Material EASE

AMPTIAC 5

Jorge Luis Romeu
IIT Research Institute
Rome, New York

STATISTICAL ANALYSIS OF MATERIAL DATA

PART I: RANDOM VARIABLES, DISTRIBUTIONS AND PARAMETERS

Introduction

Sometimes, engineers have problems understanding the philosophy behind the statistical procedures they need to apply when analyzing materials data. This is not surprising, considering that in many engineering curriculums (i) statistics is limited to one or two (3 or 4 credit hour) courses; (ii) these are rather theoretical, instead of data analysis oriented; (iii) they discuss too many statistical techniques and (iv) engineering experimentation emphasizes the physical (deterministic) rather than the stochastic laws governing the processes under analysis.

The objective of the present series of three articles is to address this situation. In this first article, we will discuss the philosophy behind random variables, statistical distributions and their parameters, including the special problem of outlier (or extreme value) detection and treatment. In the second article, we will discuss the philosophy behind parameter estimation and hypothesis testing, emphasizing goodness of fit procedures used to identify and select suitable distributions from a given set of data. In the third and last article of this series, we will apply these concepts and philosophy to some of the procedures discussed in the Composite Materials Handbook (MIL-HDBK-17) and the Metallic Materials and Elements for Aerospace Vehicle Structures (MIL-HDBK-5)[1, 2].

Statistical Distributions

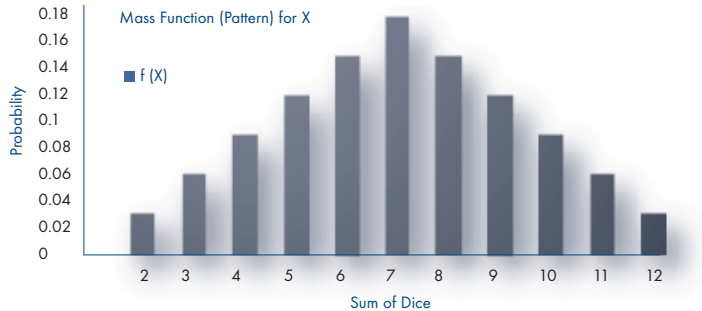
Statistics deals with the study of phenomena and processes that (i) yield more than one outcome that (ii) occur in a random fashion. These outcomes, resulting from the (conceptual) random process under observation, are called random variables (R.V.). We denote such conceptual R.V. with a capital letter, say X; their specific outcomes are called "events"; and the set of all possible R.V. outcomes is called the "sampling space" [3, 4]. For example, from the process of rolling two dice and taking their sum, we observe X, the random variable "resulting sum" and from the process of testing a given metallic specimen under specific conditions, we observe X, the random variable "maximum crack length". In the dice example, the sampling space consists of integers 2 through 12, an event is {X = 4} (rolling a sum of four) which occurs with probability P{X = 4} = 3/36 (see Table 1). For the crack length example, the sampling space consists of all positive reals, an event is {X<3.5} (observing a crack of length less than 3.5 inches) for which we can also obtain a probability.

The (graphical) frequency pattern of occurrence of such random outcomes (e.g. Figure 1) provides an intuitive way to understand what is the statistical distribution of a R.V., X. Such a graph presents, in the abscissas, the sampling space of X (all its possible outcomes) and in the ordinates, a value proportional to the frequency of such outcomes. A standardized version of such a graph of outcomes pattern (so that the area under it is unit) is called the probability density (when the sampling space of X is continuous) or mass (when discrete) function. The Distribution function F (non-decreasing, between zero and unit) of a R.V. X, is defined using the mass/density function, in the following way:

$$F(a) = P\{X \leq a\},$$

where "a" is any feasible value in the sampling space of X.

These mass/density functions (patterns) provide an objective and precise way to describe the probability mechanism governing the random process that produces them. For example, contrast the (graphical) flat pattern from rolling an honest die, where the occurrence of any of its six sides is equally likely, with that of the sum of two dice (shown in Figure 1), where a sum of 7 is more likely than that of a 12. Such outcome patterns (distributions) can be numerically described by a set of fixed numbers called parameters. In the sum of two dice example, the set (1/36, 2/36, 3/36, ... 1/36) of frequencies associated with the possible sums, uniquely describe its distribution (pattern). Thence, all random variables have a distribution, uniquely described by (one or more) parameter(s). Statistics is about investigating those distributions and parameters. In this series of articles we will discuss quantitative (as opposed to qualitative) R.V., i.e. those whose numerical outcomes exhibit mathematical properties of order and distance (and some times even have an absolute zero). These R.V. are said to have a "stronger" measurement scale level, which allow the implementation of certain statistical methods, not always appropriate for qualitative variables [5].



DICE							X	f(X)
6	7	8	9	10	11	12	2	0.028
5	6	7	8	9	10	11	3	0.055
4	5	6	7	8	9	10	4	0.083
3	4	5	6	7	8	9	5	0.111
2	3	4	5	6	7	8	6	0.139
1	2	3	4	5	6	7	7	0.167
	1	2	3	4	5	6	8	0.139
							9	0.111
							10	0.083
							11	0.055
							12	0.028

Table 1. X is the sum of two dice

Statistical distributions are discrete or continuous, according to whether the corresponding R.V. sampling space is discrete or continuous. The dice is an example of discrete, and the crack length is an example of continuous, R.V. Their corresponding (graphical) patterns yield step or continuous mass/density functions. Discrete R.V. allow calculation of (event) probabilities for individual outcomes (e.g. getting a sum of two) while continuous R.V. only allow calculation of probabilities for ranges (e.g. of getting a fracture of less than three inches long). For example, the calculation of the probability of "obtaining exactly a sum of three" (denoted P{X = 3}) or of "observing a fracture of less than three millimeters long" (denoted P{X < 3}) is obtained by adding (integrating) the discrete (continuous) mass/density function (pattern) discussed above. This yields the one-to-one relation between distributions and their corresponding mass/density functions, upon which statistical work is based.

In addition to discrete or continuous, distributions can be symmetric or skewed, according to whether their mass/density functions are/are not symmetric with respect to one point in their sampling space. Distributions can also be unimodal or multimodal, according to whether their mass/density functions have one or more than one (local) maximum. The distribution of R.V. "sum of two dice" in Figure 1, is an example of a symmetric, unimodal distribution. Its mean (and mode) is 7, about which the distribution is symmetric.

As can be surmised, the number of statistical distributions that can arise is infinite, which poses a difficult problem. In order to deal with it, well known and thoroughly studied "families" of statistical distributions, with a small and easy to interpret number of parameters, have been developed. Two examples of discrete families of distributions (and their respective parameters) are, the binomial (n, number of trials and p, probability of success of any trial) and Poisson (*l*, event rate of occurrence). Two examples of continuous distributions are the normal (with mean, *m*, and standard deviation, *s*) and exponential (with failure rate *q*).

If the (exact) distribution of a process under study can be satisfactorily approximated by one of these well known distribution families, by finding suitable parameters (i.e. we can live with the difference between the exact probability of any event and its approximation by one of these families) then we may work with the latter as if it were the exact distribution. Much statistical work is spent in (i) selection of a specifically suited family of distributions, (ii) estimation of adequate parameters, (iii) verification (testing) that such selection is correct and (iv) obtaining usable probabilistic results with them. We will see more of this type of work in the second and third articles of this series.

The above discussion shows the importance of understanding the concepts of R.V., their distributions and their corresponding parameters, as objective and precise ways of describing/prescribing a random phenomenon under study. Activities (i) to (iii) above may be performed on a given data set (as in, say, the procedures of MIL-HDBK-5 and 17) in order to provide (iv) practical and useful, probabilistic statements on "events" of interests. For example, "what specific stresses, can the population from which this sample of metal sheets comes from, withstand?" Conversely, pre-specified probabilities (or equivalently a specific distribution and its parameters) may be given by the engineering designers as minimum required performance measures (say, in the form of mean values or percentiles for a given metal characteristic). And they can be used as benchmarks against which samples of incoming materials are screened and tested for acceptance.

Distribution Parameters

Parameters are population (fixed) values that uniquely characterize and describe the distribution of a R.V. Parameters allow the graphing of the R.V. specific mass/density function (outcome patterns). In many cases, we can even directly identify the parameters in the mass/density function graph. Hence, their understanding and interpretation is of great importance. There are many types of parameters, but we will discuss here (i) location and dispersion parameters and (ii) shape, scale and threshold parameters, which are widely used.

Location parameters respond to the question "where is the distribution". A particularly useful subset of these is given by the measures of central tendency: mean, median and mode. Meaningful interpretations of these parameters are also important. The distribution mean is the R.V. outcome located at the center of gravity of the mass/density function graph. The median is the outcome such that half the population scores below (above) it. The mode is the value where the mass/density function peaks (most frequent outcome). Mean and median, if they exist, are unique. Multiple modes may coexist (i.e. in a multimodal distribution). If a distribution is symmetric and unimodal then mean, median and mode coincide (e.g. sum of 7, in the dice example, Figure 1). And this is as good as it gets.

If a distribution is skewed (non symmetric) then one tail is longer than the other and the mean loses importance to median and mode. For example, R.V. income distribution has in many countries a highly skewed distribution. Hence, its distribution mean may be of small use if say, there are several billionaires and millions of landless peasants. In such a case, (i) the median will provide an income level such that half the population income lies above (below) it, and (ii) the mode will yield an income level that is most frequent and around which there is some clustering. The latter two usually provide more useful and meaningful information about the population and the phenomenon under study. In addition, if we add (subtract) a few billionaires more to the population, the mean will be affected, whereas mode and median will be much more resilient to such types of changes. Such resilience is referred to as "robustness" of a parameter and is considered a good quality. Figure 2 represents a skewed, continuous distribution.

Other location parameters of interest are the maximum/minimum values and the percentiles. A percentile is an outcome value, within the sampling space of the R.V. such that a given percent of the population scores a result less than or equal to such outcome. For example, by definition the median is the fiftieth percentile (because 50% of the population scores less than or equal to it). Other important percentiles are the lower (upper) quartiles which define values where 25% of the population (75% of the population) are less than or equal to such value. For example, in MIL HDBK 17, there is great interest in estimating the first and tenth percentiles of the population distributions under study (for, these two percentile estimations are used to obtain the A and B basis allowables). Since A or B basis allowables represent the highest quality of data, their importance cannot be underscored. According to these (A or B) allowables, 99% (or 90%) of the corresponding population values (for a given performance measure of interest, say tensile stress) will be larger than this value.

It is very important to understand that percentiles vary with a numerical change in the value of a parameter, even within the same distribution family. To provide an example, let's obtain the 31st percentile, for a Binomial (n = 4, p) distribution (say, the distribution of correct answers in a four-question test, where every question had an unspecified probability p, of being answered correctly by any student). The sampling space consists of integers 0 through

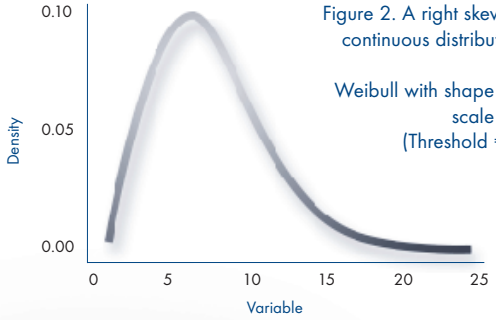


Figure 2. A right skewed, continuous distribution:

Weibull with shape = 2
scale = 8
(Threshold = 0)

4. From any Binomial table, we verify that the percentile in question is One, if parameter p = 0.5, and Zero if p = 0.25. This stresses the importance of establishing a correct distribution and of obtaining good parameter estimators for this distribution.

Dispersion parameters respond to the question: "how does the random process vary, about some location parameter". Some well known dispersion parameters are variance, range and interquartile range. The standard deviation is the square root of the variance. In a normal distribution, the standard deviation yields the distance from the mean to the inflection point of the density function. Adding and subtracting this distance respectively, once, twice and trice, from the mean provides ranges where 68%, 95% and 99% of the population lie. The Range is the difference between maximum and minimum outcomes. The Interquartile Range (IQR) is the difference between the (Upper/Lower) quartiles.

Dispersion parameters are used to characterize or compare population variability. If means of positive R.V. are the same, for example, their variances can be compared directly. But if means differ, then an indirect dispersion parameter, such as the coefficient of variation (defined as the ratio of the standard deviation to the mean) is used. Also, as distributions depart from symmetry, the usefulness of the variance loses to IQR, for the same reasons that the mean loses to median and mode.

Finally, shape and scale parameters provide the degree of curvature necessary to "adapt" a specific family to a specific population (i.e. to obtain a good fit or approximation to the exact R.V. distribution). Other useful parameters include the threshold parameter, which provides a lower bound for the R.V. range of possible values. The three-parameter Weibull [6] is a good example of such a three parameter distribution (it is worth noticing that, now, mean and variance are obtained as a function of shape and scale parameters). Skewness and kurtosis are two parameters that describe the degree of (dis)symmetry and of peakedness of a distribution. In all cases, parameters help visualize the pattern of possible R.V. outcomes.

Extreme Values or Outliers

Once, a suitable family of distributions (and its corresponding parameters) has been found to accurately characterize a random phenomenon, we can proceed to analyze its behavior in the tails. This is particularly important in hypothesis testing (which will be

discussed in the next article). For it allows us to ascertain whether an (unusual) observation may have a reasonable probability of occurrence or, on the contrary, may signal a process anomaly.

An outlier in a dataset is defined as, an observation which appears to be inconsistent with the rest of the data, in relation to an assumed statistical model [7]. Hence, an outlier either occurs with a very small probability in the assumed model or does not belong to the population (i.e. is a "contaminant"). It is therefore incorrect to believe that an outlier is always an erroneous observation or that it should be automatically removed. In the dice example, a sum of 12 occurs with probability 1/36 = 0.028. But it may occur at any trial with that probability. If we perform the dice experiment three times in a row , we may obtain three sums of 12, an event that may occur with probability 2.14×10^{-5}, very small but not zero. We could then have grounds to suspect that the dice are fixed (or otherwise that we are extremely unlucky). Such infrequent result indeed raises a red flag - but does not insure foul play. Statistics provides a useful and scientific context in which to analyze such results but not a mechanical working rule.

For example, in a sheet metal process we may observe that, with probability 0.001 a large number of cracks/sheet are obtained. If we mechanically decide to discard this information as erroneous, simply because we have detected an "extreme value" (outlier) we may be disregarding important information. It may occur, say, that a rare combination of humidity, room temperature, pressure and defective metal composition (combination that occurs with probability 0.001) always produces such poor quality material. If instead of discarding this "outlier", we collect several specimens of it, review their circumstances and submit them to laboratory, technical and statistical analysis, we could uncover the real reasons behind such rare event. We could then, by better controlling room temperature and other production factors, remove the problem (instead of the outlier that points toward the problem) and reduce the overall process variability.

On the other hand, some times there is indeed a clerical error or some other circumstance that does warrant discarding the element because it no longer represents the population under analysis. Only in this case, it is adequate to remove the element from the data set.

Conclusions and Summarization.

Statistical analysis, like most others, is more than just the mechanical application of a set of procedures and equations. Actually, many statistical procedures and equations are the result of a systematization in the process of scientific experimentation, derived under certain (statistical) assumptions and conditions. If such underlying assumptions and conditions (e.g. data normality, independence, homogeneity of variances, etc.) are not met, then the results obtained from the statistical procedures used are not valid or have different interpretation (i.e. different probabilities of occurrence).

The objective of the present article is, precisely, to provide additional insight into the statistical thinking process, so that the engineering practitioner may improve the use of statistics as an everyday analysis tool. The next issue of MaterialEASE will focus on parameter estimation, hypothesis testing and selection of suitable distributions for data. The third part of this series will concern application of these concepts to the analysis of material data.

Note: Comments or questions on this article can be posted on the AMPTIAC Materials Forum located on AMPTIAC's web site. (http://amptiac.iitri.org)

Bibliography

1. Metallic Materials and Elements for Aerospace Vehicle Structures MIL HANDBOOK 5G. November 1994.
2. Composite Materials Handbook MIL HANDBOOK 17. 1D.
3. Introduction to Probability and Statistics for Engineers and Scientists. Ross, S. M. Wiley New York. 1987.
4. An Introduction to Probability Theory and Mathematical Statistics. Rohatgi, V. K. Wiley. New York. 1976.
5. Some Measurement Problems Detected in the Analysis of Software Productivity Data and their Statistical Consequences. Romeu, J. L. and S. Gloss-Soler. Proceedings of the 1983 IEEE COMPSAC Conference. Pages 17 to 24.
6. Methods for Statistical Analysis of Reliability and Life Data. Mann, N., R. E. Schafer and N. Singpurwalla. Wiley. New York. 1974.
7. Langford, I. H. and T. Lewis. "Outliers in Multilevel Data". Journal of the Royal Statistical Society (Series A). Vol. 161, Part 2 1998). Pages 121-160.

EMAIL: amptiac@iitri.org
http://amptiac.iitri.org

PHONE: 315.339.7117
FAX: 315.339.7107

AMPTIAC is a DoD Information Analysis Center Sponsored by the Defense Technical Information Center and Operated by IIT Research Institute