critical in determining the efficacy of the transformation.

The reason behind this assertion has to do with the effect of these transformations on 0 and/or 1 as opposed to other numbers. For example, the square root of 0 and 1 are, respectively, 0 and 1, whereas the square root of 2 is 1.41, and of 3 is 1.73. Thus, a square root transformation on a distribution anchored at 0 will move a positivcly-skewed distribution toward normality because the scores on the "tail" are moved closer in toward the center of the distribution, while scores on the leftmost part of the distribution are not moved at all. This "compression" of the tail reduces skew. However, this only works due to the special properties of 0 or 1, which remain fixed. Should the minimum score of a distribution be a number other than 0 (or 1 in the case of the log and inverse transformations) then the transformation

Table 2
Variable skew as a function of the minimum score of a distribution

	Min = 0	Min = 1	Min = 2	Min = 3	Min = 5	Min = 10	Min = 100
Square Root	0.22	0.93	1.11	1.21	1.31	1.42	1.56
Log	---	0.44	0.72	0.88	1.07	1.27	1.54
Inverse	---	0.12	-0.18	-0.39	-0.67	-1.00	-1.50

Note: Skewness reported. Original variable's skewness was 1.58.

will be less effective, as the entire distribution is being moved along the number line, rather than just the right tail.

In Table 1 this becomes evident more clearly. In the table, some example scores for a variable, along with the square root, log, and inverse transformations of these scores are presented. Additionally, the "gap" between each two adjacent numbers is calculated. Looking at the results of a square root transformation, for example, one can see that transforming the numbers 0 through 9 changes the relative distance between those two scores from an original distance of 1.0 to distances ranging from 1.0 (the gap between the square root of 0 and the square root of 1) to 0.17 (the gap between the square root of 8 and the square root of 9). Thus, one can see how the tail of a positively skewed distribution is compressed down and the distribution becomes more normal. However, looking at the second set of data, 10 through 19, the gaps are much more even between the transformed numbers (ranging from 0.16 to 0.12). Thus, while the distance between the higher numbers is compressed somewhat more than the lower numbers, it is nowhere near the magnitude difference as seen in the first set. Finally, looking at the bottom set of numbers (100-109), there is virtual uniformity in the amount of compression across the range (0.05 gap, after rounding). In this case, there would be virtually no effect of a square root transformation, as the relative distances between scores remain almost as constant as the original data.

Similar effects can be seen for the other two transformations, indicating that the effectiveness of the logarithmic and inverse transformations are most effective when the minimum value of the distribution is 1.0.

In order to demonstrate the effects of minimum values on the efficacy of transformations, data were drawn from the National Education Longitudinal Survey of 1988. The variable used represented the number of undesirable things (offered drugs, had something stolen, threatened with violence, etc.) that had happened to a student, which was created by the author for another project. This variable ranged from 0 to 6, and was highly skewed, with 40.4% reporting none of the events occurring, 34.9% reporting only one event, and less than 10% reporting more than two of the events occurring. The initial skew was 1.58, a substantial deviation from normality, making this variable a good candidate for transformation. The relative effects of transformations on the skew of this variable are presented in Table 2.

As the results indicate, all three types of transformations worked very well on the original distribution, anchored at a minimum of 0 (or 1 for the log and inverse transformations). However, the efficacy of the transformation quickly diminished as constants were added to the distribution. Even a move from 0 to 1, or 1 to 2 dramatically diminished the effectiveness of the transformation. Once the minimum reached 10, the skew was over 1.0 for all three transformations, and at a minimum of 100 the skewness was approaching the original, non-transformed skew in all three cases. These results highlight the importance of the minimum value of a distribution should a researcher intend to employ data transformations on that variable.

While the initial discussion involved the necessity of adding constants to variables to allow for transformations should there be negative numbers (or in the case of log or inverse transformations, 0), these results should also be considered when a variable has a range of, say 200-800, as with SAT or GRE scores where non-normality might be an issue. In cases where variables do not naturally have 0 as their minimum, it might be useful to subtract a constant to move the distribution to a 0 or 1 minimum.

Conclusions

The goal of this paper was to explore the effects of data transformations on variables, particularly the extent to which the anchor or starting value affects the effect of the transformation. This is something that, to my knowledge, is not adequately addressed in statistical texts, and may profoundly affect the benefit or effect of a transformation if researchers do not attend to this issue.

The examples above demonstrate that as the leftmost value (anchor value) of a distribution moves from 0 or 1, the efficacy of the transformation diminishes exponentially.

References

Baker, G. A. (1934). Transformation of non-normal frequency distributions into normal distributions. *Annals of Mathematical Statistics, 5,* 113-123.

Bartlett, M. S., (1947). The use of transformation. *Biometric Bulletin, 3,* 39-52.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Finney, D. J. (1948). Transformation of frequency distributions. *Nature, London, 162, 898*

Judd, C. M., & McClelland, G.H. (1989). *Data analysis: A model-comparison approach.* San Diego, CA: Harcourt Brace Jovanovich.

Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44,* 473- 486.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research.* Harcourt Brace: Orlando, FL.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics.* New York: Harper Collins.

Zimmerman, D. W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. *Journal of Experimental Education, 64,* 71-78.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67,* 55-68.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM033745

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The effects of Minimum Values on Data Transformations

Author(s): Jason W. Osborne

Corporate Source:

Publication Date: AERA 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPROOUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:	Printed Name/Position/Title: Jason Osborne Asst Prof	
Organization/Address: North Carolina State U	Telephone: 919 515 1714	FAX:
	E-Mail Address: jason_osborne@ncsu.edu	Date: 3/18/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com