Adv Physiol Educ 42: 343–347, 2018;
doi:10.1152/advan.00018.2018.

STAYING CURRENT

Explorations in statistics: the log transformation

Douglas Curran-Everett

Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, Colorado; and Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Denver, Colorado

Submitted 26 January 2018; accepted in final form 26 March 2018

Curran-Everett D. Explorations in statistics: the log transformation. *Adv Physiol Educ* 42: 343–347, 2018; doi:10.1152/advan.00018.2018.—Learning about statistics is a lot like learning about science: the learning is more meaningful if you can actively explore. This thirteenth installment of *Explorations in Statistics* explores the log transformation, an established technique that rescales the actual observations from an experiment so that the assumptions of some statistical analysis are better met. A general assumption in statistics is that the variability of some response Y is homogeneous across groups or across some predictor variable X. If the variability—the standard deviation—varies in rough proportion to the mean value of Y, a log transformation can equalize the standard deviations. Moreover, if the actual observations from an experiment conform to a skewed distribution, then a log transformation can make the theoretical distribution of the sample mean more consistent with a normal distribution. This is important: the results of a one-sample t test are meaningful only if the theoretical distribution of the sample mean is roughly normal. If we log-transform our observations, then we want to confirm the transformation was useful. We can do this if we use the Box-Cox method, if we bootstrap the sample mean and the statistic t itself, and if we assess the residual plots from the statistical model of the actual and transformed sample observations.

bootstrap; Central Limit Theorem; normal quantile plot; residual plots

INTRODUCTION

This thirteenth paper in *Explorations in Statistics* (see Refs. 7–17, 19) explores the log transformation,[1] a long-standing technique that rescales the sample observations—the actual measurements—from an experiment so that the assumptions of some statistical analysis are better met (1, 6, 33). As you might expect from its lengthy history, the log transformation is featured in textbooks of statistics (20, 28, 29, 31, 32), papers on clinical statistics (2–4, 25), and papers in the formal statistical literature (1, 6, 26, 33).

The Log Transformation: An Overview

There are two main ways a log transformation can help sample observations better meet the assumptions of some statistical analysis. First, a general assumption in statistics is that the variability of some response Y is homogeneous across groups or across some predictor variable X (see Refs. 20 and 31).[2] If the variability—the standard deviation—varies in rough proportion to the mean value of Y, a log transformation of the actual observations can equalize the standard deviations (Table 1 and Fig. 1). In more formal statistical parlance, a log transformation can stabilize the variance (1, 6, 20, 25, 31).

Second, in our earlier explorations (10, 17) we learned that the results of a one-sample t test are meaningful only if the theoretical distribution of the sample mean is roughly normal. When we explored the bootstrap (10) we learned that a log transformation of skewed C-reactive protein values (Fig. 2) was the optimal transformation (Fig. 3). Although the log-transformed values were less skewed, their distribution was still inconsistent with a normal distribution (see Fig. 2). Instead, the real impact of this log transformation was on the theoretical distribution of the sample mean (Fig. 4). A log transformation can help the distribution of the observations themselves be more normal (2–4, 6, 10, 18, 20, 25), or—perhaps more often—it can make the theoretical distribution of the sample mean more normal.

Although these advantages of a log transformation have been long established, three recent papers (22–24) claim that a log transformation may fail to reduce—in fact, it may even exacerbate—skewness.[3] In other words, a log transformation may fail to do what it is purported to do. Needless to say, the initial paper (23) triggered a spirited exchange (5, 24) about the value of a log transformation.

In this exploration we will investigate whether the arguments in Refs. 23 and 24 are sufficient to warrant caution in using the log transformation. First, we need to review the software we will use to help us do that.

R: Basic Operations

The first paper in this series (7) summarized R (30) and outlined its installation. For this exploration there are three more steps: download the R script Advances_Statistics_Code_Log.R and the data file Table_1_Data.csv[4] to your Advances folder, confirm you installed boot and MASS in our previous explorations (10, 11, 17, 19), and install the new package moments (27).[5]

Address for reprint requests and other correspondence: D. Curran-Everett, Division of Biostatistics and Bioinformatics, M222, National Jewish Health, 1400 Jackson St., Denver, CO 80206–2761 (e-mail: EverettD@NJHealth.org).

[1] The log transformation is just one kind of transformation (see Refs. 1, 6, 18).

[2] This assumption is often called homogeneity of variance or homoscedasticity.

[3] The genesis for this paper was Ref. 22 which amalgamates Refs. 23 and 24. I include Ref. 22 only because it may be more accessible than Refs. 23 and 24.

[4] These files are posted as Supplemental Data at the end of the article posted on the *Advances in Physiology Education* website.

[5] Some of our previous explorations (10, 11, 14–16, 19) detail how to install an extra R package.

Table 1. *Plankton data*

	Actual: y				Transformed: log y				
	1	2	3	4	1	2	3	4	
	895	1520	11,000	43,300	2.952	3.182	4.041	4.636	
	540	1610	8,600	32,800	2.732	3.207	3.934	4.516	
	1020	1900	8,260	28,800	3.009	3.279	3.917	4.459	
	470	1350	9,830	34,600	2.672	3.130	3.993	4.539	
	428	980	7,600	27,800	2.631	2.991	3.881	4.444	
	620	1710	9,650	32,800	2.792	3.233	3.985	4.516	
	760	1930	8,900	28,100	2.881	3.286	3.949	4.449	
	537	1960	6,060	18,900	2.730	3.292	3.782	4.276	
	845	1840	10,200	31,400	2.927	3.265	4.009	4.497	
	1050	2410	15,500	39,500	3.021	3.382	4.190	4.597	
	387	1520	9,250	29,000	2.588	3.182	3.966	4.462	
	497	1685	7,900	22,300	2.696	3.227	3.898	4.438	
Ave {y}	671	1701	9,396	30,775	2.803	3.221	3.962	4.478	Ave {log y}
SD {y}	234	357	2,326	6,689	0.150	0.098	0.099	0.098	SD {log y}

Values represent estimated numbers of 4 types of plankton from 12 separate hauls (after Ref. 31). The standard deviations of the actual numbers, SD {y}, vary in rough proportion to their corresponding averages, Ave {y}. In contrast, the standard deviations of the transformed numbers, SD {log y}, were nearly equal. The commands in *lines 29–40* of Advances_Statistics Code_Log.R read in these data and compute these averages and standard deviations. To obtain these results, highlight and submit the lines of code from `Table 1: first line to Table 1: last line`.

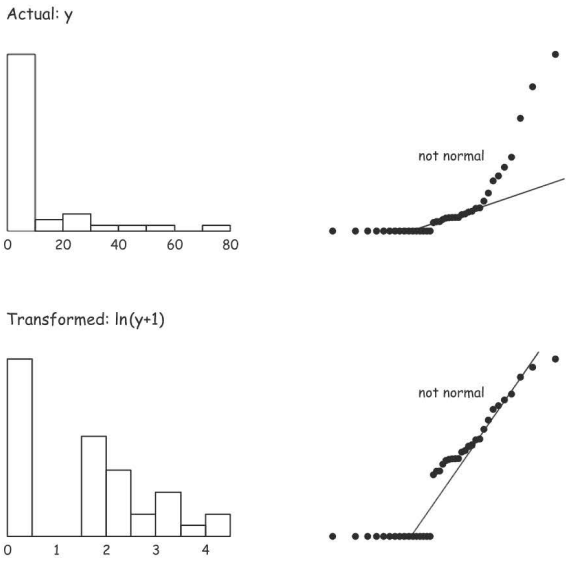

Fig. 2. Distributions (*left*) and normal quantile plots (*right*) of the actual and transformed C-reactive protein observations in Ref. 10. The transformation changed the distribution of the observations, but the transformed values remained inconsistent with a normal distribution. Advances_Statistics Code_Log.R does not create this data graphic (adapted from Ref. 10).

To run R commands. If you use a Mac, highlight the commands you want to submit and then press ⌘↵ (command key+enter). If you use a PC, highlight the commands you want to submit, right-click, and then click Run line or selection. Or, highlight the commands you want to submit and then press Ctrl+R.

The Log Transformation: A Possible Counterargument?

In their initial paper (Ref. 23, p. 233), Feng et al. mention that a log transformation does not always reduce skewness: "[It] is easy to find distributions for which the log transformation actually introduces more skewness." But they fail to offer an example of such a distribution. In their response to comments (24), they remedy that. We can recreate their simulation of 10,000 observations from a distribution for which a log transformation does indeed exacerbate skewness (Fig. 5). The real question is, is a log transformation of these simulated data even warranted?

In some of our earlier explorations (10, 17), we learned that the theoretical distribution of the sample mean could be roughly normal—an assumption for the results of a t test to be meaningful—regardless of the distribution of the actual observations as long as our sample size was big enough. The sample size of 10,000 from the skewed distribution used by Feng et al. (23) as a counterargument for the log transformation *is* big enough (Fig. 6). In other words, a log transformation of these 10,000 observations is unnecessary.

Suppose we pretend that the sample size of 10,000 is not big enough for the theoretical distribution of the sample mean to be roughly normal. If we use the Box-Cox method (6, 20) to estimate an appropriate transformation of these observations,

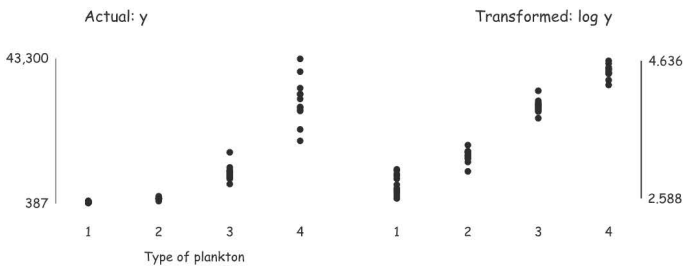

Fig. 1. Plankton data from Table 1. The commands in *lines 49–68* of Advances_Statistics_Code_Log.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from `Figure 1: first line to Figure 1: last line`.

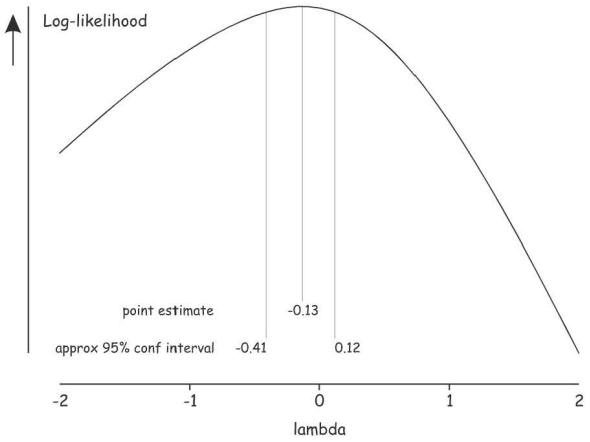

Fig. 3. Likelihood approach to data transformation. For each value of λ, the log-likelihood is calculated in a manner similar to *Eq. 1* in Ref. 10. For the C-reactive protein values, the log-likelihood estimate of λ that maximizes the log-likelihood is −0.13, and an approximate 95% confidence interval for λ is [−0.41, +0.12]. Because this interval includes 0, a log transformation of the C-reactive protein values is reasonable (6, 18). The commands in *lines 98–104* of Advances_Statistics Code_Log.R return these point and interval estimates of λ which are identical to those we obtained using the equivalent maximum likelihood approach (see Fig. 5 in Ref. 10). The commands in *lines 109–117* of Advances_Statistics Code_Log.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from `Figure 3: first line to Figure 3: last line`.

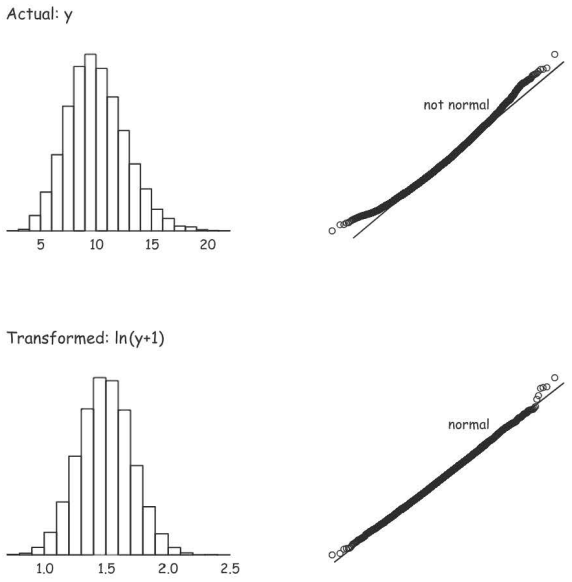

Fig. 4. Bootstrap distributions (*left*) and normal quantile plots (*right*) of 10,000 sample means, each with 40 observations. The bootstrap replications were drawn at random from the actual or transformed C-reactive protein observations in Ref. 10. The bootstrap sample means from the actual observations are inconsistent with a normal distribution. In contrast, the bootstrap sample means from the transformed observations are consistent with a normal distribution. Advances_Statistics Code_Log.R does not create this data graphic (adapted from Ref. 10).

we obtain a point estimate of $\lambda = 0.52$ and an approximate 95% confidence interval for λ of [0.50, 0.54] (Fig. 7). Because this interval excludes 0, a log transformation of these observations is inappropriate.[6]

Practical Considerations

When we explored the assumption of normality (17), we said that if our sample size is not big enough then we can transform the observations in the hopes that the theoretical distribution of the sample mean will be more normally distributed. How do we think about transformation in practice? To simplify our lives, suppose we have sample observations—some data—that we anticipate analyzing with a one-sample *t* test.

The first thing we do is plot the data to get a sense of their distribution (see Ref. 13, p. 351, *Rule 1*). We might also construct a normal quantile plot to assess whether the data are consistent with a normal distribution (see Fig. 2, *top*). If they are, then a transformation is likely to be unnecessary. If the data are inconsistent with a normal distribution (see Fig. 2, *top*), then we bootstrap the sample mean $\bar{y}$ and the statistic *t* (10, 17). If these bootstrap distributions are inconsistent with a normal distribution (see Fig. 4, *top*), then the results from the one-sample *t* test—its *P* value and confidence interval—are likely to mislead us.

We have now arrived at the point where we may want to transform the data. As we first discovered when we explored the bootstrap (10), the Box-Cox method (6, 20) is an effective tool with which to identify an appropriate transformation. If our sample observations happen to be the C-reactive protein values in Fig. 2, then the Box-Cox method identifies that a log transformation is reasonable (see Fig. 3).

If we identify that a particular transformation is appropriate, then we do not want to blindly assume the transformation is beneficial as Feng et al. (23) appear to suggest. Rather, we want to confirm it: so we bootstrap the sample mean $\bar{y}$ and the statistic *t* of the transformed values (10, 17). The goal: to assess how well the assumptions of some statistical procedure have been satisfied. For our plankton data, a log transformation appears to be beneficial: the variability of the estimated numbers is more homogeneous across the four types of plankton (see Table 1 and Fig. 1). For our C-reactive protein values, a log transformation is beneficial: the theoretical distribution of the sample mean is now consistent with a normal distribution (see Fig. 4).

Last, we also want to use a general technique akin to residual plots in regression (13). In regression, residual plots help us decide if our provisional statistical model of the relationship between *Y* and *X* is appropriate. If we want to assess whether a transformation is appropriate, then we want to examine the residual plots from a statistical model of the actual and transformed data. In this situation, for each observation, the residual is the difference between the observed value *y* and the value estimated by the statistical model. If a particular transformation is appropriate, then there is no obvious pattern to the residuals (see Ref. 13, Fig. 6). Residual plots confirm that the log transformation of our plankton data is appropriate (Fig. 8).

Contrary to urban legend, if we transform our data before we analyze them, we are not complicit in some kind of hanky-panky. The goal of transformation is not to identify a rescaling

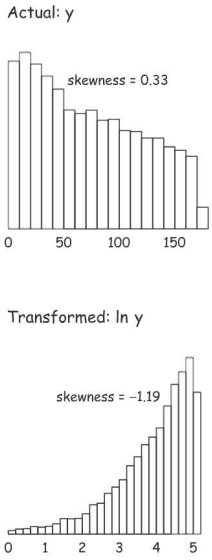

Fig. 5. Distributions of 10,000 simulated actual and transformed observations (after Ref. 24). We simulate the actual observations in two steps. First, we draw at random a value *u* from a uniform distribution (0, 1) (see Ref. 17, Fig. 1). Then, we calculate the actual observation *y* as $y = [100\ (e^u - 1)] + 1$; see Ref. 24. For example, if $u = 0$, then $y = 1$, and if $u = 1$, then $y = 172.8$. We obtain the transformed observations by taking the natural logarithm of the actual observations: ln *y*. The skewness of these distributions mirrors the skewness of the distributions depicted in Ref. 24: 0.34 and −1.16. The commands in *lines 129–162* of Advances_Statistics Code_Log.R create these distributions and this data graphic. Your distributions will differ slightly. To generate this data graphic, highlight and submit the lines of code from `Figure 5: first line` to `Figure 5: last line`.

[6] Instead, if the sample size of 10,000 were not big enough, we might opt for a square-root transformation: $\lambda = 0.50$.

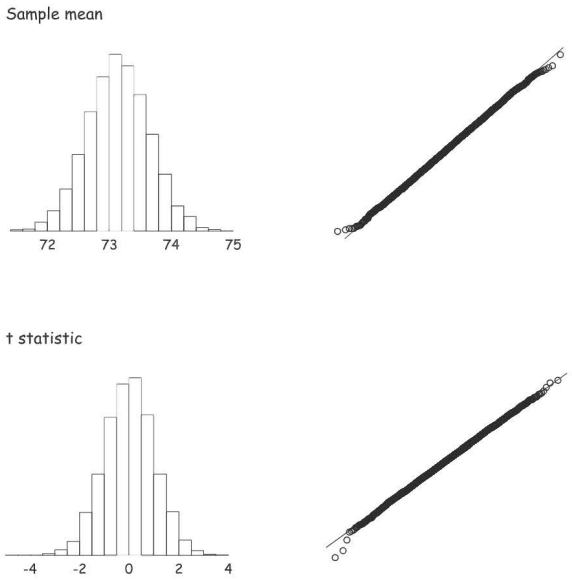

Fig. 6. Bootstrap distributions (*left*) and normal quantile plots (*right*) of 10,000 sample means (*top*) and their corresponding one-sample t statistics (*bottom*). The bootstrap replications were drawn at random from the 10,000 actual observations depicted in Fig. 5 (*top*). The bootstrap distributions of the sample mean and t statistic are consistent with a normal distribution: this means the inference we make from a normal-theory hypothesis test or confidence interval is justified. The commands in *lines 175–222* of Advances_Statistics Code_Log.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from `Figure 6: first line to Figure 6: last line`. In the theoretical distribution of t with 10,000 − 1 degrees of freedom, 2.5% of the possible values of t are less than −1.960, and 2.5% of the possible values of t are greater than +1.960. In the bootstrap distribution of t, 2.6% of the possible values of t^* are less than −1.960, and 2.5% of the possible values of t^* are greater than +1.960. The commands in *lines 232–237* of Advances_Statistics Code_Log.R return these values. Your percentages will differ slightly.

of the data that produces a statistically meaningful result. Instead, the goal of transformation is to identify a rescaling of the data so they better meet the assumptions of some statistical procedure (1, 6, 33).

On the other hand, if a transformation fails to identify a constructive rescaling of the data, then all is not lost: we can use the bootstrap to estimate a confidence interval (10), and we can use a permutation method to test a scientific null hypothesis (14). And, as Efron has written (21), "When there *is* something to permute ... it is a good idea to do so, even if other methods like the bootstrap are also brought to bear."

Summary

As this exploration has demonstrated, a log transformation can rescale the actual measurements from an experiment so that *1*) the variability of some response is more homogeneous, or *2*) the theoretical distribution of the sample mean is consistent with a normal distribution. In each situation, a log transformation can help the sample observations better satisfy the assumptions of some statistical analysis. As we have also seen, however, if we log-transform our sample observations, then we want to confirm the transformation was useful. We can do this if we use the Box-Cox method, if we bootstrap the sample mean and the statistic t itself, and if we assess the residual plots from the statistical model of the actual and transformed sample observations.

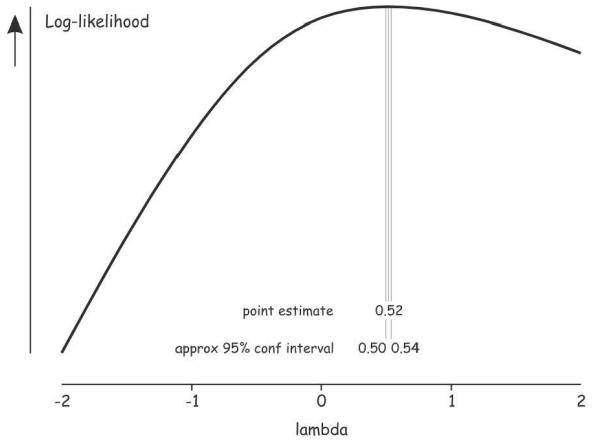

Fig. 7. Box-Cox approach (6) to data transformation applied to the actual observations in Fig. 5. The estimate of λ that maximizes the log-likelihood is 0.52, and an approximate 95% confidence interval for λ is [0.50, 0.54]. Because this interval excludes 0, a log transformation of the actual observations in Fig. 5 is inappropriate. The approximate 95% confidence interval for λ is narrow because there are 10,000 observations. The commands in *lines 251–257* of Advances_Statistics Code_Log.R return these point and interval estimates of λ. Your values will differ slightly. The commands in *lines 262–270* of Advances_Statistics_Code Log.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from `Figure 7: first line to Figure 7: last line`.

ACKNOWLEDGMENTS

I thank Bryan Mackenzie (University of Cincinnati, Cincinnati, OH) for suggesting this paper, and I thank Gerald DiBona (Göteborgs Universitet, Göteborg, Sweden and University of Iowa College of Medicine, Iowa City, IA), for helpful comments and suggestions.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author.

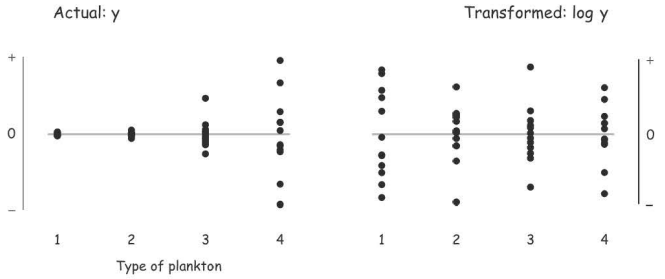

Fig. 8. Residual plots from a statistical model of the actual and transformed plankton data (see Table 1 and Fig. 1). For the ith type of plankton, the jth observation y_{ij} can be modeled as $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ for $i = 1, 2, 3, 4$ and $j = 1, 2, ..., 12$, where μ is a constant that represents a common underlying mean, α_i is a component that represents the type i effect, and ε_{ij} is a component that represents random error. We assume that ε_{ij} is distributed normally with mean 0 and standard deviation σ. The residual e_{ij} is the difference between the observed value y_{ij} and the value estimated by the statistical model. If the statistical model appears to be appropriate—if the assumptions of the analysis appear to have been met—then there is no pattern to the residuals. If the assumptions of the analysis appear to have been violated, then a residual plot depicts some sort of pattern (see Ref. 13, Fig. 6). The statistical model of the actual plankton data is inappropriate: the variability differs across plankton type. The statistical model of the transformed plankton data is appropriate: there is no discernible pattern to the residuals. The commands in *lines 283–307* of Advances_Statistics_Code_Log.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from `Figure 8: first line to Figure 8: last line`.

AUTHOR CONTRIBUTIONS

D.C.-E. conceived and designed research; performed experiments; analyzed data; interpreted results of experiments; prepared figures; drafted manuscript; edited and revised manuscript; approved final version of manuscript.

REFERENCES

1. **Bartlett MS.** The use of transformations. *Biometrics* 3: 39–52, 1947. doi:10.2307/3001536.
2. **Bland JM, Altman DG.** Transformations, means, and confidence intervals. *BMJ* 312: 1079, 1996. doi:10.1136/bmj.312.7038.1079.
3. **Bland JM, Altman DG.** Transforming data. *BMJ* 312: 770, 1996. doi:10.1136/bmj.312.7033.770.
4. **Bland JM, Altman DG.** The use of transformation when comparing two means. *BMJ* 312: 1153, 1996. doi:10.1136/bmj.312.7039.1153.
5. **Bland JM, Altman DG, Rohlf FJ.** In defence of logarithmic transformations. *Stat Med* 32: 3766–3768, 2013. doi:10.1002/sim.5772.
6. **Box GEP, Cox DR.** An analysis of transformations. *J R Stat Soc Series B Stat Methodol* 26: 211–243, 1964.
7. **Curran-Everett D.** Explorations in statistics: standard deviations and standard errors. *Adv Physiol Educ* 32: 203–208, 2008. doi:10.1152/advan.90123.2008.
8. **Curran-Everett D.** Explorations in statistics: confidence intervals. *Adv Physiol Educ* 33: 87–90, 2009. doi:10.1152/advan.00006.2009.
9. **Curran-Everett D.** Explorations in statistics: hypothesis tests and *P* values. *Adv Physiol Educ* 33: 81–86, 2009. doi:10.1152/advan.90218.2008.
10. **Curran-Everett D.** Explorations in statistics: the bootstrap. *Adv Physiol Educ* 33: 286–292, 2009. doi:10.1152/advan.00062.2009.
11. **Curran-Everett D.** Explorations in statistics: correlation. *Adv Physiol Educ* 34: 186–191, 2010. doi:10.1152/advan.00068.2010.
12. **Curran-Everett D.** Explorations in statistics: power. *Adv Physiol Educ* 34: 41–43, 2010. doi:10.1152/advan.00001.2010.
13. **Curran-Everett D.** Explorations in statistics: regression. *Adv Physiol Educ* 35: 347–352, 2011. doi:10.1152/advan.00051.2011.
14. **Curran-Everett D.** Explorations in statistics: permutation methods. *Adv Physiol Educ* 36: 181–187, 2012. doi:10.1152/advan.00072.2012.
15. **Curran-Everett D.** Explorations in statistics: the analysis of ratios and normalized data. *Adv Physiol Educ* 37: 213–219, 2013. doi:10.1152/advan.00053.2013.
16. **Curran-Everett D.** Explorations in statistics: statistical facets of reproducibility. *Adv Physiol Educ* 40: 248–252, 2016. doi:10.1152/advan.00042.2016.
17. **Curran-Everett D.** Explorations in statistics: the assumption of normality. *Adv Physiol Educ* 41: 449–453, 2017. doi:10.1152/advan.00064.2017.
18. **Curran-Everett D, Taylor S, Kafadar K.** Fundamental concepts in statistics: elucidation and illustration. *J Appl Physiol (1985)* 85: 775–786, 1998. doi:10.1152/jappl.1998.85.3.775.
19. **Curran-Everett D, Williams CL.** Explorations in statistics: the analysis of change. *Adv Physiol Educ* 39: 49–54, 2015. doi:10.1152/advan.00018.2015.
20. **Draper NR, Smith H.** *Applied Regression Analysis* (2nd ed.). New York: Wiley, 1981.
21. **Efron B, Tibshirani RJ.** *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993, p. 218. doi:10.1007/978-1-4899-4541-9.
22. **Feng C, Wang H, Lu N, Chen T, He H, Lu Y, Tu XM.** Log-transformation and its implications for data analysis. *Shanghai Arch Psychiatry* 26: 105–109, 2014. doi:10.3969/j.issn.1002-0829.2014.02.009.
23. **Feng C, Wang H, Lu N, Tu XM.** Log transformation: application and interpretation in biomedical research. *Stat Med* 32: 230–239, 2013. doi:10.1002/sim.5486.
24. **Feng C, Wang H, Lu N, Tu XM.** Response to comments on "Log transformation: application and interpretation in biomedical research". *Stat Med* 32: 3772–3774, 2013. doi:10.1002/sim.5840.
25. **Healy MJR.** Data transformations. *Arch Dis Child* 69: 260–264, 1993. doi:10.1136/adc.69.2.260.
26. **Keene ON.** The log transformation is special. *Stat Med* 14: 811–819, 1995. doi:10.1002/sim.4780140810.
27. **Komsta L, Novomestky F.** *moments: moments, cumulants, skewness, kurtosis and related tests* (Online). R package, version 0.14, 2015. http://www.r-project.org, http://www.komsta.net/.
28. **Moore DS, McCabe GP, Craig BA.** *Introduction to the Practice of Statistics* (6th ed.). New York: WH Freeman, 2009, p. 119–121, 435–438.
29. **Moses LE.** *Think and Explain with Statistics*. Reading, MA: Addison-Wesley, 1986, p. 162–163.
30. **R Core Team.** *R: A Language and Environment for Statistical Computing* (Online). Vienna: R Foundation for Statistical Computing, 2017. http://www.R-project.org.
31. **Snedecor GW, Cochran WG.** *Statistical Methods* (7th ed.). Ames, IA: Iowa State Univ. Press, 1980.
32. **Sokal RR, Rohlf FJ.** *Biometry* (3rd ed.). New York: WH Freeman, 1995, p. 413–415, 533–536.
33. **Tukey JW.** On the comparative anatomy of transformations. *Ann Math Stat* 28: 602–632, 1957. doi:10.1214/aoms/1177706875.