ECONOMICS

Anesthesiology
2000; 92:1160–7

Modeling the Uncertainty of Surgical Procedure Times

Comparison of Log-normal and Normal Models

David P. Strum, M.D., Jerrold H. May, Ph.D,† Luis G. Vargas, Ph.D.‡*

Background: **Medical institutions are under increased economic pressure to schedule elective surgeries efficiently to contain the costs of surgical services. Surgical scheduling is complicated by variability inherent in the duration of surgical procedures. Modeling that variability, in turn, provides a mechanism to generate accurate time estimates. Accurate time estimates are important operationally to improve operating room utilization and strategically to identify surgeons, procedures, or patients whose duration of surgeries differ from what might be expected.**

Methods: **The authors retrospectively studied 40,076 surgical cases (1,580 Current Procedural Terminology–anesthesia combinations, each with a case frequency of five or more) from a large teaching hospital, and attempted to determine whether the distribution of surgical procedure times more closely fit a normal or a log-normal distribution. The authors tested goodness-of-fit to these data for both models using the Shapiro–Wilk test. Reasons, in practice, the Shapiro–Wilk test may reject the fit of a log-normal model when in fact it should be retained were also evaluated.**

Results: **The Shapiro–Wilk test indicates that the log-normal model is superior to the normal model for a large and diverse set of surgeries. Goodness-of-fit tests may falsely reject the log-normal model during certain conditions that include rounding errors in procedure times, large sample sizes, untrimmed outliers, and heterogeneous mixed populations of surgical procedure times.**

Conclusions: **The authors recommend use of the log-normal model for predicting surgical procedure times for Current Procedural Terminology–anesthesia combinations. The results help to legitimize the use of log transforms to normalize surgical procedure times before hypothesis testing using linear statistical models or other parametric statistical tests to investigate factors affecting the duration of surgeries. (Key words: log transformations; normal probability plots; Shapiro–Wilk tests.)**

MEDICAL institutions are under increased economic pressure to schedule elective surgeries efficiently to contain the costs of surgical services. Scheduling is complicated by variability inherent in the surgical procedures. A good statistical model for surgical procedure times is important for several purposes. It could be used retrospectively, together with existing methods,[1] to identify those surgeons or procedures whose surgical times are unusually slow or fast. Identifying these may allow surgical managers to eliminate sources of variability or, alternatively, to identify the outliers and to schedule them separately. In real-time surgical suite management, models for surgical times could be used to identify surgeries that are proceeding unexpectedly quickly or slowly, so that support personnel involved in subsequent care can be alerted that the patient will arrive much earlier or later than expected.[2] The most appropriate statistical distribution and accurate estimates of its parameters are also critical to surgical scheduling systems,[3] especially if surgical suite availability is limited.[4]

Modeling surgical procedure times has been of interest for at least 35 yr. For example, Rossiter and Reynolds[5] noted that waiting times visually appear to fit a log-normal distribution; others, including senior administrators we have observed, routinely assume that surgical times are normally distributed and provide summaries using typical parametric statistical tests (means ± SD). The log-normal distribution is one whose logarithms are normally distributed. It can take on values from zero to infinity and is skewed with a long right tail; therefore, it

* Associate Professor of Anesthesiology, Department of Anesthesiology, Queens University.

† Professor of Operations, Decision Sciences, and Artificial Intelligence and of Intelligent Systems, The Joseph M. Katz Graduate School of Business, University of Pittsburgh.

‡ Professor of Operations, Decision Sciences, and Artificial Intelligence, The Joseph M. Katz Graduate School of Business, University of Pittsburgh.

Received from the Department of Anesthesiology, Queens University, Kingston, Ontario, Canada, and Operations, Decision Sciences, and Artificial Intelligence, The Joseph M. Katz Graduate School of Business, the University of Pittsburgh, Pittsburgh, Pennsylvania. Submitted for publication May 13, 1999. Accepted for publication November 3, 1999. Supported by a grant from the Institute for Industrial Competitiveness of the Joseph M. Katz Graduate School of Business, Pittsburgh, Pennsylvania, and the Department of Anesthesia at the University of Arkansas for Medical Sciences, Little Rock, Arkansas.

Address reprint requests to Dr. Strum: Department of Anesthesiology, Queens University, KGH, 76 Stuart Street, Kingston, Ontario, K7L 2V7, Canada Address electronic mail to: strumd@post.queensu.ca

is attractive for time estimation in a surgical environment in which a small number of procedures may take much longer than average. The literature includes the use of the normal[6] and log-normal[7–9] distributions for modeling surgical durations. In the current study, we rigorously compare the normal and log-normal models based on goodness-of-fit tests. We also briefly discuss why, in practice, goodness-of-fit tests may reject a log-normal model of surgical times when in fact it should be retained.

Using the normal distribution where the log-normal is more appropriate can distort results derived from commonly used statistical tools. Linear statistical models, for example, assume that variability of the predicted variable is normally distributed. If, instead, it is log-normally distributed, then a log transform must be applied before the tool is used, or inferences based on the analysis may be biased. The inverse is also true; if a log transform is applied to data that really follows the normal distribution, then statistical analyses based on the transformed values will be misleading. Although the problem of model selection could be avoided by using nonparametric procedures, such as using Kruskal–Wallis instead of analysis of variance, those analyses are typically less powerful than comparable parametric ones when the data are normally distributed. In the current study, we fit surgical times to the log-normal model to help lay a foundation for use of log-normal transforms used to normalize surgical procedure times before statistical testing using linear statistical models.

Materials and Methods

We retrospectively reviewed all recorded surgical cases from a large teaching hospital performed over a 7-yr period from 1989–1995. Use of anonymous patient records was approved by the human subjects review committee of the institution that collected the data. Data were collected using a previously described computerized system.[10] We analyzed total procedure time (TT), the time from entry into the operating suite until emergence from anesthesia, and surgical procedure time (ST), the time from incision to closure of the surgical wound. Preliminary analyses[11,12] indicated that to obtain a better model fit, our data should be subdivided into more homogeneous subgroups by Current Procedural Terminology (CPT) code[13] in combination with anesthesia type (general, local, monitored, or regional), as opposed to being fitted by CPT alone.

Of 60,643 total case records in the initial database, 779 were omitted from analysis because of incomplete data, leaving 59,864 surgeries that included between one and three CPTs. There were 46,322 patients with only one CPT code (5,125 different CPT-anesthesia combinations), 10,740 with exactly two different CPT codes, and 2,802 patients with three CPT codes. To eliminate a potential confounding factor, we considered only surgical procedures with a single CPT code. We therefore confined our analysis to only 40,076 cases (1,580 CPT-anesthesia combinations), all with only one CPT code and each CPT-anesthesia combination with case frequencies of five or more (*i.e.*, enough to fit a probability distribution). We used goodness-of-fit tests to compare the fit of the normal and log-normal distributions to TT and ST for each of those 1,580 CPT-anesthesia combinations.

Statistics

We initially used Shapiro–Wilk[14] and Lilliefors[15] goodness-of-fit tests. Both determine whether a set of data are consistent with a normal distribution. Taking logs of the data before performing the test measures consistency with a log-normal distribution because the logs of log-normal data are normally distributed. Using the Shapiro–Wilk and the Lilliefors tests, the null hypothesis is that the model distribution fits the data; a large P value indicates that the data fit the model well. Because the Shapiro–Wilk test is considered the single best general purpose test of normality,[16] we report its results in preference to the Lilliefors results that we reported elsewhere.[17,18]

To perform the Shapiro–Wilk tests, we used the International Mathematical and Statistical Libraries Fortran routine SPWILK (Visual Numerics, Inc., Houston, TX). We used normal probability plots to examine those CPT-anesthesia combinations that were not well-fitted by either the normal or log-normal models. We used Friedman tests to compare Shapiro–Wilk P values (goodness-of-fit) for the log-normal *versus* normal model for ST and TT. The Friedman test is the nonparametric equivalent of a paired t test. We also compared ST and TT for modeling procedure times using both the normal and the log-normal models.

Determining the Model

The log-normal model is a random variable whose logarithms are normally distributed. The shape of the log-normal, like that of the normal, is given by two parameters: The mean and variance. An additional third

Table 1. Tabular Comparisons of Sample Size and Shapiro-Wilk Goodness-of-fit *P* Values for the Lognormal Model for Surgical Procedure Time and Total Procedure Time

Category	$P < 0.01$	$0.01 \leq P < 0.1$	$P \geq 0.1$	Row Totals
Surgical procedure time				
Small ($n < 30$)	49 (3.10)	170 (10.8)	1105 (70.0)	1324 (83.8)
Medium ($30 \leq n \leq 200$)	45 (2.85)	28 (1.77)	159 (10.1)	232 (14.7)
Large ($n > 200$)	12 (0.76)	2 (0.13)	10 (0.63)	24 (1.52)
Column totals	107 (6.71)	200 (12.7)	1274 (80.6)	1580 (100)
Total procedure time				
Small ($n < 30$)	28 (1.77)	161 (10.2)	1135 (71.8)	1324 (83.8)
Medium ($30 \leq n \leq 200$)	24 (1.52)	41 (2.59)	167 (10.6)	232 (14.7)
Large ($n > 200$)	10 (0.63)	7 (0.44)	7 (0.44)	24 (1.52)
Column totals	62 (3.92)	209 (13.2)	1309 (82.9)	1580 (100)

Data are number of procedures (% total); n = 40,076 cases, 1,580 CPT-anesthesia combinations.

CPT = current procedural terminology.

parameter, termed the location parameter, is the amount by which its minimum value is shifted away from the origin. In this article, we assumed the shift to be zero and fit the two-parameter model that provides probability estimates for all intervals between zero and positive infinity. Surgical procedures often have a large minimum duration; therefore, it may be very desirable to fit a more general log-normal model to them. Doing so, however, necessitates estimation of the third parameter, which in practice may be a difficult problem. Issues related to general log-normal modeling of surgical times are discussed elsewhere[19-21]; in the current study, we investigated only the two-parameter log-normal model.

To examine goodness-of-fit tests for the normal and log-normal models, we cross-tabulated the Shapiro-Wilk test results for all CPT-anesthesia combinations by sample size and *P* value of the Shapiro-Wilk tests. The results are shown in tables 1-3. To detect the influence of sample size on the Shapiro-Wilk tests, we divided the sample size arbitrarily into small ($n < 30$), medium ($n = 30$-200), and large ($n > 200$) categories. Because commonly used levels of significance for hypothesis testing are between 1 and 10%, a frequently used rule of thumb is to regard a *P* value of at least 0.10 as leading to the retention of the null hypothesis (the model fits well) and a *P* value < 0.01 as always leading to its rejection (the model fits poorly). We interpreted *P* values between 0.01 and 0.1 as a mediocre fit for the model. Table 1 shows how well the data fit the log-normal model as a function of sample size, and table 2 does the same for the normal model.

We compared the overall performance of the log-normal and normal models using qualitative (tabular comparisons) and quantitative interpretations (Friedman tests). To determine whether the log-normal performed better than the normal on some CPT-anesthesia combinations, we compared the performance of the two models on the same data sets (table 3). We also used Friedman tests to compare the goodness-of-fit between the log-normal and normal model for ST and TT (table 4).

Table 2. Tabular Comparisons of Sample Size and Shapiro-Wilk Goodness-of-fit *P* Values for the Normal Distribution for Surgical Procedure Time and Total Procedure Time

Category	$P < 0.01$	$0.01 \leq P < 0.1$	$P \geq 0.1$	Totals
Surgical procedure time				
Small ($n < 30$)	148 (9.37)	258 (16.3)	918 (58.1)	1324 (83.8)
Medium ($30 \leq n \leq 200$)	140 (8.86)	33 (2.09)	59 (3.73)	232 (14.7)
Large ($n > 200$)	21 (1.33)	1 (0.06)	2 (0.13)	24 (1.52)
Column totals	309 (19.6)	292 (18.4)	979 (62.0)	1580 (100)
Total procedure time				
Small ($n < 30$)	107 (6.77)	243 (15.38)	974 (61.65)	1324 (83.80)
Medium ($30 \leq n \leq 200$)	125 (7.91)	43 (2.72)	64 (4.05)	232 (14.68)
Large ($n > 200$)	22 (1.39)	0 (0.00)	2 (0.13)	24 (1.52)
Column totals	254 (16.08)	286 (18.10)	1040 (65.82)	1580 (100)

Data are number of procedures (% total); n = 40,076 cases, 1,580 CPT-anesthesia combinations.

CPT = current procedural terminology.

Table 3. Tabular Comparisons of the Lognormal (Ln) and Normal (N) Models, as Measured by Shapiro-Wilk Goodness-of-fit Test *P* Values for Surgical Procedure Time and Total Procedure Time

	Normal Models			
Category	$P < 0.01$	$0.01 \leq P < 0.1$	$P \geq 0.1$	Totals
Surgical procedure time				
Ln ($p < 0.01$)	40 (2.53)	21 (1.33)	45 (2.85)	106 (6.71)
Ln ($0.01 \leq p < 0.1$)	58 (3.67)	54 (3.42)	88 (5.57)	200 (12.7)
Ln ($p \geq 0.1$)	211 (13.3)	217 (13.7)	846 (53.5)	1274 (80.6)
Column totals	309 (19.6)	292 (18.5)	979 (70.0)	1580 (100)
Total procedure time				
Ln ($p < 0.01$)	30 (1.90)	14 (0.89)	18 (1.14)	62 (3.92)
Ln ($0.01 \leq p < 0.1$)	93 (5.89)	60 (3.80)	56 (3.54)	209 (13.2)
Ln ($p \geq 0.1$)	131 (8.29)	212 (13.4)	966 (61.1)	1309 (82.9)
Column totals	254 (16.1)	286 (18.1)	1040 (65.8)	1580 (100)

Data are number of procedures (% total); n = 40,076 cases, 1,580 CPT-anesthesia combinations.

CPT = current procedural terminology; Ln = lognormal.

Limitations of the Goodness-of-fit Results

Examination of failed model fits using normal probability plots revealed that the log-normal model was sometimes rejected inappropriately because of rounding of the surgical procedure times. We investigated briefly why rounding of these values should lead to rejection of the log-normal model when it fact it should have been accepted.

In practice, surgical times are observed and recorded only to a convenient level of precision. Pearson *et al.*[22] observed that if the unit of measure to which observations are rounded is large relative to the variability of the data, goodness-of-fit tests on normal data may be unreliable. To investigate the effect of rounding errors on goodness-of-fit tests applied to log-normal data, we generated a representative normal sample with 50 values and then exponentiated those values to produce a log-normally distributed series with a mean of 30 min and a SD of 10 min. We subsequently rounded these log-normally distributed values to the nearest 1, 3, 5, 10, and 15 min. This resulted in five log-normally distributed series that differed only in the interval to which series values were rounded. Values from each series were then log transformed, and Shapiro-Wilk and Lilliefors tests were applied to the values. The *P* values for those tests were plotted against the rounding interval and compared. We also used normal probability plots to illustrate graphically the effect of progressively rounding the series values. Finally, to assure that adjusting the rounding interval did not alter the mean of any of the series, each of the rounded data series was compared with the original series using paired Friedman tests.

Results

Table 1 displays the results of fitting cases involving exactly one CPT-anesthesia combination to the log-normal model. The ST data fit the log-normal distribution for > 93% of total CPT-anesthesia combinations. We noticed a decrease in the proportion of combinations that fit the log-normal model as the sample size increased. ST fitted the log-normal distribution well ($P = 0.1$) for 87% (1,105 CPT-anesthesia combinations) of

Table 4. Friedman Test Results for Paired Comparisons of the Shapiro-Wilk Goodness-of-fit Test *P* Values for the Lognormal and Normal Models for Surgical and Total Procedure Times

Hypotheses	Ln TT = Ln ST	N TT = N ST	Ln ST = N ST	Ln TT = N TT
Friedman test statistic	4.90	554	155	228
Kendall coefficient of concordance	0.003	0.351	0.098	0.144
Rank sum	2414 (Ln TT)	2838 (N TT)	2617 (Ln ST)	2670 (Ln TT)
Rank sum	2326 (Ln ST)	1902 (N ST)	2122 (N ST)	2070 (N TT)
P value	≤0.027	<0.000	<0.000	<0.000

n = 1,580 different CPT-anesthesia combinations with 1 degree of freedom. TT fitted both the lognormal and normal models better than ST; the lognormal model fits both ST and TT better than the corresponding normal model.

CPT = current procedural terminology; Ln = lognormal; TT = total procedure time; ST = surgical procedure time; N = normal.

small-sized samples, for 12% (159 CPT-anesthesia combinations) of medium-sized samples, and for < 1%[10] of large-sized samples. The results for TT were similar to those for ST. We return to possible reasons for the decreased proportions of goodness-of-fit tests later.

Table 2 cross-tabulates analogous results for the normal distribution. The ST data fit the normal distribution for about 80% of CPT-anesthesia combinations. Similar to the log-normal model, there is a decrease in the proportion of combinations for ST that fit the normal as the sample size increases. ST fits the normal model well ($P = 0.1$) for 94% of small-sized samples, for 6% of medium-sized samples, and for < 1% of large-sized samples (n > 200). The results for TT were similar to those for ST.

To determine situations in which the normal distribution fit but the log-normal did not, we compared Shapiro-Wilk results for the two models using tabulations. In table 3, we cross-tabulated the P values for the log-normal and the normal Shapiro-Wilk tests on the same CPT-anesthesia combinations. From the first row of table 3, there are only 4% of samples that fail to fit the log-normal ($P < 0.01$) but fit the normal at least a mediocre level ($P = 0.01$). The second row shows that there are only 88 cases in which the fit to the normal is superior to that to the log-normal. Conversely, the first column shows that the data fit the log-normal well in 211 of the 309 cases that fail to fit the normal, and the data fits the log-normal well in 217 of the 292 cases in which the fit to the normal is mediocre. The frequencies below the diagonal, in which the fit to the log-normal is superior to that of the normal, dominate the corresponding frequencies above the diagonal, in which the fit to the normal is superior to that of the log-normal.

Table 4 is a paired comparison of the log-normal and normal models using Friedman tests. The log-normal model was superior to the normal for modeling ST and TT. These two comparisons are the quantitative equivalent of the tabular comparisons in table 3. In addition, persons wishing to model surgeries will discover TT is better estimated than ST when modeling using both the log-normal and the normal distributions.

To contrast the appearance of actual data with the ideal models, we compared the actual (empirical) data (illustrated as a frequency histogram) with the fitted log-normal and normal models for a single CPT-anesthesia combination. The procedure illustrated (fig. 1) was exploration of the chest for postoperative bleeding during general anesthesia. To show that the actual data differ from estimates using the normal and log-normal models, we compared estimates for the log-normal and normal models for values of the tenth, thirtieth, fiftieth, seventieth, and nintieth percentiles of the actual (empirical) data for the same CPT-anesthesia combination. The 50th percentile estimate for the normal model is both the mean and the median, whereas for the log-normal model it is only the median. The log-normal model produced estimates closer to the actual data than the normal model for all percentiles computed (table 5).

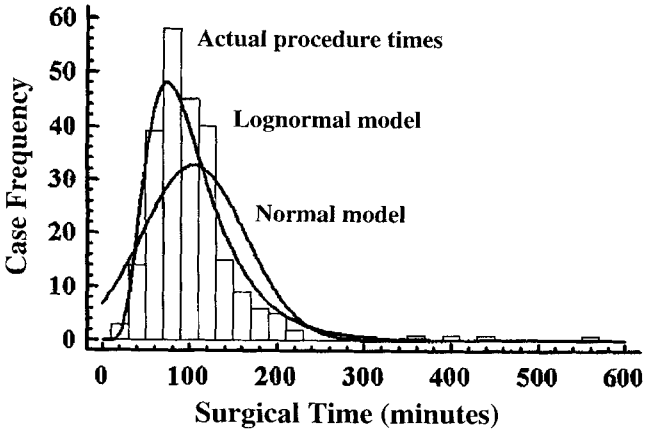

Fig. 1. To illustrate the differences between models, we compared the actual (empirical) data (illustrated as a frequency histogram) with the fitted ideal log-normal and normal models (the surgical procedure is exploration of the chest for postoperative bleeding during general anesthesia; n = 241 surgical cases).

Limitations of Goodness-of-fit Tests

The Shapiro-Wilk and Lilliefors goodness-of-fit tests were similarly affected by increased rounding of the log-transformed values (fig. 2). Using a rounding interval of 1 min, both tests accepted the hypothesis that the data fit the distribution. As the rounding intervals increased, the P value for both tests decreased, but the Lilliefors test proved more likely to reject the log-normal model incorrectly, as evidenced by the corresponding normal probability plots. Using rounding intervals of > 3 min, the Lilliefors test incorrectly rejected a fit to the log-normal each time ($P < 0.01$). Conversely, the Shapiro-Wilk test proved less sensitive and did not incorrectly reject the fit to the log-normal model until the rounding interval increased to 10 min ($P < 0.01$). None of the sample means of the five rounded series differed from the initial (unrounded) series (Friedman tests, $P < 0.00$).

Normal probability plots of simulated data series indicated that smooth probability plots become lumpy and exhibit what we termed a "stepping" phenomenon (fig. 3) as series values were progressively rounded from 1 to

Table 5. Surgical Times (min) From the Actual Data (empirical distribution) Are Compared With Estimates (min) Using Various Percentiles of the Lognormal and Normal Models

Percentile	Actual Data	Lognormal Model	Normal Model
10th	59	51	29
30th	75	75	74
50th	95	95	105
70th	120	120	135
90th	155	168	180

Note the 50th percentile estimate for the normal model is the mean and for the lognormal model; it is the median (n = 241 cases, the procedure modeled is exploration of the chest for postoperative bleeding with general anesthesia).

15 min. Figure 3A appears to be a smooth plot, whereas in contrast, figure 3B illustrates the "stepping" phenomenon caused by the accumulation of duplicate values in the rounded series (the two identical series differed only by their rounding intervals).

Discussion

Choosing the most appropriate model is an important step in forecasting surgical procedure times. For cases with only one combination of surgical procedure and anesthesia, our research indicates that the log-normal model is significantly better than the normal.

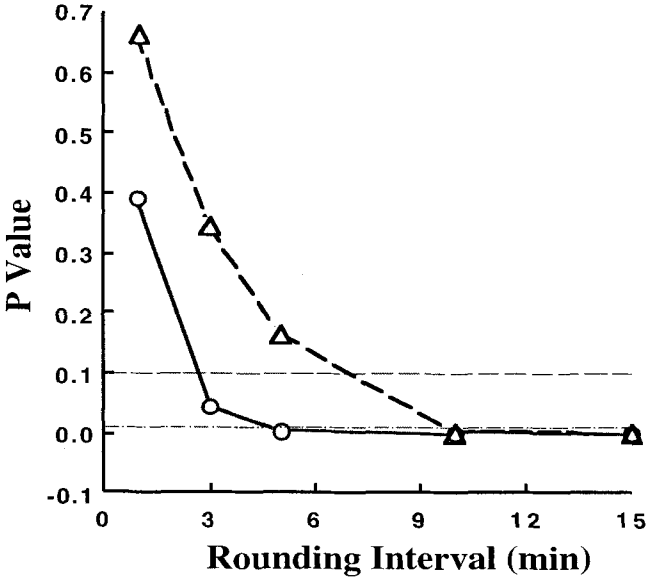

Fig. 2. Rounding interval influenced the goodness-of-fit *P* value for two common tests of normality: the Lilliefors test (open circle) and Shapiro–Wilk test (open triangle). The test data series was an ideal log-normal distribution (n = 50; mean = 30 min; SD = 10 min) rounded to the nearest 1-, 3-, 5-, 10-, or 15-min interval and log transformed before testing.

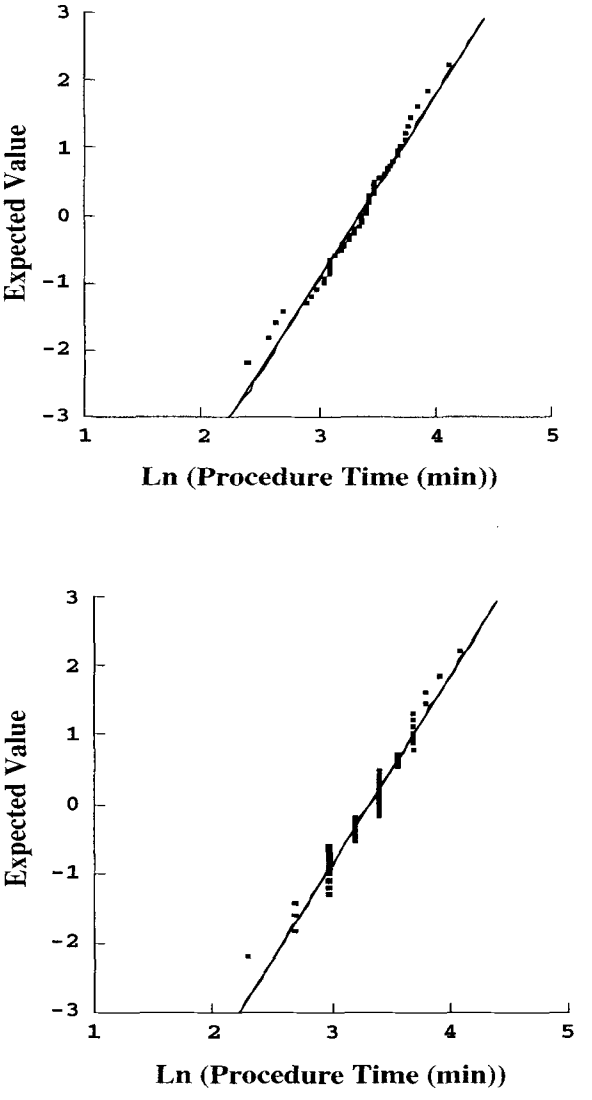

Fig. 3. Normal probability plots of two identical data series derived from the same ideal log-normal distribution (n = 50; mean = 30 min; SD = 10 min). Compare the smooth plot in *A* (values rounded to the nearest 3 min) with the "lumpy" appearance of the plot in *B* (values rounded to the nearest 5 min). (*B*) Illustrates a phenomenon termed "stepping." Stepping is increased with larger rounding intervals. Ln = the natural logarithm.

To schedule surgical procedures, we must have a statistical model that accounts for the variability inherent in surgical times. We previously showed how prevailing cost structures could determine the percentile point of the time model used to allocate surgical subspecialty block times.[23] In an analogous procedure, a similar minimal cost analysis can be used to allocate time to a single surgical procedure. The theory and application are sim-

ilar, except that single procedures are better modeled by the log-normal distribution. If overtime costs are 1.5 times regular, then during ideal conditions, the 60th percentile point estimate of the time model should be chosen to minimize the cost of scheduling the procedure. Different point estimates may be chosen for different cost structures, and fitting a statistical model to surgical procedure times is the best way to obtain these. It should be noted, however, that this analogy may not apply to the scheduling of consecutive surgical procedures. That is a more complicated problem and thus necessitates additional research.

We believe that clinician managers will increasingly explore surgical data to discover methods to improve the efficiency of surgical services. Linear statistical models may be used to explore this data looking for factors that increase variability in procedures and that can be either manipulated or alternatively scheduled. If the predicted measure is log-normal, then one should be careful to use such techniques to forecast and model logs of times rather than the times themselves. The current study, and additional studies,[11,24,25] lays a foundation intended to help legitimize log-normal transformations as tools used for the exploration of surgical procedure times.

Limitations of Goodness-of-fit Tests

The Shapiro–Wilk tests rejected the log-normal model for almost 7% of total CPT-anesthesia combinations tested. Examination of rejected combinations using normal probability plots revealed that most were indeed compatible with a log-normal model. We briefly discuss the variety of reasons why goodness-of-fit tests sometimes rejected the log-normal model when in fact it should have been accepted.

Goodness-of-fit tests may inappropriately reject preferred models in a variety of known circumstances if they are the only tools used for model selection.[14–16,22,26,27] The Shapiro–Wilk test was intended by its creators to be used in conjunction with graphical methods such as normal probability plotting, not as a substitute for them.[28] In practice, selection of a model should be based on an examination of a normal probability plot (a possibly subjective procedure) in conjunction with a formal goodness-of-fit test (a more objective measure).

Some samples that at first appear not to fit a model can be fit after a simple transformation of the measured values.[26] Potential additional causes of poor fits include rounding of shorter procedures times, large sample sizes, untrimmed outliers, and failure to properly segment sample mixtures.

Rounded procedure times may distort the results of goodness-of-fit tests. Anesthesia records record time to the nearest 5 min, and some information systems record times only to the nearest 15 or 30 min. The larger the rounding interval, the greater the likelihood that the modeled distribution will be rejected by goodness-of-fit tests. Pearson *et al.*[22] studied the sensitivity of several tests of normality, including the Shapiro–Wilk test, to rounding and numeric ties. If the ratio of the sample SD to the rounding interval was large (≥ 10) for a sample of data, he found the Shapiro–Wilk test result was little affected. However, if the ratio was small (2–3), the test was significantly less reliable. If data are recorded to the nearest 15 min, for example, the goodness-of-fit test results cannot be relied on unless the SD of the values is at least 150 min. Our experiment confirms that rounding is also an important factor when testing log-transformed data for goodness-of-fit; however, additional research is necessary to find the range of values for which goodness-of-fit tests are reliable when applied to log-normal data.

Outliers, or extreme values in tails of a distribution because of nonrecurring factors, also affect goodness-of-fit tests. Tukey[29] proposed deleting 2 to 5% of observations equally from both ends of all larger samples when contamination of the sample by outliers is suspected. Conversely, deleting valid data points can lead to acceptance of a model fit when it should be rejected. We did not trim the tails of our data sets because we had no information that would support doing so.

Common goodness-of-fit tests are designed for small samples (n = 100). The sensitivity of the tests increases roughly linearly with the log of the sample size,[14] so that stricter standards are applied to larger samples than to smaller ones. This effect might contribute to the decreasing *P* values we observed for larger samples.

A goodness-of-fit test compares a set of data to a single model and is meaningful only if the data set is homogeneous except for a random component. If, instead, a data set contains values from two different populations, the data should be separated into two homogeneous groups before model fitting because goodness-of-fit tests are not designed to detect heterogeneity. We separated data by CPT code and anesthesia. For our data, it was possible that a single CPT-anesthesia combination included cases from several surgeons, one of whom operates much more quickly than the others (a heterogeneous population). If we had segmented such heterogeneous CPT-anesthesia combinations by additional factors such as

surgeon, the fit to the log-normal model might have been even better.

Our analysis of procedure times from a large and diverse surgical database leads to the following conclusions. Procedure times (surgical time and total time) fit the log-normal distribution significantly better than they do the normal. Percentile estimates derived using the log-normal model may be very different from those derived from the normal. The Shapiro-Wilk goodness-of-fit test results should be confirmed using normal probability plots before rejection of an expected model if the sample size is large, short procedures times are rounded, the sample is a mixture of two or more populations, or outliers are present. In addition to statistical model fitting, our research implies that statistical tools such as regression and analysis of variance should be applied to log transforms of the procedure times, as opposed to being applied to the untransformed times.

The authors wish to thank Dr. Gerard Bashein for his assistance with this article.

References

1. Barnett V, Lewis T: Outliers in Statistical Data, 3rd edition. New York, John Wiley & Sons, 1994

2. Strum DP, Vargas LG, May JH: Design Of RCSS: Resource coordination systems for surgical services using distributed communications. J Am Med Inform Assoc 1997; 4:125-35

3. Magerlain JM, Martin JB: Surgical demand scheduling: A review. Health Serv Res 1978; 13:418-33

4. Gerchak Y, Gupta D, Henig M: Reservation planning for elective surgery under uncertain demand for emergency surgery. Management Science 1996; 42:321-34

5. Rossiter CE, Reynolds JA: Automatic monitoring of the time waited in out-patient departments. Med Care 1963; 1:218-25

6. Barnoon S, Wolfe H: Scheduling a multiple operating room system: a simulation approach. Health Serv Res 1968; 3:272-85

7. Hancock WM, Walter PF, More RA, Glick ND: Operating room scheduling data base analysis for scheduling. J Med Syst 1988; 12:397-409

8. Robb DJ, Silver EA: Scheduling in a management context: Uncertain processing times and non-regular performance measures. Decision Sciences 1996; 6:1085-106

9. Zhou J, Dexter F: Method to assist in the scheduling of add-on surgical cases: Upper prediction bounds for surgical case durations based on the log-normal distribution. ANESTHESIOLOGY 1998; 89:1228-32

10. Bashein G, Barna C: A comprehensive computer system for anesthetic record retrieval. Anesth Analg 1985; 64:425-31

11. Strum DP, May JH, Sampson AR, Vargas LG: Surgeon and type of anesthesia affect the scheduling of surgical procedures (abstract). ANESTHESIOLOGY 1997; 87:A1019

12. Strum DP, Sampson AR, May JH, Vargas LG: Type of anesthesia affects the duration and scheduling of surgical procedures (abstract). Anesth Analg 1999; 88:S47

13. Kirschner CG, Burkett RC, Marcinowski D, Kotowicz GM, Leoni G, Malone Y, O'Heron M, O'Hara KE, Scholten KR, Willard DM: Physicians Current Procedural Terminology 1995, 4th edition. Chicago, American Medical Association, 1995

14. Shapiro SS, Wilk MB, Chen HJ: A comparative study of various tests for normality. J Am Stat Assoc 1968; 63:1343-72

15. Lilliefors HW: On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J Am Stat Assoc 1967; 64:399-402

16. D'Agostino RB: Tests for the normal distribution, Goodness-of-fit Techniques. Edited by D'Agostino RB, Stephens MA. New York, Marcel Dekker, 1986, pp 367-419

17. Strum DP, May JH, Vargas LG: Surgical procedure times are well modeled by the lognormal distribution. Anesth Analg 1998; 86:S47

18. Strum DP, May JH, Vargas LG: Lognormal distributions are a best fit model for scheduling surgical procedure times (abstract). Anesth Analg 1997; 84:S56

19. Johnson NL, Kotz S, Balakrishnan N: Continuous Univariate Distributions, 2nd edition. New York, Wiley, 1994

20. Muralidhar K, Zanakis SH: A simple minimum-bias percentile estimator of the location parameter for the gamma, weibull, and log-normal distributions. Decision Sciences 1992; 23:862-79

21. May JH, Strum DP, Vargas LG: Fitting the lognormal distribution to surgical procedure times. Decision Sci 2000; in press

22. Pearson ES, D'Agostino RB, Bowman KO: Tests for departure from normality: Comparison of powers. Biometrika 1977; 64:231-46

23. Strum DP, Vargas LG, May JH: Surgical subspecialty block utilization and capacity planning: A minimal cost analysis model. ANESTHESIOLOGY 1999; 90:1176-85

24. Strum DP, May JH, Sampson AR, Vargas LG: Variability amongst surgeons increases with duration of the surgical procedure (abstract). ANESTHESIOLOGY 1997; 87:A1002

25. Strum DP, Vargas LG, Sampson AR, May JH: Individual surgeon variability is a multiplicative function of surgical time (abstract). Anesth Analg 1999; 88:S48

26. Bliss CI: Provisionally normal distributions, Statistics in Biology. New York, McGraw-Hill, 1967, pp 152-85

27. Madansky A: Testing for normality, Prescriptions for Working Statisticians. New York, Springer-Verlag, 1988, pp 15-31

28. Shapiro SS, Wilk MB: An analysis of variance test for normality (complete samples). Biometrka 1965; 52:591-611

29. Tukey JW: A survey of sampling from contaminated distributions, Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling. Stanford, Stanford University Press, 1960, pp 448-85