Explorations in statistics: the assumption of normality

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 8 May 2017; accepted in final form 26 June 2017.

Curran-Everett D. Explorations in statistics: the assumption of normality. Adv Physiol Educ 41: 449–453, 2017; doi:10.1152/advan.00064.2017.—Learning about statistics is a lot like learning about science: the learning is more meaningful if you can actively explore. This twelfth installment of Explorations in Statistics explores the assumption of normality, an assumption essential to the meaningful interpretation of a t test. Although the data themselves can be consistent with a normal distribution, they need not be. Instead, it is the theoretical distribution of the sample mean or the theoretical distribution of the difference between sample means that must be roughly normal. The most versatile approach to assess normality is to bootstrap the sample mean, the difference between sample means, or t itself. We can then assess whether the distributions of these bootstrap statistics are consistent with a normal distribution by studying their normal quantile plots. If we suspect that an inference we make from a t test may not be justified—if we suspect that the theoretical distribution of the sample mean or the theoretical distribution of the difference between sample means is not normal—then we can use a permutation method to analyze our data.

Procedures to Assess Normality

The most versatile approach to assess normality is to bootstrap the sample mean (for a one-sample t test) or to bootstrap the difference between sample means (for a two-sample t test) (6, 10). We can then assess whether the distribution of either bootstrap statistic is consistent with a normal distribution by generating a normal quantile plot (1, 6); see Fig. 3. We did this when we explored the bootstrap (Fig. 4) and permutation methods (Fig. 5). Why is this the most versatile approach? Because what we really care about is not the distribution as our sample size n is big enough (see Fig. 1).2 This more general way is a manifestation of the Central Limit Theorem (14, 19, 21).

Most populations have a mean μ and standard deviation σ. As we did in our inaugural exploration (3), suppose we draw from one of these populations an infinite number of samples, each with a-big-enough–n observations. The infinite number of sample means, \( \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n \), will be distributed normally with mean μ and standard deviation \( \sigma / \sqrt{n} \). In other words, the average of the sample means, Ave \( \{ \bar{y} \} \), will be the population mean μ, but the standard deviation of the sample means, SD \( \{ \bar{y} \} \), will be smaller than the population standard deviation σ by a factor of \( 1 / \sqrt{n} \) (Fig. 2).

With this brief overview, we are almost ready to begin our exploration of the assumption of normality. First, we need to review the software we will use to help us learn about it.

R: Basic Operations

The first paper in this series (3) summarized R (22) and outlined its installation. For this exploration there are three more steps: download Advances_Statistics_Code_Normal.R3 to your Advances folder, confirm you installed boot in our previous explorations (6, 15), and install the extra package nortest (16).4

To run R commands. If you use a Mac, highlight the commands you want to submit and then press \( \text{⌘}+\text{J} \) (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.

1 The results of a two-sample t test are meaningful only if the theoretical distribution of the difference between sample means is roughly normal (10). 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).

2 There are no hard-and-fast rules for how big big enough is. See Practical Considerations.

3 This file is posted as a Supplemental File under Figures & Data at the Advances in Physiology Education website.

4 Some of our previous explorations (6, 7, 10–12, 15) detail how to install an extra R package.
the observations themselves, but the theoretical distribution of the sample statistic: the sample mean or the difference between sample means. And we know that a bootstrap distribution estimates the theoretical distribution of some sample statistic (6).

We know also that if our sample is too small, then the bootstrap distribution is likely to differ—perhaps a lot—from the theoretical distribution of our sample statistic (6). If we suspect our sample is too small for the bootstrap to be useful, what recourse do we have? We can create a normal quantile plot of the observations themselves (Fig. 6). In this situation, however, our interpretation of the normal quantile plot should be tentative.

We can also assess normality using formal statistical tests (2, 16, 20). For each of these tests the implicit null and alternative hypotheses, \( H_0 \) and \( H_1 \), are:

- \( H_0 \): The sample observations are consistent with having come from a normal distribution.
- \( H_1 \): The sample observations are not consistent with having come from a normal distribution.

We can get a sense of these tests if we use the samples of 100 observations depicted in Fig. 3. Suppose we define our critical significance level \( \alpha \), the probability that we reject a true null hypothesis, to be 0.10 (5, 13). The observations drawn from the normal distribution are consistent with having come from a normal distribution: 0.50 \( P \) 0.67 (Table 1). In contrast, the observations drawn from the log-normal and uniform distributions are not consistent with having come from a normal distribution: 0.0001 \( P \) 0.08. These formal results reinforce our interpretations of the corresponding normal quantile plots (see Fig. 3).

We can also apply these statistical tests to two-sample problems (Table 2).

When we assess normality, it is important to use graphical techniques and formal statistical tests in tandem: the mere numeric result of a formal statistical test can mislead us (see Ref. 5). With a large number of observations, a statistical test may detect a small, inconsequential departure from normality. With a small number of observations, a statistical test may have insufficient power (8) to detect a meaningful departure from normality.
The assumption of normality is roughly normal (3–6, 10, 14, 21). This assumption that the theoretical distribution of the sample does not create this data graphic (adapted from Ref. 10). Advances in Physiology Education.

Practical Considerations

At the crux of meaningful results from a \( t \) test is the assumption that the theoretical distribution of the sample mean\(^5\) is roughly normal (3–6, 10, 14, 21). This assumption can be satisfied regardless of the distribution of the actual observations as long as our sample size is big enough. If our sample size is not big enough—if the Central Limit Theorem fails to hold—then we can transform the observations in the hopes that the theoretical distribution of the sample mean will be more normally distributed (see Fig. 4 and Ref. 6).

It is now time that we tackle the question, how big is big enough?

If our sample observations happen to be drawn from a population that is distributed normally, then any sample size is big enough for the theoretical distribution of the sample mean to be distributed normally (19, 21). See Fig. 7. If we analyze just three observations using a \( t \) test, then we are banking on this tenuous assumption.\(^6\) With three observations there is simply no way to know whether the underlying population is distributed normally (see Ref. 10).

If our sample observations are drawn from a population that is not distributed normally, then the answer to the question, how big is big enough?, is, it depends. The conventional rule-of-thumb is that a sample size of 30 is big enough for the theoretical distribution of the sample mean to be distributed roughly normally, even when the underlying population is skewed. But this convention is mere urban legend (17, 18, 21). When we explored the bootstrap, we discovered that a sample size of 40 was not big enough (see Ref. 6, Fig. 7).

In 1986 Moses wrote that a sample size might need to be quite large for the theoretical distribution of the sample mean to be distributed roughly normally—for the results of a \( t \) test to be meaningful—when the distribution of some underlying population was not normal (21). In 2014 Hesterberg, using simulations, confirmed that a sample size might need to exceed 5,000 observations when a population was even moderately skewed (17).

If we draw sample observations from a skewed population, the log-normal distribution in Fig. 3, we can see for ourselves that the empirical distribution of sample means is not normal, even for samples of 512 observations (Fig. 8).

\(^5\) Or the theoretical distribution of the difference between sample means.

\(^6\) This is also true if we compare two groups of three observations.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Normal</th>
<th>Log-Normal</th>
<th>Uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson-Darling</td>
<td>0.64</td>
<td>0.001(^-)</td>
<td>0.003</td>
</tr>
<tr>
<td>Cràmer-von Mises</td>
<td>0.67</td>
<td>0.001(^-)</td>
<td>0.01</td>
</tr>
<tr>
<td>Lilliefors</td>
<td>0.50</td>
<td>0.001(^-)</td>
<td>0.08</td>
</tr>
<tr>
<td>Shapiro-Francia</td>
<td>0.65</td>
<td>0.001(^-)</td>
<td>0.005</td>
</tr>
<tr>
<td>Shapiro-Wilk</td>
<td>0.62</td>
<td>0.001(^-)</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Values represent \( P \) values associated with procedures designed to assess the null hypothesis that sample observations are consistent with having come from a normal distribution. The sample observations are the 100 observations drawn from each distribution in Fig. 3. A value of 0.001\(^-\) denotes \( P < 0.001 \). The commands in lines 202–235 of Advances_Statistics_Code_Normal.R analyze the samples drawn from each distribution in Fig. 3. Your \( P \) values will differ slightly. To obtain the analyses summarized in this table, highlight and submit the lines of code from Table 1: first line to Figure 6: last line.
Table 2. **Statistical tests of normality: simulated data**

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Fish</th>
<th>Meat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lilliefors</td>
<td>0.85</td>
<td>0.02</td>
</tr>
<tr>
<td>Shapiro-Francia</td>
<td>0.81</td>
<td>0.03</td>
</tr>
<tr>
<td>Shapiro-Wilk</td>
<td>0.65</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Values represent *P* values associated with procedures designed to assess the separate null hypotheses that the 7 observations in the fish group and the 5 observations in the meat group (see Ref. 10, Table 2) are consistent with having come from a normal distribution. In their nortest algorithms, the Anderson-Darling and Cramér-von Mises procedures require at least 8 observations. The commands in lines 239–267 of Advances_Statistics_Code_Normal.R analyze the data in the fish and meat groups. To obtain these *P* values, highlight and submit the lines of code from Table 2: first line to Table 2: last line.

Because a bootstrap distribution of the sample mean can depart from normality in subtle ways, Hesterberg advocates bootstrapping *t* itself in addition to the sample mean (17, 18).

In our second exploration (5), we used *t* to assess whether our sample observations were consistent with having come from a population with mean *μ*. We calculated *t* as

\[ t = \frac{\bar{y} - \mu}{s/\sqrt{n}}, \]

where \( \bar{y} \) is the sample mean, *s* is the sample standard deviation, and *n* is the sample size, the number of observations in our sample.

We can generate a bootstrap distribution of *t* as

\[ t_{j} = \frac{\bar{y}_{j} - \bar{y}}{s_{j}/\sqrt{n}}, \]

where \( \bar{y}_{j} \) and \( s_{j} \) represent the mean and standard deviation of bootstrap sample *j* and where \( \bar{y} \) represents the original sample mean. Suppose we draw \( n = 1,024 \) sample observations from the log-normal distribution in Fig. 3 and then create \( j = 10,000 \) bootstrap samples. The bootstrap distribution of the sample mean is skewed, and the bootstrap distribution of *t* is even more so (Fig. 9). In the theoretical distribution of *t* with \( 1,024 - 1 \) degrees of freedom, 2.5% of the possible values of *t* are less than \(-1.962\), and 2.5% of the possible values of *t* are greater than \(+1.962\). In the skewed bootstrap distribution of *t* (see Fig. 9), 5.8% of the possible values of \( t_{j} \) are less than \(-1.962\), and 0.7% of the possible values of \( t_{j} \) are greater than \(+1.962\). The commands in lines 310–337 of Advances_Statistics_Code_Normal.R return these values. Your values will differ slightly.

---

**Fig. 7.** Normal quantile plots of 10,000 sample means computed from samples of 2, 32, or 512 observations drawn from the normal distribution in Fig. 3. For each sample size, the empirical distribution of the 10,000 sample means is consistent with a normal distribution. The commands in lines 271–306 of Advances_Statistics_Code_Normal.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.

**Fig. 8.** Normal quantile plots of 10,000 sample means computed from samples of 2, 4, 8, 16, 32, 64, 128, 256, or 512 observations drawn from the log-normal distribution in Fig. 3. For each sample size, the empirical distribution of the 10,000 sample means is inconsistent with a normal distribution. The commands in lines 310–337 of Advances_Statistics_Code_Normal.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.

**Fig. 9.** 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 1,024 observations taken from the log-normal distribution in Fig. 3. The bootstrap distributions of the sample mean and *t* statistic are skewed: this means an inference we make from a normal-theory hypothesis test or confidence interval is unjustified and likely to mislead us. The commands in lines 401–408 of Advances_Statistics_Code_Normal.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figure 9: first line to Figure 9: last line.
distribution of the sample mean is not normal—then we can use a permutation method in conjunction with the traditional test (10).\(^7\) If our conclusion from permutation matches our conclusion from the traditional test, then the normality assumption for the traditional procedure is at least reasonably well met. If our conclusion from permutation conflicts with our conclusion from the traditional test, then we should suspect that the normality assumption has not been met. We want to opt not for the statistical procedure that produces our dream result, but for the statistical procedure that has its assumptions best satisfied (10).

**Summary**

We knew already that the \(P\) value and confidence interval from a \(t\) test are meaningful only if the theoretical distribution of the sample mean—or the theoretical distribution of the difference between sample means—is roughly normal. As this exploration has demonstrated, the most versatile approach to assess normality is to bootstrap the sample mean, the difference between sample means—is roughly normal. As this version of manuscript.

**ACKNOWLEDGMENTS**

I thank Arindam Ganguly (Cornell University, Ithaca, NY) for suggesting this paper, and I thank Gerald DiBona (Göteborgs Universitet, Göteborg, Sweden and University of Iowa College of Medicine, Iowa City, IA), Arindam Ganguly, Andrew Greene (Medical College of Wisconsin, Milwaukee, WI), and Calvin Williams (Clemson University, Clemson, South Carolina) for their helpful comments and suggestions.

**DISCLOSURES**

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

**AUTHOR CONTRIBUTIONS**

D.C.-E. drafted manuscript; edited and revised manuscript; approved final version of manuscript.

---

\(^7\) Permutation is one kind of nonparametric procedure. Other nonparametric procedures include the sign test, the Wilcoxon signed-rank test, and the Wilcoxon rank-sum test (also known as the Mann-Whitney \(U\) or the Mann-Whitney-Wilcoxon test).

**REFERENCES**