Session 1664

Introducing Statistics to Mechanical Engineers in a Materials Science Course

Scott R. Short, Ph.D., P.E.
Northern Illinois University

Many schools are reducing the number of credit hours in undergraduate engineering programs so students have a better chance of graduating in four years. However, a majority of engineering educators feel that certain fundamental engineering topics such as materials engineering and statistics should be included in the curriculum. In an effort to minimize the number of credit hours required to graduate yet still cover these two important topics, the Department of Mechanical Engineering at Northern Illinois University is incorporating statistics in the laboratory section of their required materials science course. This match is a natural one because the laboratory projects require data acquisition, reduction, and statistical analysis. *Probability paper plots* and Rockwell hardness tests are used to introduce the student to the fundamental building block of statistics, the frequency distribution. An often-overlooked graphical statistical technique, the use of probability paper plots, is a potent teaching tool.

Introducing Statistics to Students

Statistics and its associated foundation of probability are vitally important in engineering. Engineers often must make decisions based upon relatively little information. Decision making requires fast, effective, and practical methods of data reduction and analysis and correlated methods of feedback of the resulting conclusions. An understanding of the interrelationships between mathematics, processes, and statistics can contribute to improvements in data reduction analysis and problem-solving logic.[1]

The importance of statistics in engineering being a given, the critical issue becomes how to present this complex topic to undergraduates in a concise manner. Even the most introductory statistics textbooks are very mathematical in nature and contain a plethora of notation. Faced with having to learn statistics to complete their assigned materials science laboratory projects, most students succumb to the temptation to instead merely plug their data into a spreadsheet computer program (e.g., EXCEL) and command the software to perform a few basic canned statistical operations. Moreover, since most of the software statistical routines are based on the normal distribution, students are led to believe that if experimental data are not normally distributed, then "something is wrong." Simply put, the majority of undergraduates do not realize that the underlying foundation of statistics is the frequency (probability) distribution which may take any of several possible shapes depending on the processes and measurement techniques involved.

One of the most efficient ways to introduce undergraduates to frequency distributions and their associated statistics is through the use of *probability paper plots* (PPP). This often-overlooked

graphical technique is a powerful statistical tool in addition to being a potent teaching tool. King[2] presents an excellent overview of probability distributions in a way different from most statistics textbooks. He uses plots of the cumulative probability distribution function (probability paper plots) to categorize various naturally occurring processes. He further develops one statistical distribution after another, proceeding from the simple to the more complicated, based upon the inherent mathematical processes that underlie the distributions. King's approach emphasizes that associated with all naturally occurring processes are many kinds of statistical distributions. Within a reasonable approximation, the more complex distributions are mathematically definable combinations of the simpler distributions.

Included in the students' laboratory manual is a table[3] (see Table 1, Appendix 1) summarizing statistical distributions. Table 1 shows that the mathematical operation of *counting* results in a discrete statistical distribution called the *binomial* distribution. The mathematical operation of *addition* results in the continuous distribution referred to as the *normal* distribution, and the mathematical operation of *multiplication* results in the continuous distribution referred to as the *lognormal* distribution.

The Binomial Distribution

An example of a binomial distribution is the *binary* distribution, the simplest of all distributions. It is descriptive of a two-state situation, for example, "heads or tails." The students are introduced to the binomial distribution via a simple coin-toss experiment.

The Normal (Gaussian) Distribution

Students are introduced to the normal distribution via a handout which includes the following model. A process based on several two-state systems, i.e., one in which an input variable can take one of two possible states (e.g., a cutting tool being either sharp or dull), can be developed by studying the outcome of the simultaneous tossing of several unbiased coins.

Assume that nine people toss a group of twenty coins simultaneously a total of ten times. (The two sides, head and tails, represent an ideal two-state system.) It can be imagined that the twenty coins represent twenty possibilities for random variation of the binary (discrete) process inputs, while the nine students represent possible process variations. The aggregate results simulate the occurrence of ninety (90) events during a sequence of ten repeat operations, or trials. (By shaking a group of coins simultaneously and allowing them to fall together, one, in effect, creates a function generator producing a set of completely random results for each shake.) By counting the number of occurrences of one of the possible states (heads), it is possible to define a random output variable.[4] This model illustrates how the mathematical operation of addition results in the normal distribution. Students are then taught how to generate histograms by tabulating the number of heads in each trial.

In the development of probability paper plots (PPP) and their prerequisite histograms, there arises the problem of properly arranging, or grouping the data into meaningful increments. This is an important detail that is assumed to be obvious in the general literature of statistics and data

analysis. However, many engineers new to such efforts encounter confusion in preparing data for subsequent presentation and analysis. In many data displays, the use of too few, or too many intervals obscures the form of the distribution.[5]

Sturges' Rule is a method of determining the optimum number of groups, or intervals, for arranging the data into a graphical summary (histogram). This rule states that the optimum number of intervals is found from

$$\text{Number of intervals} = 1 + 3.3 \log n$$

where log *n* is the base 10 logarithm of the number of items of data and *n* is the sample size.

King[6] gives several other tips on developing histograms, e.g., determining histogram interval widths, balancing interval beginning and end points, and defining interval end points correctly.

Students are reminded that, assuming Sturge's Rule has been appropriately applied, the general appearance of a histogram results from the manner of combination (simple addition, in this case) of the random variables. Hence, it is seen that even though the random variables in the coin-toss example are binary (discrete) independent random variables, the distribution which results from adding them together is a bell-shaped distribution referred to as the *normal* (*Gaussian*) distribution.

Many physical properties that are continuous or regular in time and/or space also follow the symmetrical, bell-shaped curve of the *normal* frequency distribution. Normal distributions describe measurements which vary due to *precision* error. Precision error can be affected by the measurement system (repeatability and resolution), the measurand (temporal and spatial variations), the process (variations in operating and environmental conditions), and the measurement procedure and technique (repeatability).[7]

The Lognormal Distribution

Students are introduced to the lognormal distribution via a handout. They are provided with a model similar to that given above for generating the normal distribution, but in this case, the *products* of face counts from tosses of four dice are tabulated. The resultant histogram (using Sturge's Rule) approximates a distribution that is *lognormal*, (positively skewed; sloping to the left, flat on the right) (see Figure 1).

As shown in Table 1, when many random (errors) variables combine *multiplicatively*, the result is a skewed distribution called the *lognormal* distribution. There are many situations in nature where it is more reasonable to suggest that the process underlying change or growth is multiplicative rather than additive. The lognormal distribution is said to obey the "law of proportionate effect" and is appropriate when the effect of the random change in the variables at any step is proportional to the previous value of the quantity.[9]

The lognormal distribution adequately describes many distributions occurring in nature, including economic, biological, and engineering processes. Furthermore, by taking the standard

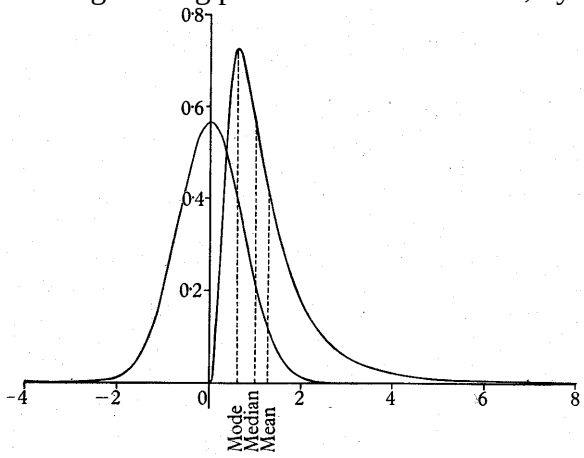

Figure 1. Frequency Curves for the Normal and Lognormal Distributions.[8]

deviation of the logarithms of the variate, σ_l, small enough, it is possible to construct a lognormal distribution closely resembling any normal distribution. In fact, the lognormal is as fundamental a distribution in statistics as is the normal distribution, despite the stigma of the derivative nature of its name. The lognormal distribution forms a fundamental basis for other important distributions and should not be ignored by engineering students. For example, the lognormal distribution has been found to be a serious competitor to another very important distribution in engineering, the Weibull distribution, in representing lifetime distributions of manufactured products. The lognormal distribution is possibly the handiest adjustable wrench in the toolbox of statistical distributions.[10]

Probability Paper Plots and Rockwell Hardness Tests

Although tensile and fatigue test data on many statistical sample groups would produce optimum continuous data distributions, for the purpose of instruction it was thought to be more important that the students perform the data acquisition, as well as the data reduction. Therefore, a much simpler test, the Rockwell hardness test, was used to generate test data.

To introduce the simple, but powerful probability-paper-plot graphical method, each lab group was instructed to make 25 Rockwell hardness measurements evenly covering the cross-sectional face of samples cut from two square steel bars. (Some lab groups received steel bars of identical heat treatment while other groups received bars of differing heat treatment.) The students were instructed to tabulate the acquired data, form histograms, plot the frequency distributions, and determine if the data followed a normal distribution, lognormal distribution, neither, or both.

The students were reminded that it is imperative to always determine the type of distribution that experimental data follow, rather than blindly assuming that the normal distribution describes the data. There are several methods that can be used to determine the type of distribution described by a data sample. One of the simplest methods is to use probability paper plots (PPP).[11]

Frequently, PPP will identify problems and specific characteristics of data which are not discoverable by other analytical methods. These advantages occur because PPP use all of the available data which results in the display of the entire data set along with the relationship of each item of data to all items. Such data displays are far more informative than the exercise of reducing the data set to single abstract numbers, such as the mean and standard deviation. Statistical analysis via PPP is especially potent because the method is capable of determining whether data is non-random or homogenous. A great deal of real-world data contain non-random errors caused by mistakes, carelessness, misunderstanding, and erratic or unstable processes. Blind, mechanistic "cranking out of the numbers" will produce a high percentage of "nonsense numbers." Data proofing is a direct and unique benefit of PPP.[11]

Probability paper plots are convenient and easy to generate because one axis of the probability graph paper has been purposely scaled to match the probability scale corresponding to the cumulative distribution function being targeted. Hence, if a normal distribution is suspected, the data are plotted on normal probability paper. And if a straight line results, then the distribution is indeed normal.

The PPP method is simple because only the following operations need be performed.

1. Arrange the data in increasing order.
2. Group into logical intervals using Sturge's Rule. (Note: no histogram needs to be plotted.)
3. Compute the plotting position.
4. Choose a probability paper.
5. Plot the data.
6. Fit a line to the data.
7. Analyze the results.

The method is efficient because:

1. The correct choice of probability paper results in a straight line of the plotted points.
2. The incorrect choice of paper (type of distribution) is immediately obvious.
3. Failure to obtain a straight line can result from the following desirable information:
   - An incorrect choice of the expected type of distribution,
   - Nonlinearities caused by nonrandom sampling,
   - Nonlinearities caused by truncation resulting from inspection, selection, or other kinds of screening of data, and
   - "Wild" points indicative of errors in obtaining or recording of data.
4. Linear plots allow visual determination of sample statistics such as the mean, median, and standard deviation.
5. Shape patterns are indicative of certain types of distributions and process trends. For example, Figure 2 shows the appearance of normal and lognormal distributions plotted on both normal and lognormal probability paper.[12] In general, the distribution plots made on a probability paper of a lower mathematical complexity than the distribution of interest are concave up, while the more complex distributions are concave down. For example, when data of a normal

distribution are plotted on lognormal paper, the curve is concave down, while when data of a lognormal distribution are plotted on normal paper, the curve is concave up. (When all of the possible distributions that data may take are considered, this investigative method is especially powerful.)

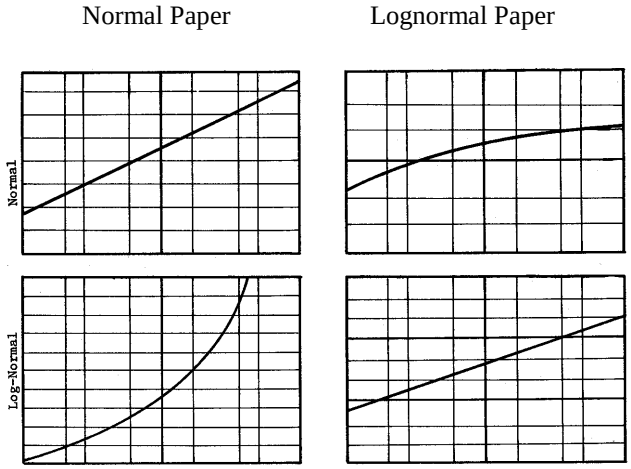

Figure 2. Comparative Probability Plots for the Normal and Lognormal PPP.[12]

After properly plotting the Rockwell hardness experimental data using PPP, the students are able to address the role that the resolution of the measuring device (Rockwell hardness tester) plays in the experimental data acquisition. For example, when measuring an output variable (dimension) of several machine parts with a crude tool like a yardstick, the measured dimension of all parts would be the same, producing a uniform distribution. But, when the output variable (dimension) is instead measured with a very precise tool (e.g., a laser), the variation inherent in the production process would be apparent because a distribution of values of the output variable (dimension) would result. This variation is defined as precision error, which is attributable to several reasons, including variation in the measurement due to the variation of measuring equipment itself and the variation of the output variable itself (which is what we really want to measure). If the variation of the output variable is truly random and the measuring equipment has adequate resolution nor is biased in its sensitivity and repeatability, then the distribution will be normal. Another case occurs if the variable is indeed truly constant (uniform distribution), but the resolution/repeatability/sensitivity of the measuring instrumentation is such that noise results in different readings; the result being that the output variable *appears* to vary. This noise may or may not be normally distributed, at least over a small number of readings. This variation can only be characterized via calibration.

In the case of the heat-treated steel bars, spatial variation of hardness was <u>not</u> random. For example, when traversing the cross section, gathering 25 readings, 9 measurements are located more toward the interior, while 16 measurements are located more toward the outer surface of the bar. If the outer surface is harder than the interior (very likely the case due to cooling considerations), then the measurements, even if taken with a highly precise (and accurate) measurement device, may not be normally distributed. And in the case that the measurement device itself possesses some randomness in its ability to reproduce a reading (not so precise), then an added

element of randomness will enter in the data, causing the distribution to be more normally distributed. The actual distribution obtained depends on the relative influence of each of these factors (among other sources of error, e.g., operator skill, etc.).

Conclusion

A simple materials engineering laboratory project, Rockwell hardness testing, produced experimental data that allowed students to learn about normal and lognormal probability distributions. It was evident that students had misconceptions about the normal distribution. Students felt that all experimental data must conform to the normal distribution. They associated the word "normal" with "accurate" data. Students did not realize that a distribution is determined by how the input variables interact. They also did not understand how the sensitivity of the measurement equipment affects the experimental data distribution. After completing the laboratory project, students had a much better understanding of statistical analyses.

References

1,2,3,4,5,6 King, J.R., *Probability Charts for Decision Making*, Technical and Engineering Aids for Management, Tamworth, NH, 1971, Revised 1981, pp, 266, 1, 267, 9, 13, 15, respectively.

7 Figliola, R.S., Beasley, D.E., *Theory and Design for Mechanical Measurements*," John Wiley & Sons, 1991, p. 14.

8 Aitchison, J. Brown, J.A.C., *The Lognormal Distribution*, Cambridge at the University Press, 1966, p 9.

9 Johnson, N.L., Kotz, S., *Distributions in Statistics, Continuous Univariate Distributions - 1*, Houghton Mifflin, Co., 1970, p. 14.

10 Sachs, L., *Applied Statistics - A Handbook of Techniques*, Springer-Verlag, 1982, p. 107.

11, 12 King, J.R., *Probability Charts for Decision Making*, Technical and Engineering Aids for Management, Tamworth, NH, 1971, Revised 1981, pp. xii, 270, respectively.

SCOTT R. SHORT

Scott R. Short obtained his Ph.D. in Engineering from The University of Dayton, Dayton, Ohio in 1990. Prior to attending graduate school, Dr. Short was employed as a metallurgist with ARMCO, Inc. Dr. Short currently is an Assistant Professor in the Department of Mechanical Engineering at Northern Illinois University in DeKalb, IL and is a registered Professional Engineer in the State of Illinois.

APPENDIX I

Table 1. Relationships Between Mathematical Models and Resultant Statistical Distributions.[3]

Mathematical Operation	Mathematical Model	Process Description	Example	Resultant Statistical Distribution
Counting	$p = \frac{c}{n}$	Enumeration or Classification	Inspection Sorting	Binomial
Addition	$f(y) = \sum_{i}^{n} (x_i)$	Linear Additive	Addition or subtraction of materials; i.e., cutting, weighing, etc., also mechanical assembly.	Normal
Multiplication	$f(y) = \prod_{i}^{n} (x_i)$	Rate-Dependent Proportional Response	Simple chemical processes; i.e., etching, corrosion, gaseous diffusion. Simple biological processes; i.e., growth rate. Simple economic processes; i.e., distribution of income.	Log-Normal
Simple Exponentiation *or* Addition of Transcendental Terms	$f(y) = ax_0 + bx_1 + cx_2^2$ *or* $f(y) = e^{x_0} + e^{x_1} + e^{x_2}$	Algebraic Polynomial Solutions of Linear Differential Equations with Constant Coefficients.	Complex processes involving the combined effects of a number of independent causes each with a different operational form; i.e., breaking strengths, meteorological and geophysical phenomena, electronic and chemical measurements, financial data.	Extreme Value
Counting of Time Duration to an Event	$f(x; n, \lambda) = \frac{\lambda^n}{\Gamma(n)} x^{(n-1)} e^{-\lambda x}$	Waiting Time	Time required for an event(s) to occur or to obtain some service.	Gamma
Addition of Squared Normalized Vectors	$f(y) = \sum_{i}^{n} \left(\frac{x_i}{\sigma_i}\right)^2$	Vector Sums	Resultant value in a system of n-fold vector spaces from physics, space-time, and probability applications.	Chi-Square
Multiplication of Transcendental Terms	$f(y) = e^{(x_0 \cdot x_1 \cdot x_3 \ldots)}$ $f(y) = e^{(x_1/x_2)(x_3/x_4)}$	Solutions of General Differential Equations Particle Sizing	Complex exponential processes involving the interdependent effects of independent causes; i.e., breakage of particulate materials, solid state diffusion, chemical kinetics.	Log-Extreme Value
Sums, Products, and Powers of Exponents of Transcendental Terms	$f(y) = e^{\left(\frac{\omega - x}{\omega - \mu}\right)^{\beta}}$	Solutions of Differential Equations with Boundary Conditions "Upper-Limit" Distributions	Processes involving limits and maxima-minima; i.e., life/failure distributions, bounded particle size distributions, and general potential, gradient, and field problems.	Weibull