Chapter Four:
On Estimation and Testing

Jorge Luis Romeu
IIT Research Institute
Rome, NY 13440

June 10, 1999

Executive Summary

In this chapter we discuss problems and present examples related with estimation and testing. Every random process (or R.V.) follows a distinctive pattern (distribution). Such distribution can be uniquely specified by a set of fixed values or parameters. Once these two elements are established we can answer all pertinent questions regarding the random process and take the necessary decisions to control, forecast or effect its course.

Unfortunately, the R.V. distribution and its associated parameters are usually unknown. Then, the best that we can do is to observe the process (i.e. sample). Then, we use these observations (sample) to reconstruct both, the distribution and the parameters that generated them (estimation), or to confirm or reject some educated guess that we have previously formed, about these distribution and parameters (hypothesis testing).

Estimation

Sampling Revisited

Statistics is about taking (optimal) decisions under uncertainty. For example, we want to find a value, such that 90% of tensile strengths from a specific type of material is above this value with 95% confidence. Such a value is known, in materials data analysis, as a B-basis allowable. However, we deal with a random process (say, the measurement of tensile strengths in the material in question) whose distribution and parameters we ignore. We would like to establish this distribution. For then, we would be able to define an optimal strategy vis-à-vis this random process (e.g. define the B-basis allowable).

Hence, we observe (sample) the tensile strength process for as long as we can afford. Sampling's first assumption is that the process is stable (i.e. that the conditions prevailing during the observation period remain the same during the extrapolation period). Then, the sample must be taken at random from the population of interest. For, the sample has to be "representative" of the population it comes from. For example, we cannot obtain an allowable for a material fabricated under "special" lab conditions and then use it for design with materials produced under normal industrial production conditions.

Random sampling schemes share two common qualities. First, all individuals in the population (in sampling with replacement) or all possible samples (in sampling without replacement) must have the same probability of selection. Second, that sampling can be very expensive (either in time, or in money or in both). For this latter reason, often sample sizes are not very large. This size problem, as we have seen in the previous chapters, can have a large impact in the statistic's resulting underlying distribution. Once such a distribution is established, the sample has again, as we will see in this chapter, an important influence in the estimation of the distribution parameters.

Point Estimation

A random sample of size n provides n pieces of information: vector (x_1, …, x_n). For example, it can be the n=20 tensile strength lab measurements already seen. But a vector is not easy to work with. We need to synthesize its information, i.e. to create a single "statistic" that summarizes most (or possibly all) of the information in the sample vector. Since it is the result of a random (sampling) experiment, such statistic is also a R.V. Therefore, it has its distribution and parameters, which we eventually also need to find.

The "statistic" that we synthesize from the sample is called a "point estimator", for it is just a number (i.e. a point in the real line). In general, we would like this statistic or point estimator to have good qualities. For example, that it is unbiased, efficient and sufficient, among other desirable properties. The sample average and the average of the largest and the smallest values in the sample are two different point estimators of the same parameter population mean (of all possible tensile strength measurements of a given material).

Being unbiased means that the expected value or long run average of the statistic yields the true parameter it is aiming for. The difference between the long run average of the statistic and the true parameter is the bias. A statistic is "efficient" if it has a small variation (or in statistical parlance, a small variance). In the example above, the sample average is a more efficient estimator of the population mean than the average of the largest and smallest values in the sample. A statistic is "sufficient" if it uses all the information in the sample. The sample average is sufficient; the average of the largest and smallest sample elements is not. This is evident because the sample average is computed using all sample values whereas the other statistic only uses two sample points. Hence, many different samples having the same largest and smallest values will yield the same second statistic but not the first one (sample average). Additional information can be found in references [8, 10, 9 and 11] in this order of difficulty.

For the above reasons (i.e. that it is unbiased, efficient and sufficient) the sample average (denoted $\bar{x}$) is a widely used and preferred statistic. In addition, if we have a reasonably large (say, size $n \geq 30$) independent, random sample, from the same distribution (i.e. population of all possible tensile strength measurements) with mean μ and finite variance σ^2 then, by the Central Limit Theorem (CLT), the distribution of sample average $\bar{x}$ is Normal, with the same population mean μ and variance σ^2/n.

The Central Limit Theorem

Let’s illustrate this extremely important CLT theorem with a numerical example. Let’s “roll” an honest die 200 times and obtain the dotplot (sample distribution plot). Let’s redo this operation five times. The dotplots and descriptive statistics are presented below.

Each dot represents 2 points

```
                                           .
             :                             :                   .
             :         :         .         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
          ---+---------+---------+---------+---------+---------+---C1

Each dot represents 2 points
                       .
                       :         :                   :         .
                       :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
          ---+---------+---------+---------+---------+---------+---C2

Each dot represents 2 points
             .                                       :
             :                   :                   :
             :         :         :         .         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
          ---+---------+---------+---------+---------+---------+---C3

Each dot represents 2 points
                                           .
             .         :                   :
             :         :         .         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
          ---+---------+---------+---------+---------+---------+---C4
```

```
Each dot represents 2 points
                                 .         :
                                 :         :         .
             :         .         :         :         :         .
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
             :         :         :         :         :         :
          ---+---------+---------+---------+---------+---------+---C5
           1.0       2.0       3.0       4.0       5.0       6.0
```

```
                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
C1            200    3.500    4.000    3.500    1.731    0.122
C2            200    3.550    3.500    3.556    1.686    0.119
C3            200    3.425    3.000    3.417    1.708    0.121
C4            200    3.460    4.000    3.456    1.698    0.120
C5            200    3.510    4.000    3.511    1.656    0.117
```

Now let's obtain the average of five rolls, for each of the 200 rolls in each sample above, and also calculate their descriptive statistics and dotplot. They are presented below.

```
                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
C6            200   3.4890   3.4000   3.4856   0.7525   0.0532
```

```
                           .  :     .    :  .
                           :  :     :    :  :  .
                           :  :     : .  :  :  :
                           :  :  :  : :  :  :  :
                        :  :  :  :  : :  :  :  :
                        :  :  :  :  : :  :  :  :
                  .  .  :  :  :  :  : :  :  :  :  :     .
                  :  :  :  :  :  :  : :  :  :  :  :  :  :
             .    :  :  :  :  :  :  : :  :  :  :  :  :  : .  .
             :  . :  :  :  :  :  :  : :  :  :  :  :  :  : :  :
          -------+---------+---------+---------+---------+---------C6
              2.10      2.80      3.50      4.20      4.90      5.60
```

Notice the resulting sample of the average of these five rolls: it is unimodal and its mean of 3.489 is similar to (3.5) that of an individual roll. Its standard deviation (0.75) is close to (square root of variance σ^2/n=) $\sqrt{}$(2.9/5)=0.76, as stated in the CLT above. This is only for five rolls and a flat distribution (outcome pattern) very different from Normal. For, since all rolls are equally likely, the outcome pattern of a single roll is completely flat. Therefore, the average of many tensile strength measurements should tend more rapidly to Normality, since the parent distribution of an individual tensile strength measurement (original distribution) has a much closer form to a unimodal and symmetric distribution.

The CLT is a very powerful result. It provides both the statistical table (Normal Standard) we need to use, as well as the necessary parameters (μ, σ) to standardize (i.e. to take it to the Standard Normal, with has μ = 0 and σ = 1) this sample average $\bar{x}$. As seen in the previous chapter, every Normal R.V. can be standardized. Hence, we can obtain the (Standard Normal) distribution z, from the sample average $\bar{x}$, via the transformation:

$$z = \frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}} \qquad (1)$$

Usually, neither the population mean μ nor the variance σ^2 is known. However, when the sample size n is large (n ≥ 30) the point estimator of the variance:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

is close to the true variance σ^2 . Hence, we can use s^2 for σ^2, in (1) above, and its distribution remains approximately Normal Standard.

Interval Estimation of the Mean

When initially observing a random process of, say tensile strengths, we may not have a clear idea of what its distribution is, nor where its parameters lie. Our objective, then, is to "estimate" these values from the sample. We can obtain a point estimator (e.g. the sample average is a point estimator for parameter population mean) as done above. But point estimators may vary widely from one sample to another. Hence, interval estimators (i.e. random intervals that "cover" the fixed parameter with a prescribed probability) are more useful. For, they provide a confidence region where the true distribution parameter may lie, with some specified probability, namely the confidence interval (denoted c.i.).

We already know that, by the CLT, the average $\bar{x}$ of large samples is Normally distributed with mean μ and variance σ^2/n. Since $\bar{x}$ is distributed Normally, it can be standardized. In addition, the Normal property of the standard deviation determines what percentages of the population are always included, under the density function, at fixed distances of the distribution mean. For example, 68% of any normal population lies within one standard deviation above and below the mean and 95% lies within two standard deviations, etc.

But we can also invert the above process to obtain a c.i. for the unknown mean. By the CLT the population of averages $\bar{x}$, of say n>30 tensile strengths measurements from a given material, is Normally distributed with mean μ and variance σ^2/n. Let's prefix a probability, say 0.95=1-α. Then, we can state that 100(1-α)%=95% of all possible averages $\bar{x}$, of the n tensile strengths measurements, will be within approximately two standard deviations above and below the mean (and the other 100α=5% will be outside this interval). We do not know the true mean μ but we have a specific sample average $\bar{x}$. So, inverting the thinking process, we assume that this sample average is one of those 100(1-α)%=95% of all possible averages $\bar{x}$, which lies within two standard deviations of the true mean. Then, with center on this specific sample average $\bar{x}$, we add and subtract two standard deviations from $\bar{x}$ and state that, with probability 0.95=1-α, the true tensile strength population mean μ lies in the c.i. ($\bar{x}$ - 2 $\sigma/\sqrt{n}$, $\bar{x}$ + 2 $\sigma/\sqrt{n}$).

To generalize the process, instead of measuring in integer units of $\sigma/\sqrt{n}$ (the standard deviation of $\bar{x}$) we will measure in arbitrary units (say z_0). This allows us to also define arbitrary probabilities or percentages 100(1-α)% for the c.i. coverage of μ (percent of times that such sample averages will be within z_0 standard deviations $\sigma/\sqrt{n}$ of μ). In mathematics we write $\bar{x} \in (\mu - z_0\, \sigma/\sqrt{n},\ \mu + z_0\, \sigma/\sqrt{n})$. Since all probability distributions add to unit and since the Normal distribution is symmetric about μ, the value z_0 (of standard deviations $\sigma/\sqrt{n}$) will depend on the prefixed probability 1-α (or on $\alpha/2$).

In statistics one denotes the z_0 standard deviations ($\sigma/\sqrt{n}$) by $z_{\alpha/2}$. These $z_{\alpha/2}$ are obtained directly from the Normal Standard tables, once the confidence 1-α is known. For the tensile strength example, the population mean was 330 and the standard deviation, 5. So, for samples of size n=30 and 100(1-α)%=90%, we obtain: 1-α =0.9; α=0.1; α/2=0.05; $z_{\alpha/2}$=1.65. Then at least 90% of the times we sample 30 observations from this population, the sample averages fall within (330-1.65x(5/$\sqrt{}$30), 330+1.65x(5/$\sqrt{}$30))=(328.5, 331.5).

However, we are really interested in a 100(1-α)% c.i. for the unknown population mean μ, given a specific sample mean $\bar{x}$. Inverting the line of thought above developed, we say that the interval $(\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}\ \ ,\ \bar{x} + z_{\alpha/2}\sigma/\sqrt{n})$ **covers** the unknown mean μ with probability (1-α) or at least 100(1-α)% of the time. This is so because at least 100(1-α)% of all such sample averages will be, at most, $z_{\alpha/2}$ standard deviations ($\sigma/\sqrt{n}$) away from the true (unknown) population mean μ. We expect (hope) this specific $\bar{x}$ to be one of those.

We present below a numerical example, using the five (samroll) data sets of 200 die rolls each, that we introduced earlier in this chapter. Their (known) mean is 3.5 and their standard deviation, 1.7. We calculate below, five 80% c.i. for these data sets:

```
                N       MEAN     STDEV  SE MEAN    80.0 PERCENT C.I.
samroll       200      3.500     1.731    0.120  (   3.346,   3.654)
C2            200      3.550     1.686    0.120  (   3.396,   3.704)
C3            200      3.425     1.708    0.120  (   3.271,   3.579)
C4            200      3.460     1.698    0.120  (   3.306,   3.614)
C5            200      3.510     1.656    0.120  (   3.356,   3.664)
```

Notice how the five c.i. lower and upper limits vary, but the fixed parameter value μ=3.5 does not. Also, if readers calculate and plot many of these c.i., they will realize how some do cover the true mean while a few others do not. In the long run, 80% will cover μ.

Let's explain the relation between c.i. length and coverage in the following way. Assume the (fixed but unknown) parameter μ were an invisible coin, sitting on top of a table. Also assume that our 100(1-α)% c.i. were a dish of radius $z_{\alpha/2}\sigma/\sqrt{n}$ that we randomly throw, to cover the coin. Then (under certain conditions) the dish would actually cover the coin

100(1-α)% of the time. The error 100α would be the percentage of times our dish would not cover the coin. Of course, the larger the dish radius $z_{\alpha/2}\sigma/\sqrt{n}$ (or equivalently, the c.i.) the smaller the coverage error α. However, once covered by the dish, the coin can lie anywhere under it. Hence, a dish (c.i.) the size of the entire table would always cover the coin. However, such dish (c.i.) becomes useless, for we are then in the same situation we started with (e.g. the invisible coin can be again on the entire table, under the dish).

Summarizing, the procedure for obtaining confidence intervals for μ, from large samples, is based on the following. By the CLT, the distribution of the average ($\bar{x}$) of a sample of size n is Normal with (unknown) mean μ and standard deviation $\sigma/\sqrt{n}$. If we prescribe a "half width" (or distance H) in both directions of mean μ we obtain a resulting population percentage included in the interval (μ - H, μ + H). If, instead, we prescribe a percentage of the population, say 100(1-α)%=90%, to lie inside interval (μ - H, μ + H), we obtain the resulting H=$z_{\alpha/2}\sigma/\sqrt{n}$. Then, any random sample average $\bar{x}$ (of the population of all possible averages of samples of size n) will be in this interval with probability 1-α (say, 0.9). Also, the furthest $\bar{x}$ can be from μ (either by excess or defect) and still lie in the prescribed interval, is H. Therefore, inverting the above process, we can get the interval ($\bar{x}$-H, $\bar{x}$+H) that, centered in average $\bar{x}$, will "cover" (include) mean μ, with probability 1-α at least 100(1-α)% of the times**.** It is again important to underline that it is the c.i. which is random, varies, and may or may not "cover" the fixed parameter μ.

Finally, the c.i. half width H is directly proportional to $z_{\alpha/2}\,\sigma$ and inversely proportional to $\sqrt{n}$ (sample size). Then, for the large sample case of a c.i. for μ, H is given by:

$$(\bar{x} - H,\ \bar{x} + H) = (\bar{x} - z_{\alpha/2}\sigma/\sqrt{n},\ \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}) \Rightarrow H = z_{\alpha/2}\sigma/\sqrt{n}$$

This equation determines, in the above-mentioned case, the sample size n required for a coverage 1-α, when the natural variability of the R.V. is σ^2. If the tensile strength example requires a 95% c.i., with a half width H of half unit, the sample size required is:

$$n=(z_{\alpha/2}\sigma)^2/H^2 = (1.96 \times 5)^2/(0.5)^2 = 384.15$$

Confidence Bounds

In the confidence interval problem, if we let one of the two c.i. limits become infinity, we have the problem of the derivation of a confidence bound. The situation is all the same, except that now the probability of coverage 1-α is met whenever the mean is greater (for a lower bound) or smaller (for a upper bound) than one value, instead of being within two values, as in the c.i. case. The consequence is that error α is allotted entirely to one side, instead of being split and allotted equally to each side. For, now there is no coverage error except when the mean is smaller (larger) than the lower (upper) bound obtained.

In the case of the tensile strength example, lets assume that we want to derive a 90% lower bound for the population mean. This implies that we want this mean to be bounded below by a value γ, with probability $1\text{-}\alpha=0.9$ (or equivalently, at least $100(1\text{-}\alpha)\%=90\%$ of the times)**.** **T**he bound is thus defined as the semi interval ($\bar{x}$ - z_α $\sigma/\sqrt{n}$, + ∞) where now the value of $z_{\alpha} = z_{0.1}$ = 1.28 and the rest is the same.

Tolerance Limits

So far we have obtained confidence intervals and confidence bounds, which cover the true **parameter** with a pre-specified probability. If in turn, we are interested in obtaining intervals that contain a proportion of the **population** with a pre-specified probability, then we want a tolerance interval instead. The extremes of these types of intervals are called tolerance limits.

A tolerance interval (or limit) is a random interval (variable) such that, with a probability $1\text{-}\alpha$ there is at least a proportion γ of the population that is within (the tolerance interval) or above or below (the tolerance limit) according to the respective case. Materials engineers call special tolerance limit estimators A and B basis values. For, they define the first (tenth) lower percentile estimator, that bounds the mentioned lower one (ten) percent of the population with probability $1\text{-}\alpha = 0.95$.

To provide an intuitive idea of tolerance limits, we generate five samples of size 20 from the tensile strength example. Hence, they follow a Normal(330, 5) distribution. Let's call the exact tenth percentile γ and let's obtain it as: $\Phi\{ (\gamma-\mu)/\sigma \}= \Phi\{(\gamma\text{-}330)/5 \}=0.1$, where $\Phi(*)$ is the Normal distribution evaluated at (*) that is less than or equal to 0.1 (i.e. that leaves behind ten percent of the population). Hence, $(\gamma\text{-}330)/5= -1.28$ (the tenth percentile of the Normal Standard). Hence, $\gamma = 330 - 1.28 \times 5 = 323.6$ is the exact percentile value.

The set 'b-basis' contains the five random values, corresponding to the estimated 10th percentile from the 5 samples of size 20. These values approximately correspond to the 2nd order statistic of each sample (i.e. the 2nd value in the sorted samples). They are:

```
                   N     MEAN    MEDIAN  STDEV MIN      MAX      Q1       Q3
C11               20   329.47   328.89   6.06 319.74 341.24 325.87  333.74
C12               20   330.27   329.40   5.26 323.01 340.18 325.35  333.72
C13               20   329.30   329.14   4.84 319.97 339.85 326.11  331.29
C14               20   331.48   332.31   5.48 321.76 343.05 326.68  334.54
C15               20   331.90   331.70   4.65 323.37 341.17 329.45  334.61
b-basis            5   323.02   323.27   1.59 320.58 324.88 321.62  324.29
```

Hence, the (b-basis) data set consists of the five (tenth percentile) estimations. Their sample average is 323.02, close to the exact value $\gamma=323.6$, and their standard deviation is 1.59. Below, is the boxplot corresponding to these five percentile estimators, which constitute a sample from the distribution of all tenth percentiles, obtained from all possible samples of size twenty, drawn from the above tensile strength population.

```
                            --------------
  *                         I       +    I---------------
                            --------------
----+---------+---------+---------+---------+---------+-bbasis
320.80    321.60    322.40    323.20    324.00    324.80
```

We can proceed to generate several dozens of such random samples of size 20, from a Normal(330, 5) population, select the 2nd order statistic and plot it. Such plot will provide a good approximation to the true distribution of the R.V. 10th percentile from samples of size twenty, drawn from the Normal(330, 5) distribution. This example provides a picture of what the statistics work for obtaining A and B basis values for materials data, really implies and entails, and of the related risks of error.

Summarizing, it is important to note the difference between confidence intervals and confidence bounds as well as between confidence and tolerance intervals and bounds. As seen above, a c.i. provides two (lower/upper) limits within which the parameter is included at least 100(1-α)% of the times we do this. A confidence (upper/lower) bound is a value such that the (percentile) parameter in question is bounded (above/below) by this value at least 100(1-α)% of the times. Therefore, in a c.i., the coverage error α is equally divided between the regions above and below its upper/lower limits. In a confidence bound, the coverage error is committed only in one case. Hence, the entire error probability α is allotted to only one region (either upper/lower).

The main difference between tolerance and confidence intervals/limits can be explained as follows. In a tolerance interval (limit) we are concerned with the coverage of a percentage of the population, as opposed to the (c.i.) which covers a parameter. Hence, when we say that (ξ_1, ξ_2) is a p-tolerance interval for a distribution (population) F, with tolerance coefficient γ, we mean that such random interval (ξ_1, ξ_2) covers at least the pre-specified percentage 100p of the population, with probability γ.

Weibull Distribution

In the previous chapter we saw how the Weibull distribution, with shape parameter β and scale θ, i.e. Weibull (β; θ), is also frequently used in reliability and materials modeling and data analysis. However, the estimation of Weibull's parameters is by no means as straightforward as the above examples of the Normal mean and variance.

There are complex analytical methods and convergence algorithms to obtain them and several computer programs have been developed (see [6, 7]). They are out of the scope of this SOAR. There are also simpler, less exact but practical, graphical procedures that use probability paper. An excellent tutorial on such graphical methods to obtain Weibull parameters is presented on section 4.2 (Weibull Distribution) of reference [12].

Other distributions

The CLT makes the sample average $\bar{x}$ and transformation (1) above one of the most frequently used statistics. However, there are many others and their use depends on the situation at hand. For example, the application of the CLT requires having a large sample size. Only then, $\bar{x}$ is Normally distributed. Otherwise, the distribution is not necessarily Normal, nor symmetric and then, mean and variance may become less informative. In such cases we use other sampling statistics that have associated with them other sampling distributions. Some of these are Student's *t*, Chi-Square and F, also frequently used in estimation and testing.

The distribution of Student's *t statistic*:

$$t = \frac{(\bar{x} - \mu)}{s / \sqrt{n}} \qquad (2)$$

is obtained when (i) the sample size is "small" (less than 30), (ii) the variance σ^2 of the population is unknown (and estimated by s^2) and (iii) the parent distribution is Normal. Student t distribution is symmetric but "flatter" than the Standard Normal and with heavier tails. This is a consequence of having a larger uncertainty, since we have less information than before (e.g. smaller n and unknown σ). We now have to deal with the "degrees of freedom" (d.f.) parameter that is defined (due to the estimation of both mean and variance from the sample) as the number of sample points minus one (n-1).

The variance estimator is given below:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad (3)$$

It yields (via $(n-1)s^2/\sigma^2$) a Chi-Square (χ) Distribution, which can be defined as the sum of "ν" independent, squared Standard Normal R.V., and has ν degrees of freedom associated with it. In turn, the ratio of two independent Chi Square R.V., χ_1 and χ_2 divided by their corresponding d.f. ν_1 and ν_2,:

$$F = \frac{\chi_1 / \nu_1}{\chi_2 / \nu_2} \qquad (4)$$

is distributed as an (Fisher's) F, with ν_1 and ν_2 d.f.

Notice how, all three distributions above (Student t, Chi-Square and F) require that the R.V. sample average $\bar{x}$ be "centered" (e.g. subtract the population mean μ). The corresponding non-central R.V. *t*, Chi-Square and F are obtained when the originating R.V. $\bar{x}$ are not "centered" (e.g. when μ is no longer the expected value of $\bar{x}$). This difference, related to the "non-centrality parameter", is also used in several testing procedures included in the handbooks [6, 7].

Summarizing, we first draw a random sample of size n, from the population of interest. Then, according to our sample size, to the parent distribution and to the statistical objectives we are pursuing, we synthesize this sample into a statistic (e.g. sample average, sample variance etc.). Then, we determine the corresponding sampling distribution (e.g. z, *t*, Chi Square, F, etc.) and use it for estimation or testing, as needed.

Confidence Intervals for Other Distributions

Analogous philosophy underlies the calculation of c.i. for the mean, when using small samples (e.g. Student t distribution). As an example, assume we have a random sample of five tensile strengths (data set smallnor) taken from the Normal(330, 5) distribution that we have been working with. These five values are:

```
       341.354   323.668   330.221   328.849   327.889


                N      MEAN    MEDIAN    TRMEAN     STDEV    SEMEAN
smallnor        5    330.40    328.85    330.40      6.60      2.95
```

Then, a 95% c.i. for the above small sample, from the Normal distribution, is given by the equation: $(\bar{x} - t_{\alpha/2}\, s/\sqrt{n},\ \bar{x} + t_{\alpha/2}\, s/\sqrt{n})$, where $t\,(\alpha/2\ , n\text{-}1) = t\,(0.025, 4) = 2.776$

```
             N      MEAN     STDEV  SE MEAN    95.0 PERCENT C.I.
smallnor     5    330.40      6.60     2.95  (  322.20,  338.59)
```

A somewhat different approach (but still similar philosophy) is used for the calculation of c.i. for the variance, ratio of two variances, etc. In such cases we use, instead of the Normal Standard and statistic z, other distributions and statistics such as the mentioned Chi-Square, F, etc. The overall philosophy, however, of pre-establishing the coverage probability 1-α and then "inverting" this process on the statistic distribution, as done above for the c.i. of a mean of large and small samples, remains the same as before.

The specifics of such c.i. derivations are beyond the scope of this SOAR. References [8, 9, 10] in the Appendix have the methods for undertaking such derivations and the reader is referred to them for further information.

Hypothesis Testing.

Often, we do have some preconceived idea, requirement or educated guess, regarding the random process under study. Such additional information allows us to implement additional statistical methods. For example, from previous experience we may have established that a parameter (say the population mean μ of tensile strengths of a given material) is equal to a specific value (say μ_0=330). And we would like to verify whether the tensile strength measuring process (or R.V.) under study (say, an incoming batch of the same material) maintains the same value, or whether the parameter mean value has changed. In such cases we are dealing with a hypothesis-testing situation.

Testing for the Mean

To implement a hypothesis test, we first find a suitable estimator of the parameter for which we have made the conjecture (e.g. the large sample average $\bar{x}$ estimates μ, the parameter population mean). Based on our conjecture that the true population mean μ is μ_0=330 (in technical terms, the null hypothesis is $H_{0:}\ \mu = \mu_0$) we derive the sampling distribution of test statistic z, based on the sample average and given in (1) above.

Under hypothesis H_0, if the sample size n is large ($n \geq 30$), statistic z is distributed Normal Standard (see Figure 4.1). When the sample size n is small and the parent distribution is Normal (but the variance σ^2 is unknown and estimated by s^2 from the same sample) the test statistic becomes (2) and its distribution, under H_0 is now Student t, with n-1 d.f.

Our objective is to decide, based upon the result of the hypothesis test, whether our conjecture, as defined in the null hypothesis H_0 is reasonable. That is, whether the value of test statistic z is main stream within its Normal Standard distribution. This occurs when z falls toward the center of the Normal Standard distribution. Alternatively, the test result may constitute a "rare event" according to the null distribution, obtained under $H_{0.}$ In this case, observing such z result has a low probability of occurrence under H_0, which happens when z falls toward the extremes or 'tails' of the null distribution.

In such cases, one of two possibilities exist. First, that our conjecture H_0 (null hypothesis) is incorrect. Then, we fare better by rejecting hypothesis H_0 in favor of the "alternative hypothesis" H_1 (that the tensile strength population mean is other than μ_0=330). For, H_1 (i.e. $\mu \neq 330$) is always the negation of the null hypothesis. Secondly, that we have been terribly unlucky and have drawn, by pure chance, an unusually bad sample. Such sample yielded an unusually different average. Such rare event has occurred precisely to us, something that would happen, when H_0 is true, at most with probability α. We thus reject H_0 and absorb a probability α of taking this potentially wrong decision (Type I error).

The probability α is also known as the significance level or "size of the test". It is the error probability we are willing to commit if we make this (Type I) wrong decision. Probability α also serves to determine the critical value and the critical region of the test.

Summarizing, there are two types of wrong decisions, namely Types I and II errors. They correspond to rejecting H_0 when it is true and accepting H_0 when it is false, respectively. The probability α of committing Type I error is, say 0.05, if we are prepared to reject H_0 when it is true, in the long run, at most once in twenty times. If this probability α is too high, we may want to reduce it to say, one in a hundred or 0.01, etc. As with the c.i., we can reduce α to zero, by adopting the decision rule "always accept H_0". But then, we would be maximizing the Type II error, i.e. of accepting the null when it is false.

To implement any hypothesis test we must first define the two test hypotheses, the test statistic and its distribution under H_o. Then, we must define α (the significance level) and obtain the critical values and the critical regions for the test. For our tensile strength example, we pre-specify α=0.05 and allot it, symmetrically, to the upper/lower tails. This procedure defines both critical values $z_{\alpha/2}$ (see shaded areas in Figure 1) which, for this example and from the Normal Standard tables are 1.96 and -1.96. The two critical regions are the semi intervals from $z_{\alpha/2}$ to infinity and from $-z_{\alpha/2}$ to minus infinity. The decision to reject H_0 is taken if the value z of the sample test statistic (1) falls in either one of these two rejection or critical regions. If statistic z falls outside of this region, we do not reject H_0 and instead assume μ_0 is a reasonable value for the tensile strength population mean.

We have generated a (large) sample of n=30 tensile strength measurements (lgsptst) from the Uniform (320, 340) distribution, with mean of 330 and standard deviation of 5. The descriptive statistics for this data set are presented below.

```
                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
lgsptst        30   328.37   328.50   328.31     5.17     0.94
```

For the test described above ($H_{0:}\ \mu = \mu_0$) assuming known σ of 5, the results are:

```
             N      MEAN    STDEV   SE MEAN        Z    P VALUE
lgsptst     30   328.367    5.169     0.913    -1.79      0.074
```

If the significance level α=0.05, the critical values are plus/minus 1.96. Since the value of test statistic z is -1.79, we cannot reject $H_{0:}\ \mu = \mu_0$. However, if level α=0.1 then the critical values are plus/minus 1.65 and we reject H_0 because statistic z is less than -1.65.

Assume we only have available the six measurements of tensile strengths shown below:

```
        322.185   321.002   333.250   321.796   316.804   330.630
```

which were generated from a Normal(325, 5) population. Assume that the variance of the population is unknown, and estimated from this small sample (smtst). We then use (2), the small sample t-test statistic (with d.f.= n-1=6-1=5), to test the same hypothesis above:

```
             N      MEAN    STDEV   SE MEAN        T    P VALUE
smtst        6    324.28     6.29      2.57    -2.23      0.076
```

If the significance level α=0.05, the critical values t(0.025,5) are plus/minus 2.571. Since the value of test statistic t is –2.23, we cannot reject $H_{0:}\ \mu = \mu_0$. But if α=0.1, the critical values are plus/minus 2.015 and we reject H_0 because statistic t is less than –2.015.

A Conceptual Comparison

Lets explain the hypothesis testing process via a comparison with the American judicial system, using the well-known case of O. J. Simpson. Here, Judge Ito plays the role of the statistician (he directs the process and interprets the rules). There are two hypotheses.

The null (assumed) is that the defendant is innocent. Its negation or alternative is that the defendant is guilty (and must be proven beyond reasonable doubt). The evidence is the data: the bloody gloves, the DNA tests, etc. The Jury, who evaluates the evidence (data), plays the role of the test statistic The Jury can reach one of two possible decisions. It can declare the defendant guilty (reject H_0) when the evidence overwhelmingly contradicts the assumed defendant's innocence (null hypothesis). Or it can continue to assume that the defendant is not guilty, if they cannot convince themselves (i.e. beyond reasonable doubt) that it is guilty. The Jury can thus commit two types of errors. They can convict an innocent defendant (reject the null when it is true) which is Type I, or acquit a guilty person, which is Type II. The Judicial system (and the Statisticians) would like to minimize the probability of either of these two possible errors (i.e. α and β).

Justice System	**Statistical Hypothesis Testing**
Presiding Judge (Ito)	Statistician
Jury (of 12 peers)	Test Statistic (e.g. formula (1) in the text)
Jury Task: process the evidence	Test Statistic Task: synthesize the (data) information vector
Defendant (O.J. Simpson)	Parameter tested (e.g. population mean)
Possibilities: (Not Guilty and Guilty). Always assume the null (Not Guilty) is true unless disproved by data – beyond reasonable doubt	Hypotheses: (null and alternative). Always assume the null hypothesis to obtain the test statistic distribution.
Evidence (glove, DNA test, etc.) Does Evidence (data) overwhelmingly contradict the assumed null hypothesis beyond reasonable doubt?	Data collected (for the test) is synthesized by the test statistic and then compared with its (null) distribution.
Decision: acquit or convict defendant	Decision: Reject or not Reject the null hypothesis
Possible errors (misjudgment) Convict an Innocent Defendant Acquit a Guilty Defendant	Error Types (I and II) Type I: Reject the null when it is true Type II: Accept the null when it is false
Risk of Convicting an Innocent Defendant	Alpha: Probability of Type I error
Risk of Acquitting a Guilty Defendant	Beta: Probability of Type II error

Step-by-Step Testing Procedure

The following is a procedure to implement a hypothesis test:

1. Define the Null ($H_{0,}$) and Alternative (H_1) Hypotheses
2. Define the Test Statistic
3. Identify the Distribution of the Test Statistic
4. Select the Significance Level (α) of the Test
5. Obtain the Critical Value(s)
6. Define the Critical Region(s) for the Test

7. Obtain the Test Statistic Result
8. Obtain the p-value of the Test
9. Compare 7 with 5 or 8 with 4
10. Take a Decision, accordingly.

We illustrate this procedure with the above large sample test for the tensile strengths.

1. Null Hypothesis is: $H_{0:}\ \mu = \mu_0$. ; Alternative Hypothesis is: $H_{1:}\ \mu \neq \mu_0$.
2. Test statistic z is given by equation (1)
3. Distribution of statistic z is Standard Normal (under H_0)
4. Significance level α=0.05
5. Critical values $z_{\alpha/2}$ are plus/minus 1.96.
6. From $z_{\alpha/2}$ to infinity and from $-z_{\alpha/2}$ to minus infinity
7. Statistic value: $z = -1.79$
8. P-value = 0.074
9. Statistic z is not smaller than critical value - $z_{\alpha/2}$
10. Decision: do not reject (but assume) the null hypothesis

The information required for implementing the small sample tensile strength table is also in the above test. It is left to the reader to do likewise as above, as an exercise.

Other Issues About Hypothesis Testing

There are several types of hypothesis tests. Up to now, we have seen the two sided case (e.g. the ones discussed in the examples above). There is also the one sided case. Often, we are not interested in the exact value of a parameter (say that the true population mean μ is exactly μ_0). Instead, we may want to test whether the mean μ is greater or smaller than a specific value (say μ_0). In such case, the null hypothesis H_0 becomes: $\mu \geq \mu_0$ or $\mu \leq \mu_0$, accordingly. Such types of hypothesis tests, which are fairly common, are called one-sided tests and have a single critical value and critical region.

From the above discussion, we see the one-to-one relation between two sided hypothesis tests and the derivation of confidence intervals, and one-sided hypothesis tests and the derivation of confidence bounds. For example, for a given data set (sample) and a specified significance level α, if a two-sided test for mean μ_0 rejects hypothesis H_0, then the corresponding 100(1-α)% c.i. for μ does not cover μ_0 and vice-versa.

Two widely used hypothesis tests performance measures are the p-value and the Power. They both serve to assess our test decisions, when taken on a specific sample with a specific test. The p-value (also called the Observed Significance Level or OSL in MIL HDBK 17) is the probability of obtaining a test statistic result at least as extreme as the one we have, under the null distribution. It corresponds to the probability of rejecting the null hypothesis H_0 with a test statistic value, as extreme or even more extreme, than the value we have obtained from our sample. In our two sample test examples above, the p-values were, respectively, 0.074 and 0.076. If we (erroneously) reject the null hypotheses with those samples, these would correspond to the two respective probabilities of error. This is why, equivalently, if level α=0.05 (maximum error we are willing to commit) then we do not reject H_0. But if α=0.1 then we do reject the null hypothesis H_0.

The Power of the test is the probability of rejecting H_0, with the test statistic value that we have obtained from our sample. This advanced hypothesis-testing concept is beyond the scope of this SOAR. The interested reader is referred to [8, 9 and 10] in the appendix.

The above situations, regarding hypothesis testing, can only be guaranteed if all test assumptions (statistic distribution under the null, independence and assumed distribution of the raw data, etc.) are met. For example, the z-test (1) for the mean requires that the population variance is known. However, in some cases one or more test assumptions may be relaxed (to a certain point) and the test results remain acceptable. In these cases we say that the test is Robust to (violations of) such assumption. For example, the z-test is robust to the variance assumption, since the substitution of the sample variance s^2 for the population variance σ^2 still yields an approximately Normal Standard distribution for statistic z defined in (1).

When a hypothesis test is invalidated by serious violations of its assumptions, one can still resort to other procedures such as transformations of the raw data or to the use of distribution free (non parametric) tests. When the Normal distribution assumption is not met by the raw data we may, by transforming it, obtain a better fit to a more suitable distribution that fulfills the test assumptions. Otherwise, we may implement a distribution free test. These tests no longer bound the data to some distributional assumptions (e.g. Normality) which are sometimes difficult to obtain, even after implementing a transformation. Distribution free tests, however, are usually less powerful than their parametric counterparts (e.g. they do not reject H_0 when it is false, as often as their parametric counterparts do).

As with everything else, there is always a trade-off involved in test selection, and care must be exercised. Since meeting the Normality assumption definitely provides some advantages, we next discuss ways to test this assumption (via the Anderson Darling test) and other related issues such as testing for outliers in a sample.

Goodness-of-Fit Tests

We have seen that establishing the underlying distribution of a process is of crucial importance for the correct implementation of some statistical procedures. Some such procedures are the small sample t test and the c.i. for the population mean. For, such procedures require that the distribution of the underlying population is Normal (say, that the distribution of possible tensile strengths for the material under analysis is Normal). Therefore, we need to implement first, other statistical procedures to test Normality, before we can correctly implement those statistical procedures (e.g. small sample test for the mean) that require such Normality from the data.

There are several types of statistical procedures that assess a distribution, called Goodness of Fit (GoF) tests. They are essentially based on either of two distribution elements: the CDF or cumulative distribution function, or the pdf or probability density function. GoF tests assume (null hypothesis H_0) a distribution (e.g. Normal) with pre-specified parameters (say, mean and variance). Hence, this H_0 is a "composite" hypothesis (e.g. has more than one element in it, that jointly have to be true). The negation of the null hypothesis H_0 (i.e. the alternative hypothesis H_1) results by failure to achieve any one of the several elements in H_0.

Some GoF tests such as the Chi-Square, use the density function (pdf) of the assumed distribution for assessing the null hypothesis. Since the area under the pdf is unit, one can establish specific proportions of the data set in each subinterval, leading to corresponding "expected" data outcomes in such subintervals. For example, in the tensile strength problem, one can subdivide the range of outcomes into subintervals and "superimpose" the corresponding Normal pdf on it. This will yield the "proportion" of data points that should fall within each subinterval. With it, one obtains the "expected" number and compares it with the observed (actual) number of data points in every subinterval. Then one uses the Chi-Square distribution to assess the probability that such outcome of difference in patterns is mainstream or rare, under the null distribution.

Other GoF tests such as the Anderson-Darling (AD), which is used in MIL HDBK 5 and 17 to test Normality, use the second approach. AD test uses the CDF of the assumed distribution (say, Normal with parameters estimated from the sample) to compare actual versus expected distribution values. Below, we develop two examples with the AD test: one for testing Normality and another for testing the Weibull assumption. If one needs to test for Lognormality, one then Log transforms the original data and then uses the Normality test on the transformed observations, as described below.

The AD GoF test for Normality (reference [6] section 9.6.1) has the functional form:

$$AD = \left[\sum_i \frac{1-2i}{n} \left\{ \ln (F_0 (Z_{(i)}) + \ln (1 - F_0 (Z_{(n+1-i)})) \right\} - n \right]$$

where F_0 stands for the Normal distribution function (with the assumed parameters), n is the sample size, ln is the natural logarithm and subscript "i" runs from 1 to n. The null hypothesis that the true distribution is F_0 with the assumed parameters, is rejected at significance level α=0.05, if the test statistic AD is greater than:

$$AD > 0.752 / (1 + 0.75/n + 2.25/n^2)$$

We illustrate this procedure by testing the data in problem 6 of section 8.3.7 of [7] for Normality. Data set 'prob6' below is composed of six batch averages. For simplicity, we will consider them as six individual values drawn from the same population.

```
                 338.7    308.5    317.7    313.1    322.7    294.2

                    N      MEAN    MEDIAN    TRMEAN     STDEV    SEMEAN
prob6               6    315.82    315.40    315.82     14.85      6.06
```

The AD work for the above formula for the ‘prob6’ data set is presented in the spread sheet table below. Each formula component is in the correspondingly named column:

```
i       Xi      F(Xi)     ln(FXi)   n+1-i    FX(n+1-i)  1-FX(-i)    ln(1-F)

1    294.2    0.072711  -2.62126      6    0.938310   0.061690  -2.78563
2    308.5    0.311031  -1.16786      5    0.678425   0.321575  -1.13453
3    313.1    0.427334  -0.85019      4    0.550371   0.449629  -0.79933
4    317.7    0.550371  -0.59716      3    0.427334   0.572666  -0.55745
5    322.7    0.678425  -0.38798      2    0.311031   0.688969  -0.37256
6    338.7    0.938310  -0.06367      1    0.072711   0.927289  -0.07549
```

The AD statistic yields a value of 0.1699, which is non significant (i.e. does not disprove the plausibility of this sample having been drawn from a Normal(315.8, 14.9) population. For comparison we present the Normal probability plot and AD test (Figure 4.2).

Since we now consider the underlying distribution as Normal, the next step is to assess whether there are potential outliers in this sample. We implement the Maximum Normed Residual (MNR) test for outliers (reference [7], section 8.3.3) for the same example.

This method assumes the underlying distribution is Normal. Then it computes the normed residuals, for each observation, by “standardizing” the sample (i.e. subtracting the sample mean and dividing by the standard deviation). The MNR statistic is the Maximum of the absolute values of all normed residuals. This value is then compared to the critical values in Table 8.5.7 of [7] which is searched by sample size n. The test is significant at level α=0.05 if the largest absolute normed residual is larger than the table value. If a potential outlier is detected, it is removed from the sample and the test is again recalculated on the remaining values, until the sample passes the test.

It should be stressed that potential outliers detected should not be automatically removed from the final data analysis. This test, like all similar ones, signals out potential outliers. These should then be thoroughly checked for consistency. If (clerical or implementation) anomalies are found, then outliers should be corrected or removed. Otherwise, these potential outliers should be left in the original sample. An example of the procedure using the ‘prob6’ (tensile strength) data is shown below:

```
ROW       xi      MNR

   1   294.2   1.45589
   2   308.5   0.49293
   3   313.1   0.18317
   4   317.7   0.12660
   5   322.7   0.46330
   6   338.7   1.54074

   max 'MNR' =  1.5407 < MNR (n=6) = 1.887
```

Since the maximum standardized residual is smaller than 1.887, the critical or table value for this test, we conclude that this data set contains no potential outliers and proceed on.

For illustration purposes, we again use the data set ‘prob6’ (tensile strength) as example to implement the Anderson Darling GoF procedure for assessing the Weibull distribution. The reader can also find this method in section 8.3.4 of reference [7].

We use Weibull probability paper, as explained in [12], to estimate the Weibull shape and scale parameters. These two estimations are 8 and 350. Alternatively, the reader can also use the numerical procedure explained in section 8.3.4 of [7], to estimate them. For our GoF example purposes, a crude estimation is acceptable (for, we know data is Normal).

The statistic (sec. 8.6.4.3 of [7]) for the Weibull version of the AD GoF test statistic is:

$$AD = \left[\sum_i \frac{1-2i}{n} \left\{ \ln\left(1 - \exp\left(-Z_{(i)}\right)\right) - Z_{(n+1-i)} \right\} - n \right]$$

where $Z_{(i)} = [x_{(i)} / \theta*]^{\beta*}$ where the asterisks (*) in the Weibull parameters denote the corresponding estimations and where the rest of the formula is as in the Normal AD case above explained. The OSL (p-value) is given by:

$$OSL = 1/\{1+\exp[-0.1 +1.24 \ln(AD^*) + 4.48 (AD^*)]\} \quad \text{where} \quad AD^* = (1+0.2/\sqrt{n})AD$$

The results of this example are given in the spread sheet table below:

ROW	i	xi	zi	exp(-zi)	ln(1-exp)	n+1-i	z(n+1-i)	i-term
1	1	294.2	0.249228	0.779402	-1.51141	6	0.769090	0.29344
2	2	308.5	0.364332	0.694661	-1.18633	5	0.522213	0.77533
3	3	313.1	0.410129	0.663565	-1.08935	4	0.460886	1.24957
4	4	317.7	0.460886	0.630725	-0.99621	3	0.410129	1.69995
5	5	322.7	0.522213	0.593206	-0.89945	2	0.364332	2.13249
6	6	338.7	0.769090	0.463434	-0.62257	1	0.249228	2.55137

The AD (Weibull) GoF test statistic value is AD* =2.92278, highly significant. The corresponding OSL (probability of rejecting the Weibull(8,350) distribution, erroneously) is extremely small. Hence, we reject the (composite) null hypothesis that the underlined distribution (of the population from where these data were obtained) is Weibull, with shape and scale parameters, respectively, 8 and 350.

Finally, there are many more statistical tests than those we have discussed here. Since our objective is to overview the fundamentals of hypothesis testing, only the simpler cases of testing for a single mean and of GoF tests were presented. As with the other topics treated before, the reader is pointed toward references [4, 5] for further reading and examples.

Summary and Conclusions.

Statistics is about finding and taking the best decisions under uncertainty. We are in a position to do that, once the process underlying distribution and its associated parameters are established. For then, we can use this information to help answer all necessary questions and define the best strategy in dealing with such R.V. or process.

In practice, however, the true distribution of the process (R.V.) and its parameters are usually unknown. Hence, to achieve our objectives (of answering questions and defining the best strategies) we need to "estimate" them. We do this via observing the random process (R.V.) under study and then using these observations (sample) to form educated guesses regarding its unknown distribution and associated parameters. If, due to previous experience we already have some ideas regarding this distribution and its parameters, we test. If we have no idea and want to start by constructing a framework of reference, we estimate. This is what statistical sampling, estimation and testing are about.

Further Suggested Reading

Lehman, E. L. Testing Statistical Hypothesis. Wiley, NY. 1959.
Langford, I. H. and T. Lewis. "Outliers in Multilevel Data". Journal of the Royal Statistical Society (Series A). Vol. 161, Part 2 (1998).