Dealing With Non-normal Data

Kristin L. Sainani, PhD

INTRODUCTION

Although some continuous variables follow a normal, or bell-shaped, distribution, many do not. Non-normal distributions may lack symmetry, may have extreme values, or may have a flatter or steeper "dome" than a typical bell. There is nothing inherently wrong with non-normal data; some traits simply do not follow a bell curve. For example, data about coffee and alcohol consumption are rarely bell shaped. Instead, these follow a right-skewed distribution: they have a cluster of values at zero (nonconsumers), another bunch in the low-to-moderate range, and a few extreme values to the right (heavy consumers). Researchers need to be aware of whether their variables follow normal or non-normal distributions, because this influences how data are described and analyzed. Non-normal variables, particularly those with extreme right or left tails, may be better summarized (described) with medians and percentiles rather than means and standard deviations. Standard statistical tests for analyzing continuous data (*t*-test, analysis of variance [ANOVA], linear regression) may also perform poorly on non-normal data but only if the sample size is small. In these cases, alternative statistical approaches may be warranted. This article reviews how to spot, describe, and analyze non-normal data, and clarifies when the "normality assumption" matters and when it is unimportant.

SPOTTING NON-NORMAL DATA

Researchers should always plot the distributions of their variables. Just looking at a simple histogram ("eye-balling" it) will usually reveal whether or not a given variable is normally distributed (follows a bell curve), is skewed (has a left or right tail), or otherwise deviates from a bell shape (eg, is flat). For example, histograms for 2 hypothetical variables, 1 with a heavy right skew (right tail) and 1 with a bell curve, are shown in Figure 1. A normal probability plot (or Q-Q plot) can also help assess normality; if a variable is normally distributed, then this plot will approximate a straight line (Figure 2).

Researchers may also apply formal tests of normality, such as the Shapiro-Wilk test or the Kolmogorov-Smirnov test. The null hypothesis for these tests is that the variable follows a normal distribution; thus, small *P* values give evidence against normality. For example, running the Shapiro-Wilk test on the hypothetical variable from Figures 1B and 2B gives a *P* value of .80, which indicates concordance with normality; whereas running this test on the skewed variable (Figures 1A and 2A) gives a *P* value of .001, which indicates a violation of normality. Formal normality tests are highly sensitive to sample size: very large samples may pick up unimportant deviations from normality, and very small samples may miss important deviations. Thus, these tests are optional and should always be interpreted alongside histograms and normality plots.

DESCRIBING NON-NORMAL DATA

Researchers typically describe continuous variables by using means and standard deviations. However, these descriptive statistics may be misleading for skewed data. Means and standard deviations are highly influenced by extreme values. For example, take a hypothetical randomized trial of 2 weight-loss regimens (Diet A and Diet B), with 10 subjects per group. Imagine that the mean weight loss in Diet A is 34.5 lb, and the mean weight loss in Diet B is 18.5 lb. One might easily jump to the conclusion that Diet A is superior to Diet B.

K.L.S. Division of Epidemiology, Department of Health Research and Policy, Stanford University, HRP Redwood Bldg, Stanford, CA 94305. Address correspondence to: K.L.S.; e-mail: kcobb@stanford.edu

Disclosure: nothing to disclose

1934-1482/12/$36.00
Printed in U.S.A.

http://dx.doi.org/10.1016/j.pmrj.2012.10.013

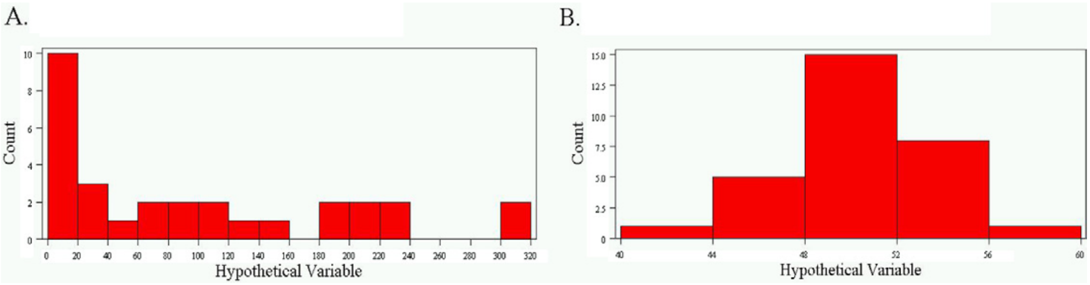

Figure 1. Histograms of a right-skewed variable and a normally distributed variable (both variables are computer generated). (A) Right-skewed data (N = 30). (B) Normally distributed data (N = 30).

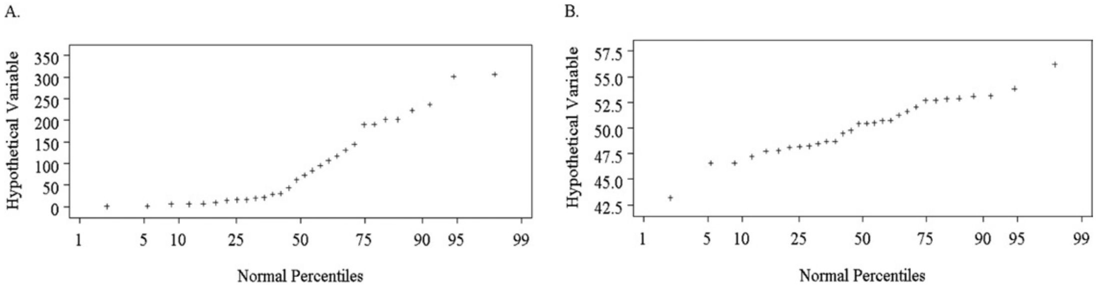

Figure 2. Normal probability plots of a right-skewed variable and a normally distributed variable. When the data follow a normal distribution, the normal probability plot will approximate a straight line. (A) Normal probability plot (skewed distribution). (B) Normal probability plot (normal distribution).

However, a closer inspection of the data reveals that this conclusion is incorrect. The distribution of weight loss in each group is shown in Figure 3. The raw data (change in pounds) are as follows:

Diet A: +4, +3, 0, −3, −4, −5, −11, −14, −15, −300

Diet B: −8, −10, −12, −16, −18, −20, −21, −24, −26, −30

The raw data and plot reveal that 1 individual in Diet A was highly successful (losing 300 lb!), but that, overall, the dieters in Diet B fared better. The mean for Diet A is misleading, because it is completely driven by a single extreme value.

When faced with such an "outlier," researchers are often tempted to simply throw away the data point. This is a mistake. If the data point is real (not simply a data recording or entry error), then it contains important information that should not be discarded. If 1 person truly lost 300 lb on Diet A, then it is possible that Diet A is highly effective for some people. Rather than ignoring this information, researchers can use statistical approaches that are not overly influenced by extreme values.

In this case, the median would be a better summary measure than the mean. The median is the middle-most value (or the average of the 2 middle-most values). The median weight change in Diet A is −4.5 lb, which means that 5 people lost fewer than 4.5 lb (+4, +3, 0, −3, and −4 lb), and 5 people lost more than this (−5, −11, −14, −15, and −300 lb). The median in Diet B is −19.0 lb (Table 1), which reflects the greater overall success in this group. Medians are

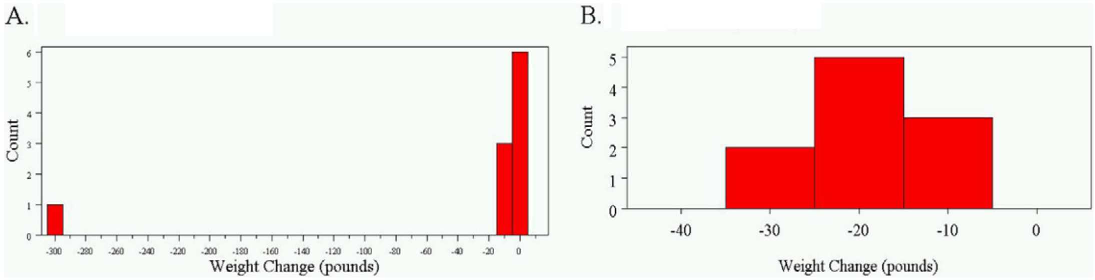

Figure 3. Histograms of the hypothetical weight-loss data from the Diet A and Diet B groups. (A) Diet A (N = 10). (B) Diet B (N = 10).

Table 1. *Descriptive statistics for weight change in the 2 hypothetical diet groups*

Diet Group	Mean, lb	Median, lb	Standard Deviation	Interquartile Range
A	−34.5	−4.5	93.5	0 to −14
B	−18.5	−19.0	7.1	−12 to −24

not influenced by extreme values; the median would remain unchanged regardless of whether the most successful dieter in Diet A lost 300, 100, or 30 lb.

Similar to the mean, the standard deviation in Diet A is highly influenced by the extreme value and does not reflect the experience of the majority. Nine of the 10 participants in Diet A had weight losses that were within 20 lb of each other, but the standard deviation, 93.5 lb, suggests much higher variability. Thus, for highly skewed data, it is preferable to report the interquartile range as a measure of variability. The interquartile range gives the middle 50% of the data, so is not influenced by extreme values (or by right or left tails). The interquartile range is 14 lb wide for Diet A and 12 lb wide for Diet B (Table 1), which suggests that the variability in weight loss is similar in the 2 groups for the bulk of participants.

ANALYZING NON-NORMAL DATA

Many researchers are aware that a certain family of statistical tests, called linear models, which include the *t*-test, ANOVA, and linear regression, have a "normality assumption." These tests are said to be appropriate only when the outcome (dependent) variable is normally distributed. In fact, this assumption is critical for small samples but is irrelevant above a certain sample size. Linear models make inferences about means; as long as the means are normally distributed, the inferences will be valid. Fortunately, means tend to follow a normal distribution even when the variable itself does not. This concept, called the Central Limit Theorem, is illustrated in Figure 4: even when the underlying trait follows a highly skewed distribution, the means approach a bell curve as the sample size increases.

What is "sufficiently large" will vary from problem to problem; but researchers estimate that, even with extreme deviations from normality, a sample size of approximately 80 is usually enough to run a *t*-test [1]; and much smaller sample sizes will suffice if the deviations from normality are more modest. The normality assumption is just 1 of the assumptions of linear models (others include homogeneity of variances, independence, and linearity). For small samples with large deviations from normality, linear models may lead to incorrect inferences, so researchers should consider alternative approaches: data transformations, nonparametric tests, or bootstrapping.

TRANSFORMING THE DATA

Often, applying a simple function, such as a square root or a log, to non-normal data will make the data more closely approximate a bell shape. For example, for skewed distributions, taking a natural log is often sufficient to remove the right or left tail (because logs rein in extreme values) (Figure 5). If an adequate transformation can be achieved, then researchers can run standard statistical tests on the transformed data. One drawback is

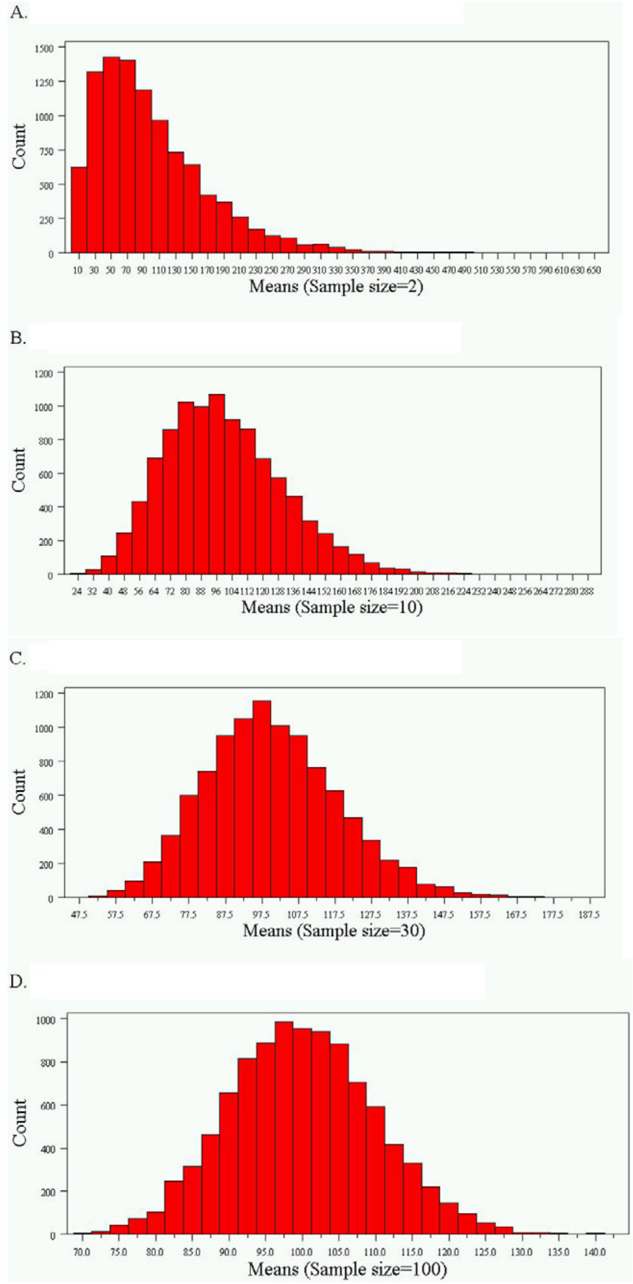

Figure 4. Illustration of the Central Limit Theorem. As sample sizes get larger, the distribution of the means approaches a bell curve even when the underlying distribution is highly skewed. (A) Distributions of the means from 10,000 samples of 2. (B) Distributions of the means from 10,000 samples of 10. (C) Distributions of the means from 10,000 samples of 30. (D) Distributions of the means from 10,000 samples of 100.

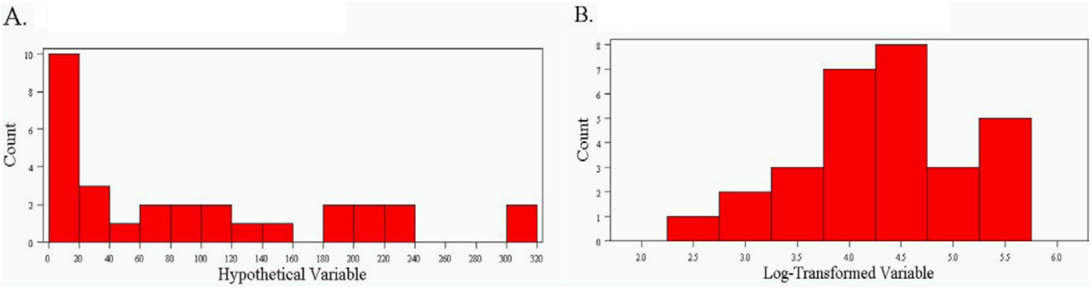

Figure 5. A right-skewed variable before (A) and after (B) applying a natural log. (A) Right-skewed data (N = 30). (B) Log-transformed right-skewed data.

that log-transformed data are more difficult for researchers and readers to intuitively understand.

NONPARAMETRIC TESTS

Another option is to analyze the data by using "nonparametric" tests, which do not make any assumption about the underlying distribution of the data. For example, the Wilcoxon signed rank test can be used in place of a 1-sample or paired *t*-test, the Wilcoxon rank sum test (also known as the Mann-Whitney *U* test) substitutes for the 2-sample *t*-test, and the Kruskal-Wallis test substitutes for ANOVA. Although these tests remain a "black box" for many researchers, the mechanics are actually easy to understand.

For example, take the hypothetical diet trial. The Wilcoxon rank sum test does exactly what the name says: it ranks the data and then sums them. First, rank the trial participants in terms of their weight loss from lowest (rank 1) to highest (rank 20), while ignoring the diet group; then sum the ranks within each group. Here are the rankings and the sums of the ranks for the diet trial:

Diet A: +4, +3, 0, −3, −4, −5, −11, −14, −15, −300 lb

Ranks: 1 2 3 4 5 6 9 11 12 20

Sum of the ranks: 1 + 2 + 3 + 4 + 5 + 6 + 9 + 11 + 12 + 20 = 73

Diet B: −8, −10, −12, −16, −18, −20, −21, −24, −26, −30 lb

Ranks: 7 8 10 13 14 15 16 17 18 19

Sum of the ranks: 7 + 8 + 10 + 13 + 14 + 15 + 16 + 17 + 18 + 19 = 137

Note that the most successful dieter gets a rank of 20, so this person gets credit for his or her success, but the ranked value does not exert undue influence on the sum. The Wilcoxon rank sum test formally tests the null hypothesis that the summed ranks in the 2 groups are equal. The sum from Diet B is sufficiently large that we can reject the null hypothesis (P = .017). In contrast, running a *t*-test on these data gives a nonsignificant result (P = .60), thus missing an important effect. The Wilcoxon signed rank and Kruskal-Wallis tests are based on similar ranking approaches. A drawback to nonparametric tests is that no effect sizes are calculated (the summed ranks have no inherent meaning); only a *P* value is generated.

BOOTSTRAPPING

A third way to deal with violations of normality is to use a technique called bootstrapping [2]. Rather than assume that the means follow a normal distribution, one can directly "observe" the distribution of the means and use this empirical distribution to make inferences. The procedure is best illustrated with a simple example. Suppose we have a data set of 30 observations that appear to be right skewed (as in Figure 1A). We can perform a bootstrap analysis as follows:

1. Extract a new sample of 30 observations from the original set of 30. The trick is to sample with replacement, such that some of the original observations are repeated and others are omitted. Here is an example of a "resampled" data set (n = 30): 48, 158, 107, 195, 195, 29, 73, 73, 73, 42, 187, 21, 61, 61, 61, 57, 57, 57, 57, 60, 212, 212, 44, 44, 244, 244, 244, 244, 43, 100.
2. Calculate the mean of this new sample of 30. For example, the mean for the above data set is 110.
3. Repeat steps (1) and (2) an arbitrarily large number of times, say 5000 times.

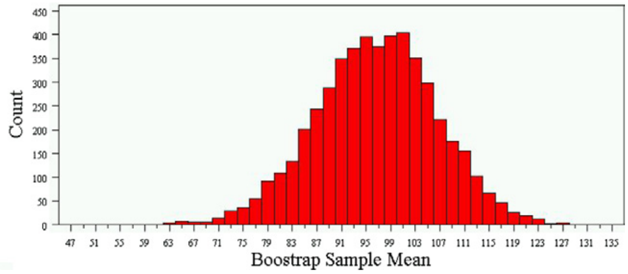

Figure 6. Empirical distribution of the means, from 5000 bootstrap samples from a right-skewed data set (N = 30).

4. Examine the distribution of the resulting 5000 means (see Figure 6).
5. Make inferences about the mean based on this observed distribution. For example, one can directly calculate the 95% confidence interval for the mean as the middle 95% of the observed 5000 means (from the 125th mean to the 4875th mean); this is: 77 to 116.

The bootstrap allows researchers to keep variables in the original units (rather than transforming them) and to calculate effect sizes.

CONCLUSION

Researchers should always understand the distributions of the variables in their data set. Summary measures that are appropriate for normal distributions may be misleading when applied to non-normal distributions, regardless of sample size. When it comes to statistical testing, normality is often less critical. The *t*-test, ANOVA, and linear regression may be inappropriate for small samples with extreme deviations from normality; in these cases, the researcher may opt to transform the data, run a nonparametric test, or perform a bootstrap analysis.

REFERENCES

1. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23:151-169.

2. Efron B. Bootstrap methods: Another look at the jackknife. Ann Stat 1979;7:1-26.