12-2020

Concerns in ID'ing a Suitable Distribution

Necip Doganaksoy
Siena College School of Business

Gerald J. Hahn

William Q. Meeker
Iowa State University, wqmeeker@iastate.edu

Concerns in ID'ing a Suitable Distribution

Abstract

Analysis of product lifetime data generally requires fitting a suitable distribution to the data at hand. The fitted distribution is used to estimate quantities of interest, such as the fraction of product failing after various times in service and selected distribution percentiles (for example, the estimated time by which 1% of the product population is expected to fail).

Disciplines

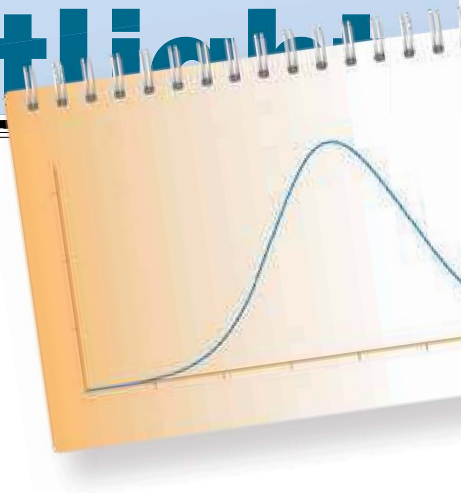
Probability | Statistical Methodology

Comments

This article is published as Doganaksoy, N., Hahn, G.J., and Meeker, W.Q., (2020), Concerns in ID'ing a Suitable Distribution, *Quality Progress*, 53, December, 72–75. Posted with permission.

Solving quality quandaries through statistics

Statistics Spotlight



DATA ANALYSIS

Concerns in ID'ing a Suitable Distribution

Product lifetimes are typically not normally distributed

by Necip Doganaksoy, Gerald J. Hahn and William Q. Meeker

Analysis of product lifetime data generally requires fitting a suitable distribution to the data at hand. The fitted distribution is used to estimate quantities of interest, such as the fraction of product failing after various times in service and selected distribution percentiles (for example, the estimated time by which 1% of the product population is expected to fail).

Many phenomena encountered in practice follow a normal distribution and, therefore, the data can be analyzed using well-known methods for statistical estimation. In fact, the normal distribution is the proverbial workhorse of statistical analysis and plays a prominent role in elementary statistics courses. Thus, it is sometimes fondly referred to as that "old familiar bell-shaped curve." See Figure 1.

Unfortunately, product lifetime data is not one of the "many phenomena" referred to above, and typically do not follow a normal distribution. The histogram in Figure 2, for example, displays the distribution of insulating fluid breakdown times for a sample of 19 test electrodes.[1] (The test was conducted at a high voltage to obtain failure information quickly.) The histogram clearly shows that a normal distribution is not a suitable model for oil breakdown times. Also, the data were plotted on normal distribution probability paper (see Figure 3, p. 74). This shows the plotted points diverging appreciably from a straight line—again suggesting that the data cannot be described by a normal distribution.

This column expands on what one of us wrote in his youth.[2] We remind readers of the theoretical justification for the normal distribution as a model for the distribution of many phenomena, provide a demonstration, and point out a misconception that we encountered. We explain why the theoretical justification usually does not hold for product lifetime. We conclude by briefly reviewing how you typically proceed when analyzing lifetime data.

Justification for normal distribution

The theoretical justification for the normal distribution as a model is the so-called central limit theorem (CLT). The CLT asserts that outcomes that are the sum of many small effects can be described by a normal distribution. Some examples of phenomena which you would expect as a consequence of the CLT to be normally distributed are:

+ The stack height of a motor (that is, the sum of the heights of the individual laminations that make up the stack).
+ The weekly sales of a business that sells many low-priced products—after the effects of trends and seasonal variations have been eliminated.
+ The number of separate lost-time accidents per year in a large industrial plant.
+ The total time required to check out a system that comprises several individual stages operating independently of each other.

iStock.com/kolotuschenko

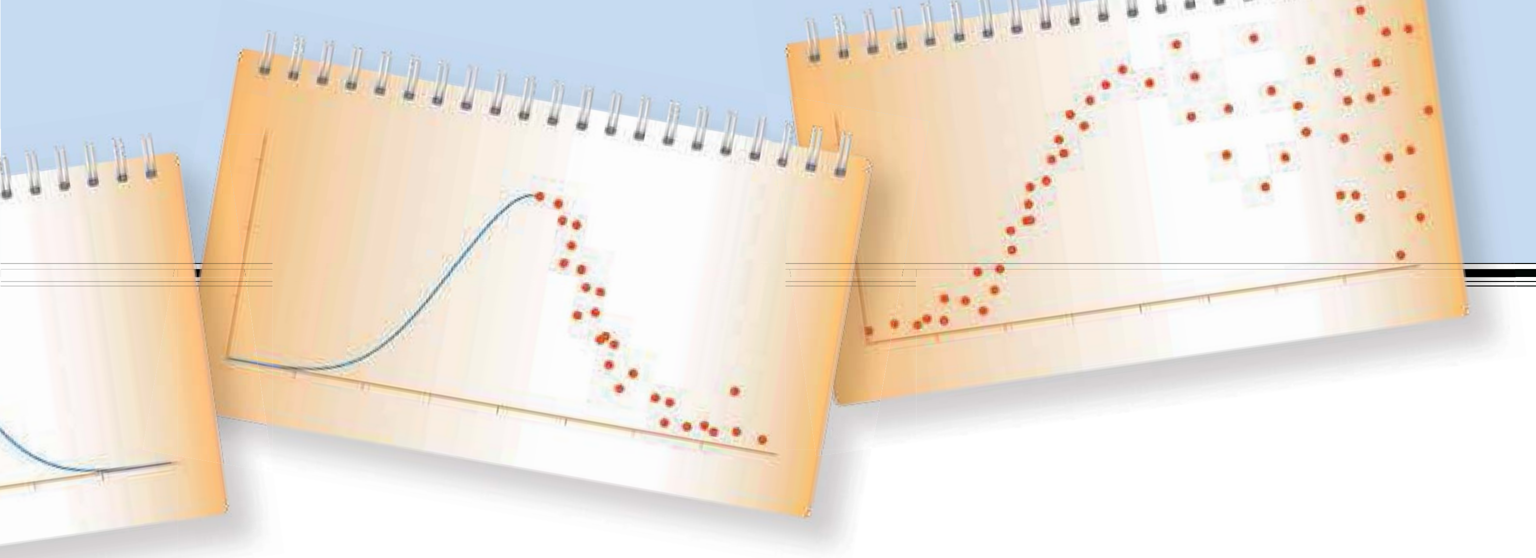


FIGURE 1

The standard normal distribution probability density function

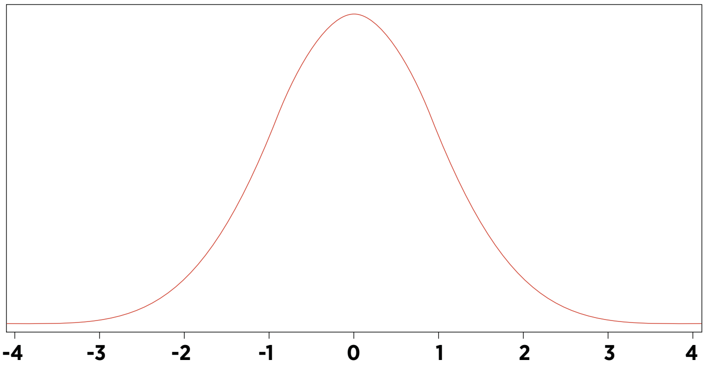
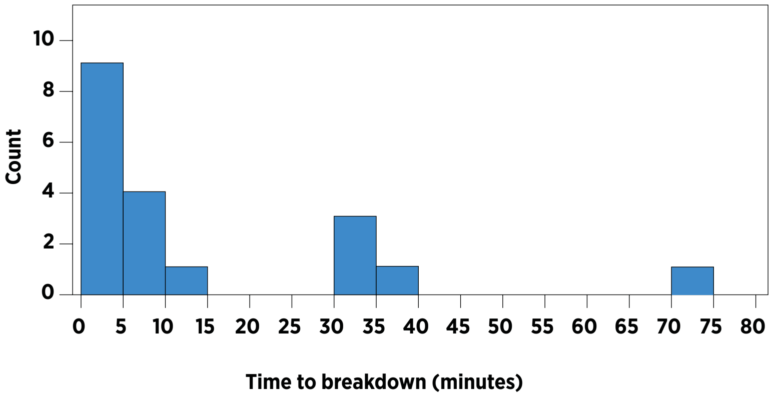


FIGURE 2

Insulating fluid breakdown times for a sample of 19 test electrodes



Other variables that cannot be explicitly expressed as the sum, or mean, of a large number of individual variables—but whose values, nevertheless, reflect such a sum—are likely to be approximately normally distributed. Examples are the heights of American adult males, molecular velocities of a gas, scores on an intelligence test, the dimensions of parts from a manufacturing process, and random electrical noise. Instrumentation and measurement errors, also, are frequently normally distributed.

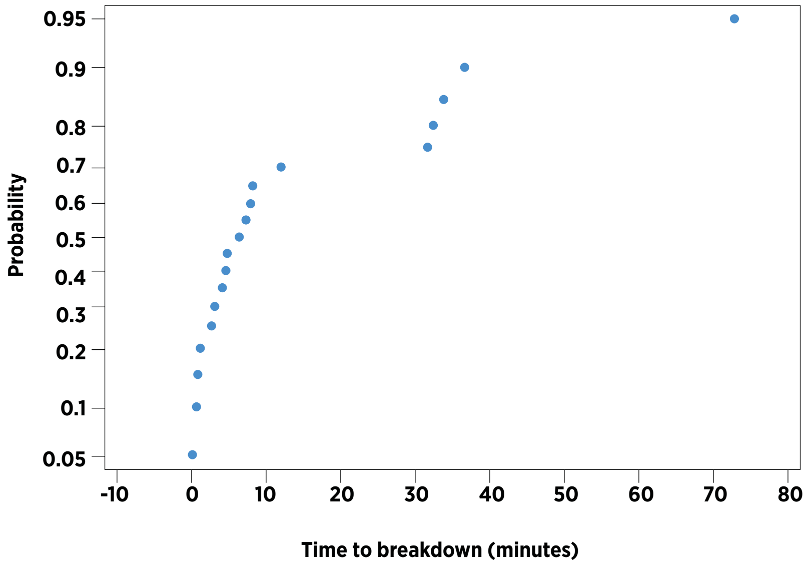
Demonstration

Here's an example that statisticians have used to demonstrate the CLT to a class of students:[3]

- Each student is asked to write down the day in the month of her birthdate and those of her two closest relatives. The histogram of 90 birthdates in five-day intervals obtained by one of the authors from a class of 30 students is shown in Figure 4 (p. 75). This clearly does not resemble a bell-shaped curve. In fact, the underlying distribution is a uniform distribution with each of the intervals having an approximately equal number of observations.
- Each student is now asked to average the birthdates of her two closest relatives (excluding her own birthdate). The results are displayed in five-day intervals for the entire class in Figure 5 (p. 75). This histogram shows, as expected, a higher frequency in the more central intervals than in the extremes.
- Finally, each student calculates the average of all of her three birthdates, now including her own, and reports her results. These are plotted in five-day intervals in Online Figure 1, which can be found on this column's webpage at qualityprogress.com. Now you can see the approximation of a bell-shaped curve emerging. Moreover, you would expect from the CLT that the approach to normality would become more

FIGURE 3

Normal distribution probability plot for insulating fluid breakdown times



and more evident as the sample on which each average is based is increased—as could be demonstrated by continuing the demonstration with added birthdates, beyond the three used here, for each student to include in her calculated average.

A misconception

The CLT does not suggest that with a large sample the variable being studied will somehow become normally distributed. We have, nevertheless, seen the suggestion that, on account of the CLT, you can assume a normal distribution in making inferences about an underlying distribution even in dealing with estimates that do not involve an averaging process—such as estimating a distribution percentile or the probability of product lifetime exceeding a specified value—as long as the sample size is sufficiently large.

This interpretation of the CLT is plain wrong. The applicability of the CLT is restricted to situations that involve an averaging process—such as in constructing a confidence interval on the mean lifetime of a product—and does not apply for other situations.

Why product lifetimes are typically not normally distributed

The CLT provides a powerful justification for the assumption of normality in many situations. Its applicability, however, is far from universal. This is because many phenomena that you may encounter in practice cannot be regarded, explicitly or implicitly, to be the sum of many small effects. Consequently, there is no reason to expect the normal distribution to apply. Product lifetime is such a phenomenon. Indeed, frequently, one or a small number of failure modes is predominant in determining product life and their effect is not necessarily additive.

Moreover, the normal distribution is defined as running from minus infinity to plus infinity. In contrast, for many products, very short lifetimes may be likely, but negative lifetimes are impossible. You might argue that, likewise, many phenomena encountered in practice that are well described by a normal distribution cannot take on negative values. The height of humans and the time to perform a task are two examples. The mass of the distribution for these and many other phenomena, however, is sufficiently far removed from zero that this restriction is not of practical consequence. This may not be the case for lifetime data.

Also, the normal distribution is represented by a unique bell-shaped curve, as shown in Figure 1. Such symmetry is uncommon for lifetime data.

In summary, unlike the case in many other situations, the normal distribution is typically not an appropriate model for lifetime data (see the sidebar "The Lognormal Distribution").

What to do?

Hopefully, the preceding comments have convinced you—should you have needed any convincing—of the inadequacy of the normal distribution as a model for product lifetime data. But what distribution might you use instead? How can you determine what fits best for your specific application?

Space limitations prohibit us from responding to these questions in detail, so brief summary comments must suffice.[4]

The Weibull and the lognormal (see sidebar "The Lognormal Distribution") are the most popular distributions for describing product lifetime data. Whether either of these or some other distribution applies to the situation at hand should, if possible, be based on physical-chemical knowledge about the failure mechanism of the product. In other situations, previous experience

FIGURE 4

90 original birthdates

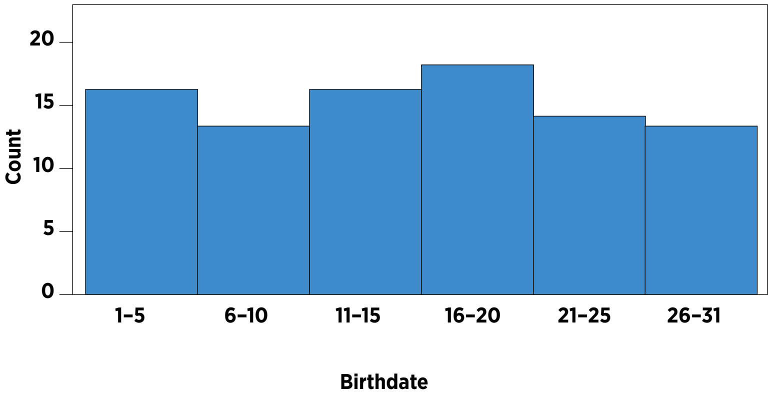
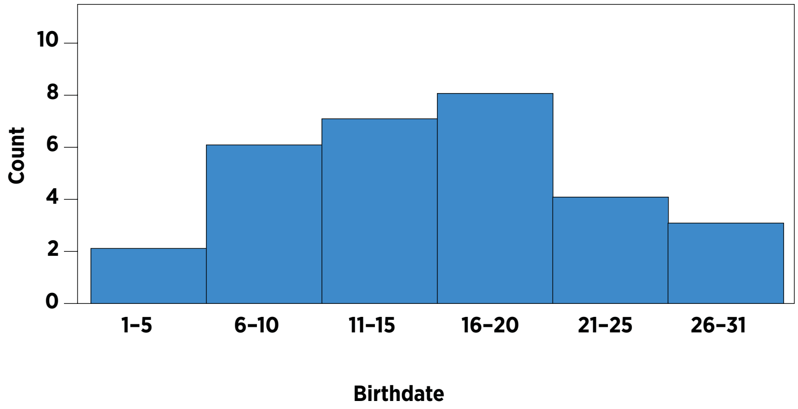


FIGURE 5

30 means of two birthdates



with similar products might suggest one (or more) distribution(s) to use.

Fortunately, statisticians have developed methods to analyze lifetime data under distributional assumptions other than the normal distribution and implemented these in computer software. Such methods also allow analysts to handle another characteristic of lifetime data, namely, so-called "censored observations." These occur in applications in which some units have not yet failed at the time of data analysis—and all that is known about them is their running times.

Then you can assess the suitability of a proposed distribution for the available data. This typically calls for graphical methods, such as probability plots, and more formal statistical analyses.[5] QP

EDITOR'S NOTE

References listed in this column can be found on the column's webpage at qualityprogress.com.

The Lognormal Distribution

Even though product lifetimes typically do not follow a normal distribution, their logarithms might do so. In particular, the lognormal distribution has been justified as a model for crack propagation when product lifetime is dictated by random shocks that increase degradation at a rate proportional to the total amount of degradation already present.

—N.D., G.J.H and W.Q.M.

Necip Doganaksoy is associate professor at Siena College School of Business in Loudonville, NY, following a 26-year career in industry, mostly at General Electric (GE). He has a doctorate in administrative and engineering systems from Union College in Schenectady, NY. Doganaksoy is a fellow of ASQ and the American Statistical Association.

Gerald J. Hahn is a retired manager of statistics at the GE Global Research Center in Schenectady. He has a doctorate in statistics and operations research from Rensselaer Polytechnic Institute in Troy, NY. Hahn is a fellow of ASQ and the American Statistical Association.

William Q. Meeker is professor of statistics and distinguished professor of liberal arts and sciences at Iowa State University in Ames. He has a doctorate in administrative and engineering systems from Union College. Meeker is a fellow of ASQ and the American Statistical Association.