Chapter Three:
Statistical Distributions
RandomVariables, Distributions and Parameters

Jorge Luis Romeu
IIT Research Institute
Rome, NY 13440

June 9, 1999

Executive Summary

Statistical distributions are used to describe the random outcomes of varying processes, to better understand them and work with them. In this chapter we discuss the meaning, interpretation and philosophy of random variables, of their statistical distributions (patterns) and of their parameters. Three specific distributions, Normal, Lognormal and Weibull, that are at the heart of materials data analysis and design, are discussed in detail. The special problem of analysis, detection and treatment of sample and distribution outliers (or extreme values) is also discussed. Illustrative numerical examples are presented and discussed.

Statistical Distributions

Generalities

Statistics deals with the study of phenomenons and processes that (i) yield more than one outcome and (ii) occur in a random fashion. These process outcomes (observations, data), resulting from the (conceptual) random process under observation (stress, strain, loads) 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" [8, 10]. For example, from the process of rolling two dice and taking their sum, we observe X, the random variable "resulting sum". 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 reason statistics is so important to the materials engineers is that processes such as stress, strain, load, deformation, etc., that take place on materials, are stochastic and not deterministic. Therefore, a fixed parameter such as the mean is not very informative. For, a large percentage of the population of all possible stresses, strains, loads, etc. that a specific

material can be subjected to, may be very different than this parameter. Thus, we also need information about its variability.

A or B basis allowables, instead, are very informative values and hence can be used for design. For, a prefixed percentage of the population outcomes are assured, with a given probability, to be above A or B basis values. From here, we can realize the importance of correctly identifying the pattern (or statistical distribution) of such (strain, stress, load, etc.) stochastic or random processes. This is the main topic of this chapter.

The (graphical) frequency or pattern of occurrence of specific random outcomes (e.g. Figure 3.1) provides an intuitive way to understand what is the statistical distribution of a R.V. X (e.g. the stress, strength, load, strain, or other random process of interest to the materials engineer). Such graph presents, in the abscissa axis, the sampling space of X (i.e. all possible outcomes of such stress, strength, load, strain, etc.) and in the ordinates, a value proportional to the frequency of occurrence of such outcomes. A standardized version of such 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 of a R.V. X, denoted F, is non-decreasing, between zero and unit, and defined using the mass/density function, in the following way:

$$F(a) = P\{X \leq a\} \text{ where "a" is any feasible value in the sampling space of X}$$

These probability mass/density functions (patterns) provide useful, objective and precise ways to describe the probabilistic mechanism governing the random processes of stress, strengths, load, strain, etc. that produces them. For example, contrast the (equiprobable) 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 3.1), where a sum of 7 is more likely than that of a 12. Such 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 chapter we discuss three special quantitative (as opposed to qualitative) R.V. They correspond to the Normal, the Lognormal and the Weibull distributions, which are frequently used to describe the patterns of materials processes of interests.

Quantitative R.V. are those whose numerical outcomes exhibit mathematical properties of order and distance (and some times even have an absolute zero). They are said to have a "stronger" measurement scale level, which allows the implementation of certain statistical methods, not always appropriate for qualitative variables. Some of such methods include regression and ANOVA, which will be discussed in further chapters,.

Statistical distributions can be discrete or continuous, according to whether their corresponding sampling space outcomes are discrete or continuous. The dice sum 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 the probability of a fracture of less than three inches long, on a given material, under a specific load).

Specifically, the calculation of the probability of "obtaining exactly a sum of three, when rolling two dice" (denoted P {X=3}) or of "observing a fracture of less than three inches long" (denoted P {X<3}) is obtained by adding (or integrating) the discrete (or continuous) mass/density function (patterns) discussed above. This illustrates the one-to-one relation between distributions and their corresponding mass/density functions, upon which statistical work is based. Hence, if we know the distribution of a R.V., we know its density. This also illustrates the importance of correctly characterizing (finding a good Fit for) the distribution of the random process (of say loads on a given material) under study.

In addition to being discrete or continuous, distributions can also be symmetric or skewed, according to whether their mass/density functions are/are not symmetric with respect to one point in their sampling space. This can be useful, for if the distribution is symmetric about a point, say the mean, its practical value rises considerably. For, the process under study will behave as if it departed at the same rate from this center of symmetry.

Distributions can also be unimodal or multimodal, according to whether their mass/density functions have one (or more than one) local maximum (e.g. they cluster about these points). For example, the distribution of R.V. "sum of two dice" in Figure 3.1, is symmetric and unimodal. Its mean and mode is 7, about which the distribution is symmetric and clustered. If one had to choose three numbers with the highest wining probability, these should be 6, 7 and 8. Unimodality is useful because one will have small ranges of values, where large percentages of the phenomenon under study tend to concentrate. Next section presents an example where large percentages of all possible tensile strength values, from a given material and under specific conditions, fall within a small interval with a high probability.

As one can imagine, the number of statistical distributions that can arise in real life is infinite, which poses a difficult problem. In order to practically 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 (with number of trials n and probability of success of any trial, p) and the Poisson (with rate of occurrence λ). Two examples of continuous distribution families are the Normal (with mean μ and standard deviation σ) and the Weibull (with scale θ and shape β). We refer to them as "families" because different patterns can be obtained, that describe different process behaviors, by varying their parameters. We will study continuous distributions in the next section.

The (exact) distribution of a process of interest (say the stress of a material) may be satisfactorily approximated by a well-known distribution (by finding some parameters that provide a similar pattern). If so, we will work with the latter as if it were the exact (but

unknown) underlying distribution. We approximate the true distributions (and we say, that we find a good Fit to them) by estimating suitable parameters (say, μ) that allow a member of the well known family of distributions (say, Normal or Weibull) to satisfactorily describe the pattern of the process outcomes (say, tensile strength variability).

In other words, we can neglect the difference between the exact probability of the occurrence of say, a specific strength on the material under study and its approximation by one of these distributions, say the Weibull. Much statistical work is spent in the (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 following chapters.

The above discussion shows the importance of understanding the concepts of R.V. (e.g. process outcomes such as strength, loads on a wing, etc.), their distributions (e.g. the pattern of such outcomes) and their corresponding parameters (fixed values that provide a good distribution Fit). These distributions then provide objective and precise ways of describing or prescribing the random phenomenon under study. Activities (i) to (iii) above are performed on a given data sample following, say, MIL-HDBK-5 & 17 procedures. The distribution and parameters found are then used to obtain A and B basis values.

Then, materials engineers or designers use these values to obtain (iv) practical and useful, probabilistic statements on "events" of interests. For example, "what stresses, does 90% of the population from which this sample of metal sheets comes from, can withstand?" Or, conversely, what pre-specified probabilities (given a specific distribution and its parameters) should be required by the engineering designers as performance measures (say, in the form of percentiles for a given metal characteristic). Finally, such values can also be used as benchmarks, against which samples of incoming materials and their test results are screened and assessed for acceptance in the data base.

Three Distributions of Interest in Materials Data Analysis

There are three main distributions of interest in MIL HDBKs 5 and 17: the Normal, the Lognormal and the Weibull. They are well studied because they actually fit many materials processes or because, for physical reasons, some process outcomes follow such patterns.

We will illustrate our discussion of these distributions using an example. We have selected it from problem 6 (pages 8-47 and 8-57 of [7]) because it deals with real life tensile strength measurements, having a mean of 330 and a standard deviation of 5. We have simulated population and sample data from several distributions with these parameters, for discussion and comparisons. For all these distribution results are close and signal out the importance and difficulties of determining the correct underlying distribution and parameters.

The Normal Distribution

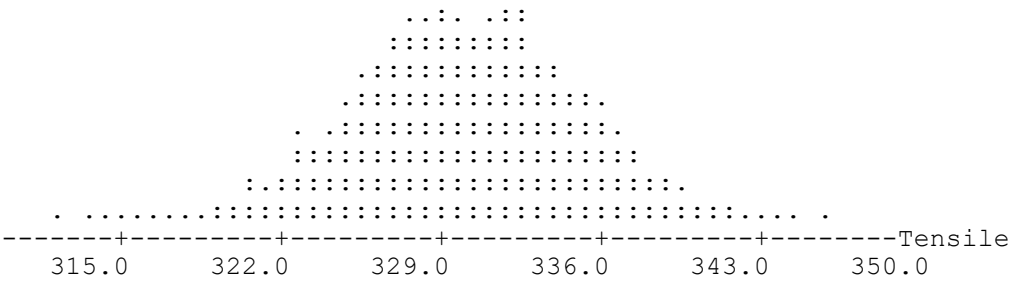
A process or R.V. X (say, stress on a sheet of metal) follows a Normal distribution, with mean μ and standard deviation σ (i.e. $N(\mu, \sigma)$) if its density function can be written as:

$$f(x) = \frac{1}{\sqrt{(2\Pi)}\,\sigma} \exp\{-(x-\mu)^2 / 2\sigma^2\} \qquad \text{for values} \quad \infty < x < \infty\ ;\ \sigma > 0$$

One of the greatest problems with using the Normal distribution to characterize materials data is the fact that the X values can conceptually be negative (which is not possible for strength, loads, stress, etc.). It does not matter how small the standard deviation σ is or how large is the mean μ. There is always a (perhaps very small) probability of occurrence of a negative value (e.g. strength). In practice, however, this probability is usually negligible.

The Normal distribution is unimodal and symmetric about the mean μ, where values tend to concentrate. Its mean, median and mode coincide. It is also standardizable, i.e. there is one distribution, the Normal Standard, $N(\mu=0; \sigma=1)$ that provides all probabilities. Any Normal value X can be "standardized" by the process: $Y = (X-\mu)/\sigma$. After this, "Y" is Normal Standard i.e. has mean unit and standard deviation one. Hence, there is only one probability table for the Normal Distribution: the Standard Normal table. Any of the textbooks given in the references provides additional information and the tables for the Normal distribution.

We present below a dot plot for 3000 data points generated from a Normal with $\mu=330$ and $\sigma=5$. Verify how it is unimodal and symmetric about 300, which is also its median value. Each plot point represents 11 tensile strength data values, from this population.



The Lognormal Distribution

A process or R.V. Y (say stress on a sheet of metal) follows a Lognormal distribution, e.g. $\Lambda(\mu,\sigma)$ if the logarithm of this process, say $X = \text{Log}(Y)$ follows the Normal distribution. The Lognormal density (Figure 3.2) has the following functional form:

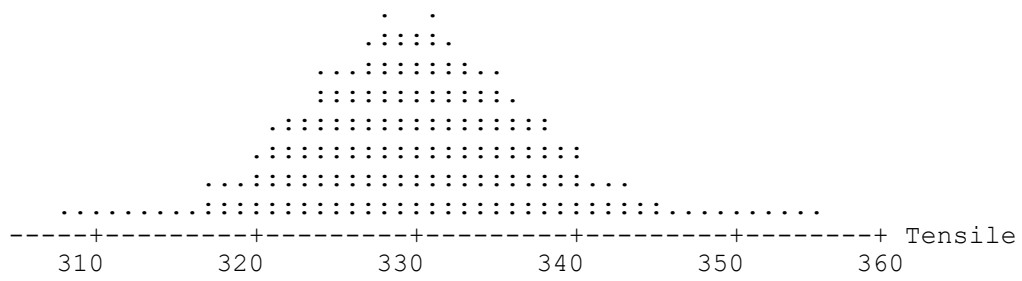
$$f(y) = \frac{1}{\sqrt{(2\Pi)}\,\sigma\, y} \exp\{-(\ln y - \mu)^2 / 2\sigma^2\} \qquad \text{for values} \quad 0 < y < \infty\ ;\ \sigma > 0$$

The Lognormal distribution does not take negative values. Therefore, it is more realistic in this sense than the Normal. It is not symmetric, but usually skewed (i.e. one tail is longer than the other). Its mean or expected value, does not coincide with either its mode or its median. Its expected value and variance are, respectively:

$$E(Y) = \exp\{ \mu + \sigma^2/2\} \quad \text{and} \quad V(Y) = \exp\{ 2\mu + \sigma^2\} (\exp \{ \sigma^2 \} - 1)$$

We don't need tables for the Lognormal since one can obtain its probabilities by taking logarithms and using the Normal tables, e.g. $P(Y \leq a) = P(X \leq \ln(a))$. An excellent treatment of the Lognormal distribution is found on pages 264 and following, of [11].

We present below a dot plot for 3000 data points generated from a Lognormal ($\ln(\mu)$=5.8 and 0.02) that yield mean of 330 and standard deviation of 6.6. It is somewhat flatter than the previous Normal (notice length of the tails). Each plot point represents 13 data values.



Weibull Distribution

A process or R.V. X (say strain on a sheet of metal) follows a Weibull Distribution, if the density function (where β is the shape and θ is the scale parameter) has the following form:

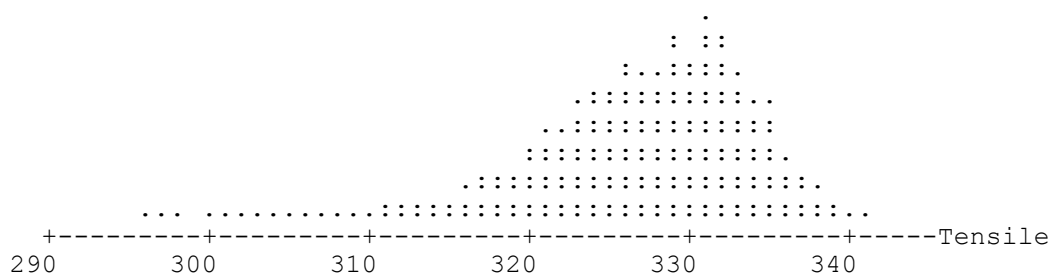
$$f(x) = \frac{\beta}{\theta} \frac{x^{\beta-1}}{\theta^{\beta-1}} \exp -\{ \frac{x^{\beta}}{\theta^{\beta}} \} \qquad \text{for values of } x, \beta, \theta > 0$$

As in the Lognormal, this is a more realistic distribution since all values x have to be positive. Also, in addition to providing an empirically good fit, the Weibull has a physics basis for its use in reliability and materials analysis. For, it is the asymptotic distribution of the smallest values from certain other distributions. Weibull is not symmetric, but skewed (one tail is longer) and its mean (expected value) median and mode do not coincide, either. As with the Lognormal, it is very flexible and can accommodate a large number of pattern shapes (Figure 3.3). Its mean and variance are:

$$E(X) = \theta \, \Gamma(1 + 1/\beta) \quad \text{and} \quad V(X) = \theta^2 [\Gamma(1 + 2/\beta) - \Gamma^2(1 + 1/\beta)]$$

Probabilities are obtained directly by the formula. An advanced treatment of the Weibull can be found on pages 184 and following of [11]; a basic treatment is found in [12].

We present below a dot plot for 3000 data points generated from a Weibull, with shape parameters, 60 and 330. The mean is 327 and standard deviation 6.8 (similar to the others). It is left skewed (notice the length of the left tail). Each plot point represents 15 data values.



Distribution Parameters

Distribution Parameters, as we have seen in the three cases above, are population (fixed) values that uniquely characterize the distribution function describing a R.V. Parameters allow the graphing of the R.V. specific mass/density function (outcome) patterns. For example, the Normal distribution can be taller/slimmer or flatter/broader, given the same mean μ, according to whether σ is larger or smaller. And the Lognormal and Weibull can be right or left skewed, peaked, unimodal or not, according to their shape and scale parameters. Examples of shapes of these distributions are presented in Figures 3.2 and 3.3.

In many cases, we can even directly identify the parameters in the mass/density function graph, as we have done above. Hence, their understanding and interpretation is of great importance. There are many parameters, but we will only discuss here the widely used location and dispersion parameters and the shape, scale and threshold parameters. We will illustrate these concepts using parameter values from our three examples above developed.

	N	MEAN	MEDIAN	STDEV	MIN	MAX	Q1	Q3
normpop	1000	330.12	330.01	4.91	312.30	347.34	326.80	333.27
logpop	1000	330.26	330.03	6.59	310.23	352.58	325.78	334.80
webpop	1000	326.77	327.88	7.01	294.96	340.66	323.17	331.75

Location parameters respond to the question “Where is the distribution?” Some very useful location parameters are the measures of central tendency: mean, median and mode. All three examples generated above had very similar measures of central tendency.

Meaningful interpretations of these parameters are also very important. In general, the distribution expected value, mean or long run average is denoted by E(X). It is also the outcome located at the center of gravity of the mass/density function graph. The median is the outcome such that half of the population scores below (above) it. The mode is the value

where the mass/density function peaks (most frequent outcome). If a distribution is symmetric and unimodal then mean, median and mode coincide (e.g. Normal Distribution).

Mean and median, if they exist, are unique; there can be many modes. Multiple modes may coexist (i.e. in a multimodal distribution) if the shape of the density shows more than one hump. This is not usual and when it occurs it often indicates that there are more than one homogeneous populations pooled together.

If a distribution is skewed (non symmetric), as with Lognormal and Weibull, then one tail is longer than the other. This is very noticeable in our Weibull example. In such cases the mean loses importance to median and mode. For no longer there is a cluster about the mean, but about the mode, instead. The median, however, is always interesting because it indicates the population value such that half all others are above it, and half are below it. That is, if we sort the population, the median is the value that occurs in the middle. Knowing where this central value is, provides many interesting uses.

For example, R.V. "income distribution" usually is highly skewed. Hence, its distribution mean is not very informative if say, there are few billionaires and millions of landless peasants. More information is provided by the median, which is the income level such that half the population income lies above (or below) it. The mode, in turn yields the income level that is most frequent and around which, there is some income clustering.

The same occurs with stress on a metal sheet. If the distribution is skewed, as with our Weibull example, then the average stress will differ with respect to the mode, about which more stress values tend to cluster. This becomes more acute as skewness increases. In such cases, the average population stress will be of small design value.

In our Weibull example, skewed to the left, the approximate value of the mean is 327, of the median is 329 and of the mode is 332. Also notice that the maximum value is 340 and the minimum is 295. They are non symmetrically situated with respect to the mean and differ from those same parameters in the Normal and Lognormal examples.

Therefore, unless the distribution is symmetric, mode and median usually provide more useful and meaningful information about the phenomenon under study, than the mean. In addition, if we add (subtract) a few extreme values, 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. For example, a few test cases where there is a particularly small or large tensile stress, will affect the average stress. However, it will not affect the modal or median stress values.

Other location parameters of interest are the maximum/minimum values and the percentiles. A percentile is an outcome, within the sampling space of the R.V. such that a given percent of the population scores less than or equal to such outcome. For example, the median stress is the fiftieth percentile of all stress values to which a material is subjected to. This is so, because half the population scores less than, and the other half more than this stress value.

Other important percentiles are the lower (upper) quartiles, also called first (third) quartiles and denoted Q1 and Q3. These percentiles define values where 25% of the population (or 75% of the population) score less than or equal to such values. In general, percentiles are of great interests because they define a value such that a pre-specified percent of the entire population scores less than or equal to it. This is what engineers need for designing parts: a value such that an overwhelming percentage scores better that this value. Then, the engineer uses it as a design parameter and calls it an "allowable".

In our three examples, we can verify how the values of the first and third Quartiles (Q1 and Q3) do differ, though not markedly. At the center, the three distributions behave somewhat similarly, as expected, since they were generated with similar means and variances. But the real difference in behavior, also as expected, is in the tails of the distributions. Such tail behavior is of extreme importance in statistics and especially in materials data analysis.

For example, in MIL-HDBK-5 & 17, there is great interest in estimating the first and tenth percentiles of the population distributions under study. These percentile estimations are then used to obtain the A and B basis allowables. Their importance is underscored by the fact that 99% (or 90%) of the corresponding population values for a given performance measure of interest (say tensile stress) will be larger than this value. In our three examples, this situation is highlighted by the values of the corresponding minimums, which differ markedly from each other, the Weibull being the smallest, for it is left skewed.

Population percentiles in the Normal distribution are obtained via the Normal Standard table. Percentiles in the Lognormal distribution are obtained using the logarithms of the values in the Normal tables. In the Weibull distribution, percentiles are obtained by calculating the values from the closed form cumulative distribution function.

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 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 it.

In the tensile strength examples developed above, the population average is approximately 330 units, with a standard deviation of 5 units. Hence, the three distributions capture the central values quite similarly (e.g. values between quartiles Q1 and Q3). However, this is not so in the tail percentiles. If the true distribution is Normal, then the approximate tenth percentile is: 313. But if the true distribution is Lognormal, the approximate tenth percentile is now: 309. If the true distribution is Weibull this approximate percentile becomes 296. Again, this signals the great importance of selecting the appropriate distribution and also of correctly estimating its parameters. For, these first and tenth percentiles will frame the A and B basis allowables, frequently used in design and in materials data analysis.

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.

An interpretation of the standard deviation of a Normal distribution is the distance from the mean to the density function inflection point, in both directions. The standard deviation also defines what percentage of the population is under the Normal density. For example, if tensile strength is normally distributed, independently of the value of its mean and variance there are always about 70% of tensile strength population values in the interval between one standard deviation below and one standard deviation above, the mean. There are about 95% of population tensile strength values in the interval between two standard deviations below and two above, the mean. In our example, there are about 70% of possible tensile strength measurements between 325 and 335 and about 95% between 320 and 340.

If the population distribution is not Normal, e.g. if it is Lognormal or Weibull, the above explained percentage population values no longer hold. Therefore, there is no specific reason to add/subtract units of the standard deviation from the mean, any more. For, the population percentages in these intervals are now totally different than before.

A better estimation of dispersion about the center, in such cases, is the Interquartile Range (IQR). This is the difference between the (Upper/Lower) quartiles: Q3-Q1. These are useful alternative measures of dispersion, when the population distribution is not symmetric, as is the case of Weibull and Lognormal. Irrespective of the population distribution, IQR always contains the 50% of the population, closest to the center (median). In the Normal example IQR=333.27-326.8=6.47; in the Lognormal, IQR=334.8-325.78=9.02, and in the Weibull example, IQR=331.75-323.17=8.57, relatively close, as the standard deviations were.

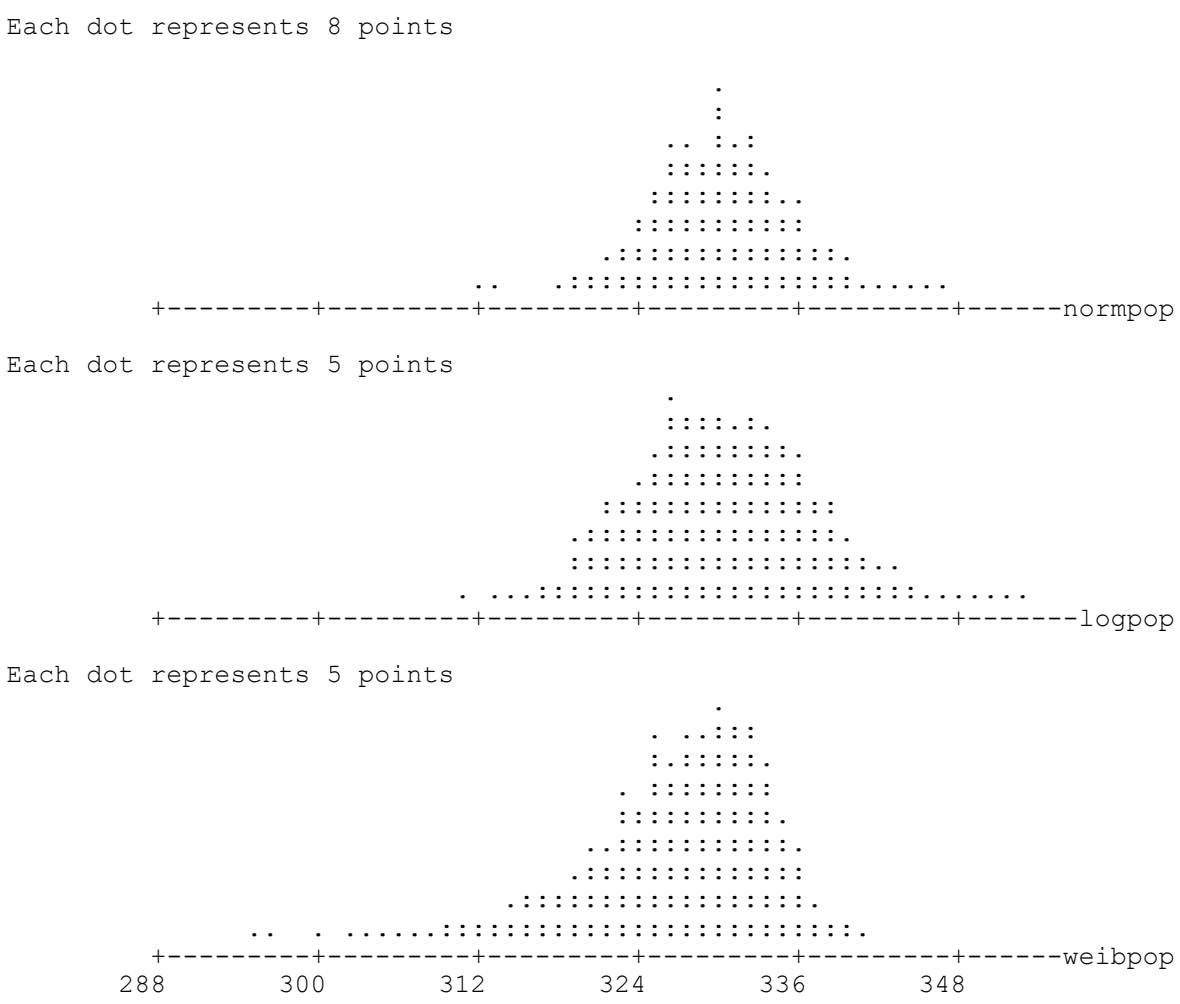
Dispersion parameters are often used to characterize or compare population variability. If means of positive R.V. are the same, their variances can be compared directly. But if the means differ, then an indirect dispersion parameter, such as the coefficient of variation (defined as the ratio of the standard deviation to the mean) is used. The usefulness of the variance loses to IQR, as distributions depart from symmetry, for the same reasons that the mean loses to median and mode. In our Normal example CV=4.91/330.12=0.014 and in the Lognormal CV=6.9/330.03=0.021 (variation is similar). But the variances could have been compared directly, since both distributions had the same mean of 330.

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, true distribution). This is especially so in the cases of the Weibull and Lognormal, where the shape and scale parameters allow them to be extremely versatile and thus useful for fitting many types of data. Examples of different Lognormal density shapes, for various shape parameters, are shown in Figure 3.2. Examples of different Weibull density shapes, for various shape parameters, are shown in Figure 3.3.

Other useful distribution parameters include the threshold, which provides a lower bound for the range of possible outcomes. The threshold is used when, for some technical reason, measurement values can never occur below this pre-specified lower limit. Three-parameter Lognormal and Weibull are good examples of such three-parameter distributions. Their discussion, however, is beyond the scope of this SOAR. The appendix reference [11] provides extensive treatment of this advanced topic and of their uses.

It is worth noticing how, in the Lognormal and Weibull distributions, mean and variance are obtained as a function of the shape and scale parameters, while in the Normal, they are obtained directly and have a specific meaning. Finally, skewness describes the degree of (dis)symmetry and kurtosis that of peakedness, of a distribution.

In all cases, the distribution parameters help visualize the density functions or patterns of possible R.V. outcomes (e.g. patterns of tensile strength, stress, strain, etc). We illustrate some of these visualizations in the comparative dotplots, below, on the same scale:



In the next section we develop EDA (exploratory data analyses) examples from the three above discussed distributions. We draw small samples, simulated from these distributions, and try to determine from which one they came from and what their parameter values are, in order to recreate their original (but supposedly unknown) pattern.

For, the main problem in statistics is that the true, underlying distribution is seldom known. Statisticians, therefore, have to observe (sample) the process of interest. Then, based on this observation, conjecture (estimate) about the unknown distribution and its parameters. Then, verify (test) these conjectures and, either use them or try another one, accordingly.

Example of Materials Exploratory Data Analysis

Random samples, usually small, are taken from the process under study with the objective of estimating the unknown distribution (pattern) of the R.V. of interest, say, tensile strength, stress, load, etc. and of its parameters. We will next develop three such examples, one for each distribution (Normal, Lognormal and Weibull) discussed above.

To visualize the underlying distribution and establish the first conjectures about it, we take samples of size twenty (reasonable in this context) and perform EDA on them. We obtain the descriptive statistics (mean, median, variance, etc.) discussed above. Then we obtain the stem-and-leaf, box-and-whiskers (boxplot) and probability plots, We contrast these results, comparing the known (simulated) distribution versus the estimated ones.

We describe below the (EDA) exploratory procedures used, via a data set (sample) of 20 equiprobable points between 320 and 340. Their true distribution (Uniform) differs from either Normal, Lognormal or Weibull, but their mean and variance are similar. The set is:

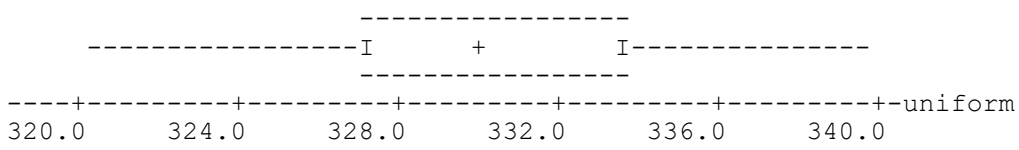
324.067	339.668	326.397	328.389	327.287	329.930	335.063
331.985	339.055	331.064	329.806	320.234	327.432	328.192
322.861	339.008	332.106	331.224	338.830	330.281	

	N	MEAN	STDEV	MIN	MAX	Q1	Q3
uniform	20	330.64	5.51	320.23	339.67	327.32	334.32

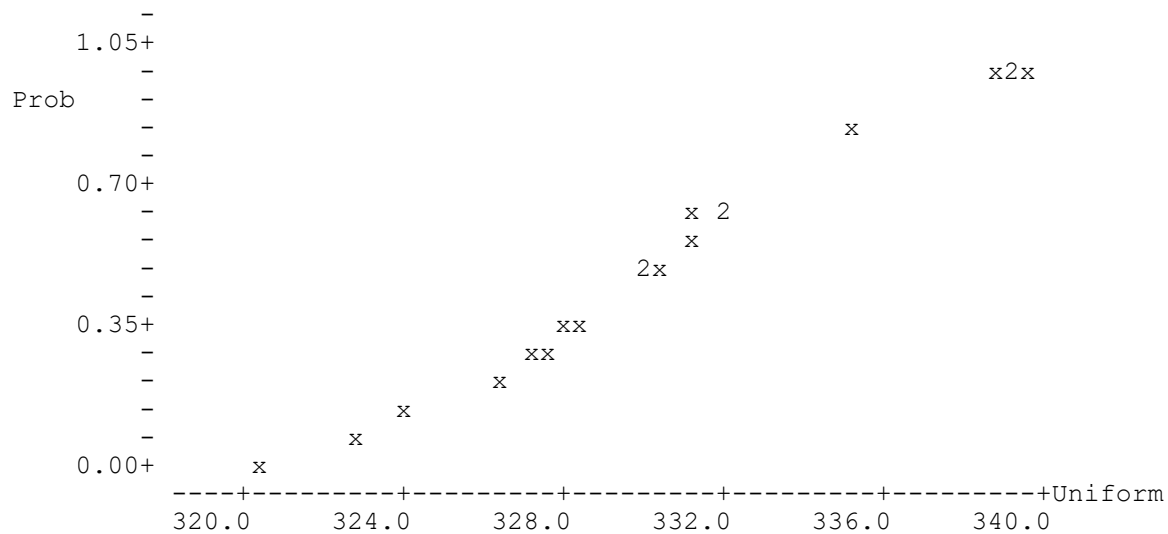
A stem-and-leaf plot is a table where data is sorted and then written down by line. Each line defines a class, as in a histogram. Each class is defined by the most significant digit(s) in the range of data. For example if data ranges from the fifties to the nineties there may be a class of 50s, then of 60s, up to 90s, where the first digit is written on the margin. Then, the second digit (say a 3 in a 53, a 5 in a 65, etc.) is written in the corresponding line. The result is a table that resembles a histogram, but which keeps the individual entries. This way, we can also obtain the quartiles, the median, and the maximum and minimums. In our example the classes are defined as 320s, 322s, 324s, etc. For example, number 324 (written 4) is the 3rd sorted and belongs to the class of 32. The 'four' appears to the right, in the same line.

```
 1   32 0
 2   32 2
 3   32 4
 6   32 677
10   32 8899
10   33 0111
 6   33 2
 5   33 5
 4   33
 4   33 8999
```

The boxplot is a plot on an axis having the measurements of the variable of interest. From the above stem-and-leaf, we obtain the five-number summary: minimum, Q1, median, Q3, maximum of the sample. We plot them on this line and draw a "box" between Q1 and Q3, where we signal the median. This plot helps visualize whether the distribution is symmetric, how dispersed it is and possible outliers. The "box" contains the 50% of centered data and ends in the corresponding quartiles Q1 and Q3. The "+" indicates the sample median value. Lower and upper single lines denote the ranges for the upper and lower 25% of the data. Outliers would be denoted by "0" or "*". The boxplot for the example data set is:



A Probability-plot is a bivariate plot of the probability of the original data (obtained under the assumed distribution and parameters) versus the original data values. The closer to a straight line, the better it fits the hypothesized distribution. In this case, the distribution hypothesized (Normal) is wrong, but the parameters (330 and 5) are correct. The P-plot is:



Additional information about how to develop and implement these exploratory data analysis methods can be obtained from the appendix references [8, 9 and 12].

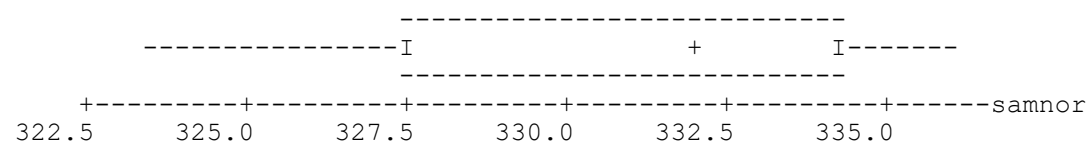
Example from the Normal Distribution: the "samnor" data set

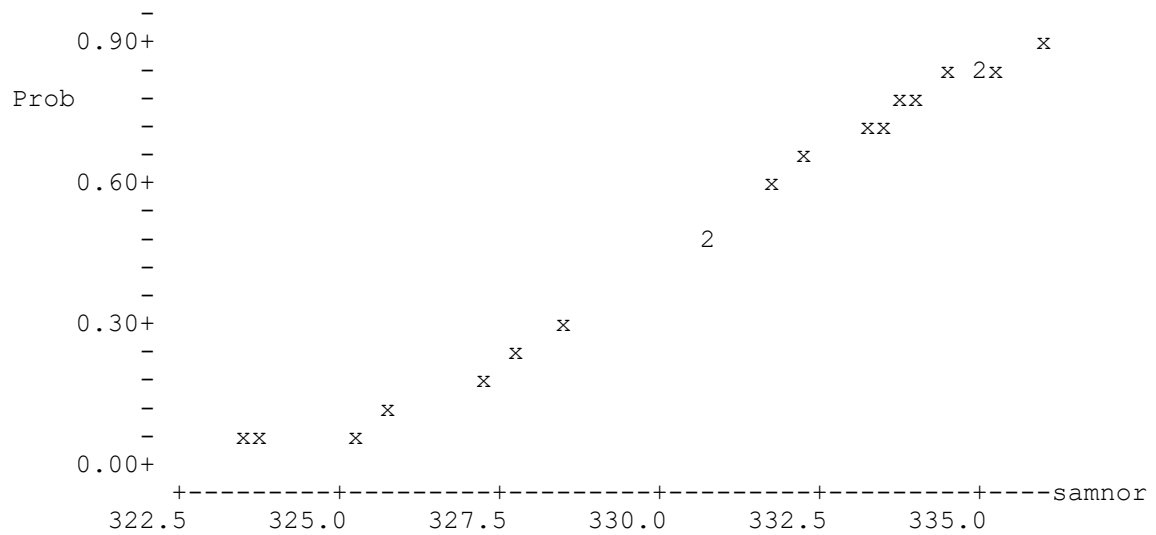
We generate a sample of twenty Normal(330, 5) data points, and perform similar analyses.

```
samnor
  331.791   332.356   334.967   333.400   333.283   323.474   325.869
  328.466   330.706   327.337   334.580   327.715   335.101   333.980
  323.853   325.257   333.776   335.254   330.874   336.022
```

We first obtain the descriptive statistics, stem-and-leaf, boxplot and Probability plot:

```
                N     MEAN   MEDIAN   STDEV  MIN    MAX     Q1     Q3
samnor         20   330.90   332.07    4.08 323.47 336.02 327.43 334.43


    2   32 33
    4   32 55
    6   32 77
    7   32 8
   10   33 001
   10   33 23333
    5   33 4455
    1   33 6
```



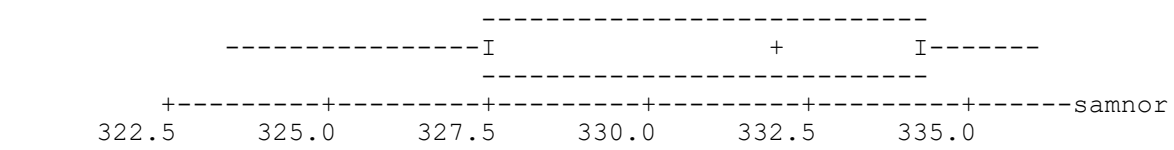


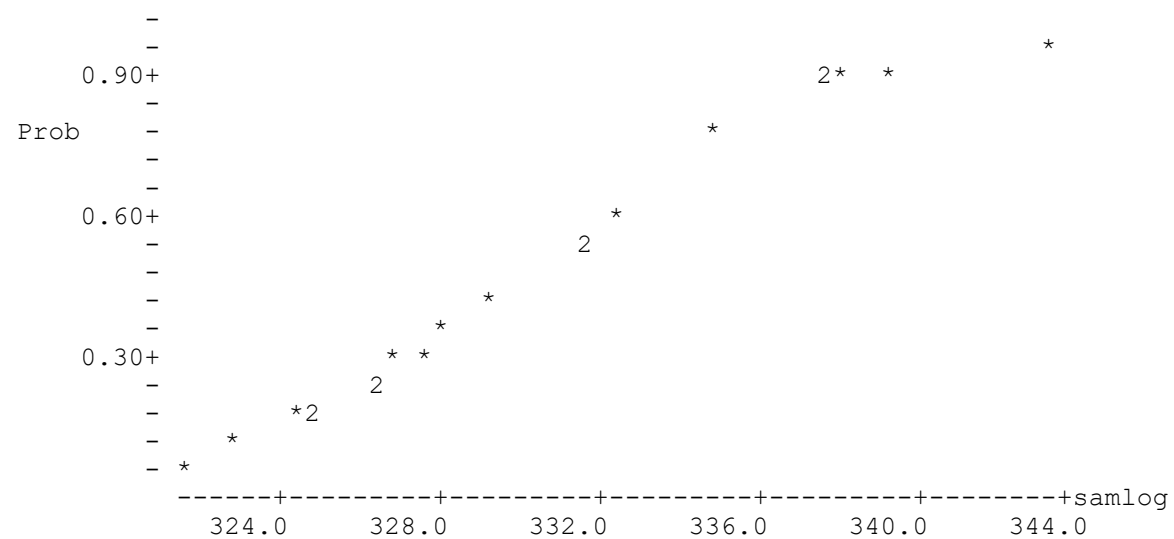
The Anderson Darling (AD) Normality test (which will be explained in the next chapter) yields a p-value of 0.059 (the test is borderline). Hence the (true) Normality of the data is not rejected. The boxplot, stem and probability plots (Figure 3.4) and the descriptive statistics, do not greatly contradict the assumption of the data being Normally distributed either, especially considering the small sample size (20) that we are working with.

Example from the Lognormal Distribution: the "samlog' data set:

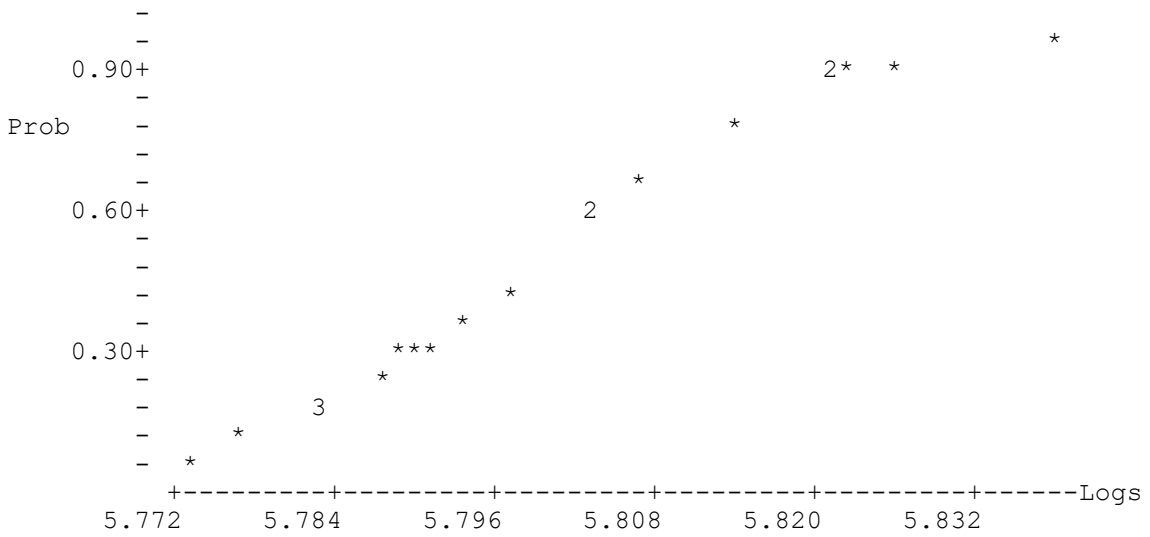
We also obtain the descriptive statistics, stem-and-leaf, boxplot and Probability plot:

```
samlog
  322.622   339.015   337.496   326.577   337.969   334.904   321.611
  326.909   331.485   327.420   332.434   331.427   343.294   326.406
  329.342   324.788   337.555   324.536   324.845   328.083

               N     MEAN   MEDIAN STDEV  MIN     MAX      Q1      Q3
samlog        20   330.44   328.71  6.15 321.61  343.29  325.24  336.85

    1   32 1
    2   32 2
    5   32 444
    9   32 6667
   (2)  32 89
    9   33 11
    7   33 2
    6   33 4
    5   33 777
    2   33 9
    1   34
    1   34 3
```





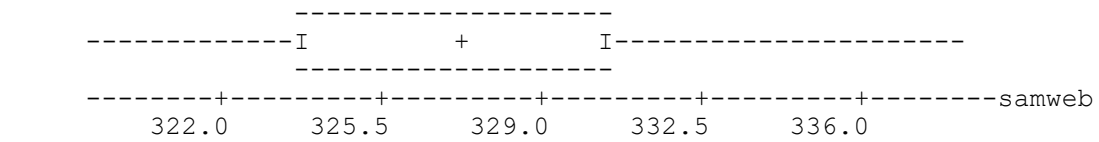
The Anderson Darling test does not reject the Normality assumption, even when these data do come from the Lognormal distribution. The same comment regarding small sample size, applies to boxplot and the other displays. The transformed (Logarithmic) data, which by definition is Normal, also passes the Normality test. The probability plot for Log data is:



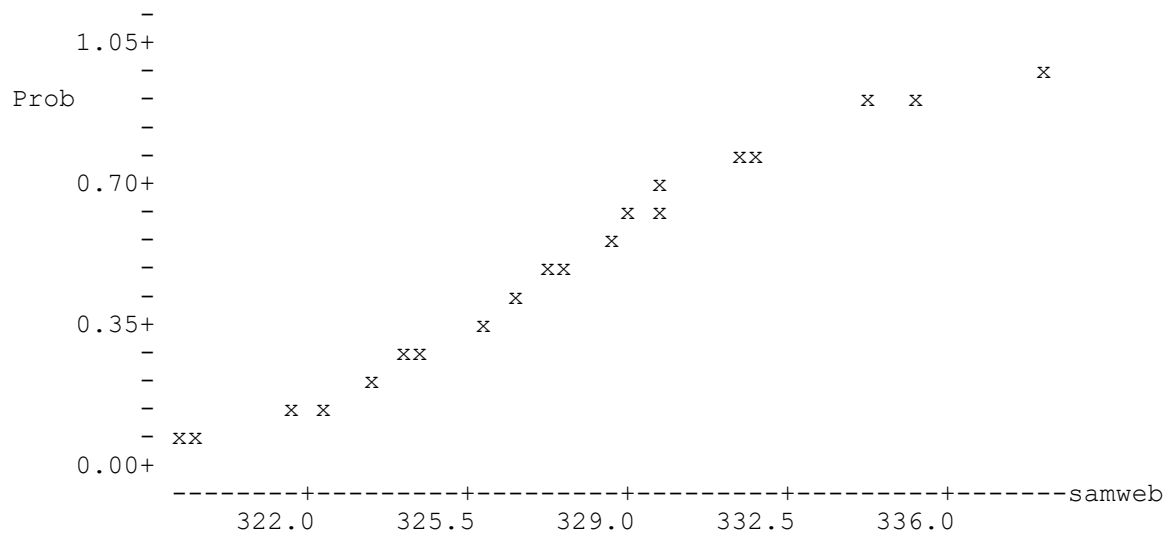
Example from the Weibull Distribution: the "samweb" data set

We also obtain the descriptive statistics, stem-and-leaf, boxplot and Probability plot:

```
samweb
  327.499   321.732   335.454   319.592   329.087   324.084   323.503
  329.799   331.301   338.228   322.344   325.816   319.262   334.181
  324.332   327.348   329.534   326.618   328.533   331.875

                N     MEAN   MEDIAN  STDEV   MIN    MAX     Q1     Q3
samweb         20   327.51   327.42   5.14 319.26  338.23 323.65 330.93
```



This sample also passed the Anderson Darling test for Normality (when in fact it does come from the Weibull distribution). In addition, notice how the right tail is longer than the left one (see boxplot) in the sample, contrary to the situation in the population. The linear trend of the Probability plot below, also provides plausibility for the Normal distribution.



Summarizing, the four small samples examined above have all been drawn from different populations (Uniform, Normal, Lognormal and Weibull) with similar means and variances (means close to 330 and standard deviations close to 5). Since the four distribution patterns are close and the sample size is small, they can all be approximated by the Normal. The real pattern difference, also shown above, lies in the tails of the distributions and affects the calculations of small percentiles such as those defining the A and B basis allowables. This illustrates the crucial problem at the heart of the statistical work in materials analysis.

Extreme Values or Outliers

When working with data, we first have to establish a distribution (with its corresponding parameters) that accurately characterizes the random phenomenon (e.g. Fit the data). Then, we proceed to analyze the sample behavior in the tails of the fitted distribution. This is particularly important in hypothesis testing (which will be discussed next chapter). For it allows us to assess whether an (unusual) observation, assuming a specific distribution and parameters, has an unreasonably small probability of occurrence: i.e. it is an "outlier".

For example, a particular tensile strength observation may be an outlier, if it is assumed to come from a pattern of tensile strengths distributed Lognormal. But it may well be within specs if the assumed distribution is Weibull. This is especially important in the tails, where the probabilities of occurrence are very small and the difference between such probabilities under two distributions, may be relatively very large. In the examples presented above, we saw that a tensile strength value of 310 had a much higher probability of occurrence, if we assumed the Weibull distribution, than if we assumed the Normal.

For, an outlier is defined as an observed value, in the tails of the assumed distribution (pattern of possible outcomes, say of tensile strength) that occurs with a very small probability. It is 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. We may perform

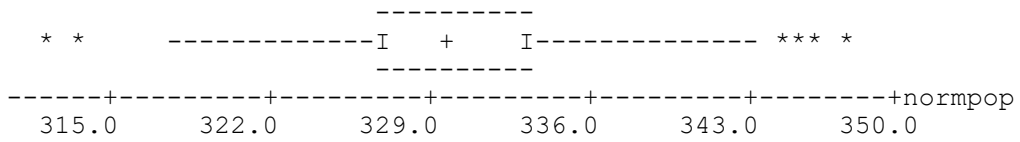
the dice experiment three consecutive times and get three sums of 12 (an event that occurs with probability 2.14×10^{-5}, very small but not zero). We may then conclude that the dice are fixed or otherwise that we are extremely (un)lucky.

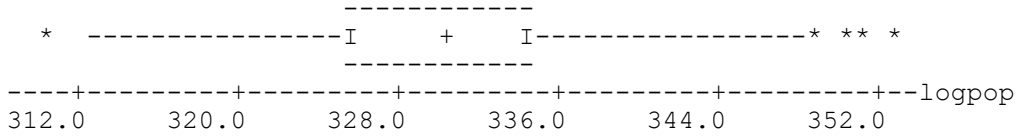
However, we may be disregarding important information, if we mechanically decide to discard it as erroneous, simply because we have detected an "extreme value" (outlier). For example, it may occur that, in a sheet metal process, we observe with probability 0.001 that a large number of cracks per sheet are obtained.

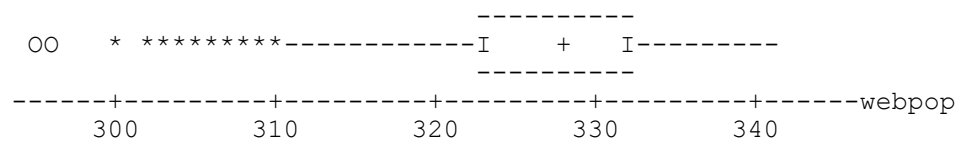
In fact, it may occur that a rare combination of humidity, room temperature, pressure and defective metal composition (combinations that occur with probability 0.001) always produces such poor quality material. However, in the Lab where this testing is taking place, for some special reason, such rare combination arises often. 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 events. We could then, by better controlling the room temperature and other production factors, remove the real problem (instead of the outlier that points toward this problem) and reduce the overall process variability. This is a better use of statistics.

Outliers or extreme values indeed raise a red flag - but do not insure foul play. Statistics provides a useful, scientific context in which to analyze such result -but not a mechanical working rule. On the other hand, in many cases there is indeed a clerical error that can be corrected, or some extraneous circumstances that warrant discarding the data, because it no longer represents the population under analysis (e.g. metal thickness is different). Only in this case it is adequate to remove it from the data set (and put it somewhere else).

Below, we show three comparative displays of boxplots, from the three above developed examples. In them, potential outliers are represented by "*" and "0". To "assess" potential outliers in boxplots, one uses the method of "fences" (see chapter 3 of [8]). One first computes the sample IQR. Then, one subtracts/adds 1.5 and 3 times the IQR to quartiles Q1 and Q3, obtaining two "fences". The rule of thumb is that observations (denoted with asterisks) lying within the 'inner' and 'outer' fences, should be regarded with care. Observations lying outside the 'outer' fence (denoted with 0) i.e. that lie outside 3 times the IQR, should be considered as potential outliers and should be reviewed very carefully.







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 and examination, 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 a different interpretation (i.e. have different probabilities of occurrence).

The objective of the present chapter is, precisely, to provide initial insight into the statistical thinking process, so that the engineer or practitioner can improve the use of statistics as an everyday analysis tool. To that effect we have developed examples from the three statistical distributions used in handbooks [6, 7] procedures: Normal, Lognormal and Weibull.

These examples show a few but important points. First, that given some general parameters such as the mean and the variance, several statistical distributions can approximate the true pattern of outcomes, especially toward the center of the distribution. Secondly, that the main difference between distribution models occurs in the tails, where the interesting statistical work in materials data analysis is done. This is so because the small tail percentiles define the A and B basis allowables, of interest in materials data analysis and design. The differences in tail behavior, between two distribution models (and even within one model, for different parameter values), may be of significant importance.

Finally, we have shown how it is extremely difficult to differentiate between two statistical distributions (models) when the sample size is not large. We have used here samples of size twenty, which in many occasions is much more than the materials engineer can afford, due to cost and time considerations, in real life data analysis or in experimental work. This is a real life constrain that we have to learn to live with and manage, as best we can.

A last word should be said about non-parametric statistics. Because of all the considerations above, it is often useful and even desirable not to specify a distribution. If the sample is large and under certain circumstances, distribution free (also called non-parametric) statistics can be obtained. These statistics are based on general probabilistic considerations (outside of the scope of this SOAR) rather than on specific parametric families of distributions such as the three discussed above. Non-parametric results are usually conservative. But when there is doubt regarding which is the correct or adequate distribution to use, non-parametric statistics is a better solution.

With this understanding of random variables, statistical distributions and its corresponding parameters, and hence, of what outliers or potential extraneous observations are, the reader is in position to advance to the next chapter, that deals with testing and estimation.

Suggested Further Reading

An Introduction to Probability Theory and Mathematical Statistics. Rohatgi, V. K. Wiley. New York. 1976.
"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.