Material EASE

AMPTIAC 6

Jorge Luis Romeu
IIT Research Institute
Rome, New York

# STATISTICAL ANALYSIS OF MATERIAL DATA
## PART II: ON ESTIMATION AND TESTING

### Introduction

In the previous (first) article of this series, random variables (R.V.), distributions and parameters were overviewed and the problem of outliers was briefly discussed. Our objective was to provide practicing engineers with a more thorough understanding of the philosophy behind the statistical procedures they need to apply in their materials work. [1,2].

In this second article we pursue further this objective by discussing problems related with sampling, estimation and testing. We have seen how every random process (or R.V.) has two or more outputs that follow a distinctive pattern (its distribution). And we have seen how such a distribution can be uniquely specified by a set of fixed values or parameters. Once these two elements are known, we can answer all pertinent questions regarding the random process and thus take the necessary decisions to control, forecast or effect its course.

Unfortunately, in almost every case the R.V. distribution and its associated parameters are unknown. Then, the best that we can do is to observe the process (i.e. sample) and use these sample observations 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).

### Sampling

Statistics is about taking (optimal) decisions under uncertainty. We deal with a random process (R.V.) whose distribution and parameters we ignore but would like to know for then we would be able to define the optimal strategy vis-à-vis this random process. Hence, we observe this process for as long as we can afford: this is sampling. Sampling's first assumption is that the process is stable (that the conditions prevailing during the observation period will remain the same during the extrapolation period). Then, the sample must be taken at random, in order for it to be "representative" of the population it comes from [3].

Sampling can take several forms. For example, we can select n subjects at random from a finite population of N individuals (e.g. n light bulbs out of a batch of N). Or we can select them from an infinite population (e.g. roll n times a pair of dice, from the infinite population of possible dice rolls). We can also sample with (or without) replacement according to whether we return (or do not return) each sample subject back to the population, after each drawing. However, (simple, 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 is very expensive (either in time, or in money or in both). For this latter reason, often sample sizes are not very large.

Once a sample of size n is obtained, we need to synthesize it, i.e. to create a "statistic." Since it is the product of a random (sampling) experiment, the statistic is also a R.V. and has its distribution and parameters. For example, the sample average (denoted $\bar{x}$) is a widely used statistic. For, if we have a reasonably large (say, 30 or more) random sample, from the same (unspecified) distribution (i.e. population) with finite mean $\mu$ and variance $\sigma^2$ then, by the Central Limit Theorem (CLT) the distribution of sample average $\bar{x}$ is Normal, with the same mean $\mu$ and variance $\sigma^2$/n. This is a very useful result, for it provides both, the statistical table (distribution) we need to use (Normal Standard) as well as the necessary parameters ($\mu$, $\sigma$) to standardize R.V. (i.e. take it to the Standard Normal, with $\mu = 0$ and $\sigma = 1$). Since every Normal R.V. can be standardized, we obtain the Standard Normal distribution from $\bar{x}$, via the transformation:

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

Due to the CLT, the sample average $\bar{x}$ and transformation (1) above are among the most frequently used statistics. However, there are many others and their use depends on the situation. First, average $\bar{x}$ requires a large sample size. Then $\bar{x}$ is an estimator of the population mean. And as discussed in our first article, mean and variance may become less informative, as the population distribution becomes less symmetric. In such cases we may use other sampling statistics that have associated 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:

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

is obtained when (1) the sample size is "small" (less than 30), (2) the variance $\sigma^2$ of the population is unknown (and estimated by $s^2$) and (3) the parent distribution is Normal. Student t distribution is "flatter" than the Standard Normal, 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 $\sigma$). We now have to deal with the "degrees of freedom" (d.f.) parameter, which depends on the number of sample points (n) minus one (due to the estimation of both mean and variance from the sample).

The variance estimator:

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

yields (via (n -1)s$^2$/$\sigma^2$) a Chi Square ($\chi$) Distribution which can be defined as the sum of "n" independent, squared Standard Normal R.V. and has n degrees of freedom associated with it. The ratio of two independent Chi Square R.V., $\chi_1$ and $\chi_2$ divided by their corresponding d.f. $\nu^1$ and $\nu^2$,:

$$F = \frac{\chi_1 / \nu_1}{\chi_2 / \nu_2} \tag{4}$$

is distributed F, with $\nu^1$ and $\nu^2$ 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 $\mu$). The corresponding non-central R.V. t, Chi Square and F are obtained when the originating R.V. are not "centered" (e.g. when $\mu$ 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 [1,2].

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

## Estimation

In the initial observation of a random process or R.V., we may not have a firm 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). However, point estimators may vary widely from sample to sample. Hence, interval estimators, i.e. random intervals that "cover" the fixed parameter with a prescribed probability, are more efficient. For, they provide a region where the distribution parameter may lie with some specified probability, namely the confidence interval (c.i.).

It is known, by the CLT, that for large samples, the interval ($\bar{x}$ - $z_{\alpha/2}\sigma/\sqrt{n}$, $\bar{x}$ + $z_{\alpha/2}\sigma/\sqrt{n}$), covers the mean $\mu$ with probability (1 - $\alpha$) or 100(1 - $\alpha$)% of the time ($\alpha$ defined as the non coverage probability). This means, for example, that if the (fixed but unknown) parameter $\mu$ were an invisible coin, sitting on top of a table, and our c.i. were a plastic dish (of radius $z_{\alpha/2}\sigma/\sqrt{n}$) that we were throwing, to cover the coin, then (under certain constraints) the dish would actually cover the coin 100(1 - $\alpha$)% of the time. The error 100$\alpha$ would be the percentage of times our dish would not cover the coin. Of course, the larger the dish radius (or c.i.) the smaller the coverage error $\alpha$. However, once covered by the dish, we no longer see where the coin, sitting under it, lies. So, a dish (c.i.) the size of the table would always cover the coin. Only that such c.i. becomes useless, for we are back again in the same situation we started with (e.g. the coin can be on the entire table under the dish).

The procedure for obtaining an interval estimator (c.i.) for $\mu$, from a large sample, 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 $\mu$ and standard deviation $\sigma/\sqrt{n}$. If we prescribe a "half width" or distance H, from both directions of $\mu$, we obtain the population percentage included in this interval ($\mu$ - H, $\mu$ + H). If, instead, we prescribe a percentage of the population (say 90%) to lie inside an interval ($\mu$ - H, $\mu$ + H), we can equally obtain the necessary H. Hence, 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 - $\alpha$ (say, 0.9). Also, the furthest $\bar{x}$ can be from $\mu$ (either by excess or defect) and still lie in the prescribed interval, is H. Hence, inverting the above process, we can equally say that the interval ($\chi$ - H, $\chi$ + H) centered in average $\chi$, will "cover" or include mean $\mu$, with probability 1 - $\alpha$, i.e. 100(1 - $\alpha$)% of the times. Again, it is important to emphasize that it is the c.i. which is random, varies, and may or may not "cover" the fixed parameter $\mu$.

Analogous philosophy underlies the calculation of c.i. for the mean, when using small samples, or for the variance, the ratio of two variances, etc. In such cases, we use some of the other above mentioned distributions and statistics (Student t, Chi Square, F, etc.) instead of the Normal Standard and z. But the philosophy of pre-establishing a coverage probability 1 - $\alpha$ and then "inverting" the process on the statistic distribution, remain the same as explained above.

It is important to recognize that, all other factors remaining equal, the half width H is inversely proportional to sample size. For the large sample c.i. for µ, we see that:

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

And this equation determines the sample size n, required for a coverage 1 - $\alpha$, when the natural variability of the R.V. is $\sigma^2$, in the above mentioned case.

Finally 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) bounds within which the parameter is included 100(1 - $\alpha$)% of the times we do this. A confidence (upper/lower) bound is a value such that the parameter in question is (above/below) this bound 100(1 - $\alpha$)% of the times. Therefore, in a c.i., the coverage error $\alpha$ is equally divided between the regions above and below its upper/lower bounds. In a confidence bound, the coverage error is committed only in one case, hence the entire error probability $\alpha$ is allotted to only one region (either upper/lower).

The main difference between tolerance and confidence intervals/ bounds can be explained as follows. In a tolerance interval (bound) we are now concerned with the coverage of a percentage of the population, as opposed to the (c.i.) coverage of a parameter. Hence, when we say that ($\xi_1$, $\xi_2$) is a tol-

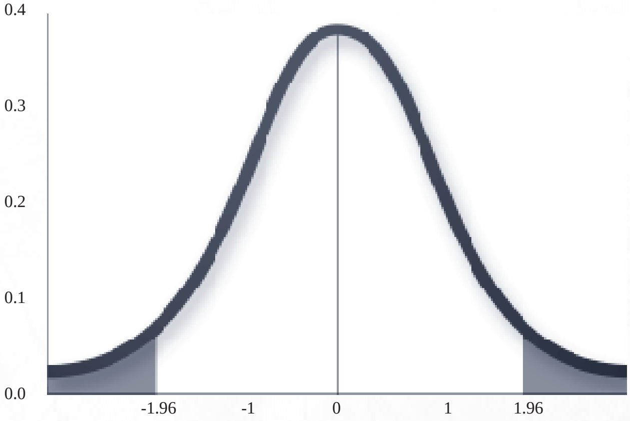

Figure 1. Distribution of the Test Statistic Under the Null Hypotesis ($H_0$).

erance interval for a distribution (population) F, with tolerance coefficient $\gamma$, we mean that, with probability $\gamma$ such random interval covers at least a pre-specified percentage (e.g. 100%) of the population.

### Testing.

Often, we do have some preconceived idea or educated guess, regarding the random process under study. For example, previous experience may have established that a parameter (say the population mean $\mu$) is equal to a given value (say $\mu_0$). And we would like to verify whether the current process (or R.V.) under study maintains this value or whether it has changed. In such cases we are dealing with a hypothesis testing situation.

We first find a suitable estimator of the parameter for which we have made the conjecture (say, large sample average $\overline{x}$ for the population mean $\mu$). Based on our conjecture that the true population mean is $\mu_0$ (in technical terms, the null hypothesis $H_0$: $\mu = \mu_0$) we derive the sampling distribution of test statistic z, given in (1) above. Under $H_0$, z will be distributed Normal Standard (see Figure 1). When the sample size n is small, the parent distribution is Normal and the variance $\sigma^2$ is unknown but estimated by $s^2$ from the sample, the test statistic becomes (2) and its distribution under hypothesis $H_0$ is Student t, with n - 1 d.f.

Our objective here is to decide, based upon the result of the hypothesis test, whether our conjecture, as defined in the null hypothesis $H_0$ is reasonable (e.g. whether the value of the test statistic z, is mainstream in its distribution). Alternatively, the test result may constitute a "rare event" according to the hypothesized null distribution (i.e. this result has a very low probability of occurrence under $H_0$). In such case, one of two possibilities exist. First, our conjecture $H_0$ (null hypothesis) is incorrect. Here, we fare better rejecting $H_0$ in favor of the "alternative hypothesis" $H_1$ (negation of the null, or in this example that the true population mean is other than $\mu_0$). Secondly, that we have been terribly unlucky and such rare event has occurred precisely to us (something that would happen, under $H_0$, at most with probability $\alpha$). Hence, we reject $H_0$ and absorb a probability $\alpha$ of (Type I) error.

Hence, the probability $\alpha$, "size of the test" or significance level is the error we commit if we take this wrong decision. This probability also determines the critical value and the critical region of the test. There are two types of wrong decisions, namely Types I and II errors: rejecting $H_0$ when it is true and accepting $H_0$ when it is false, respectively. The probability $\alpha$ 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 $\alpha$ 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 Type I error to zero by adopting the decision rule "always accept $H_0$". But then we would be maximizing Type II error: accepting the null when it is false.

Once the test hypotheses, the test statistic and its distribution under $H_0$ and $\alpha$ (significance level) are defined, we obtain the critical values and the critical regions for the test. For our example we pre-specify $\alpha = 0.05$ and divide it symmetrically into two upper/lower tails. This procedure defines $z_{\alpha/2}$ (see shaded areas in Figure 1). Hence, both critical values $z_{\alpha/2}$ for this example will be (from the Normal Standard tables) 1.96 and -1.96 and the critical regions, the semi intervals from $z_{\alpha/2}$ up and lower than $-z_{\alpha/2}$. The decision to reject $H_0$ is taken if the value z of test statistic (1) falls in either one of these two rejection or critical regions. In any other case, we do not reject $H_0$ (and hence assume it is reasonable).

Lets explain the hypothesis testing process via a comparison with the judicial system (Table 1). In the well known case of O. J. Simpson, Judge Ito plays the role of the statistician (he directs the process and interprets the rules). There are two hypotheses. The null (assumed) is the defendant is innocent. Its negation or alternative is: 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, plays the role of the test statistic who evaluates the evidence (data). The Jury then reaches 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 declare the defendant not guilty, if they cannot convince themselves (i.e. beyond reasonable doubt) that the defendant is guilty. The Jury can commit two types of errors. They can convict an innocent (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.

There are two types of hypothesis tests: two sided (as the one discussed in the example above) and one sided. Often, we are not interested in the exact value of a parameter (say that the true population mean $\mu$ is exactly $\mu_0$). Instead, we may want to test whether the mean $\mu$ is greater or smaller than a given value (say $\mu_0$). In such case, the null hypothesis $H_0$ becomes: $\mu \geq \mu_0$ or $\mu \leq \mu_0$, accordingly. These hypothesis tests are called one-sided and have a single critical value and critical region.

From the above discussion, we can see that there is a 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 sample and significance level $\alpha$, if a two-sided test for $\mu_0$ rejects hypothesis $H_0$, then the corresponding 100(1 - $\alpha$)% c.i. for $\mu$ does not cover $\mu_0$ and vice-versa.

Two widely used hypothesis tests performance measures are the p

Table 1. Hypothesis testing process

| 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 | Statistic Task: synthesize the (data) information |
| Defendant (O.J. Simpson) | Parameter tested (e.g. population mean) |
| Verdicts (Not Guilty and Guilty, always assume the null –Not Guilty – is true unless disproved by data – beyond reasonable doubt | Hypothesis (null and alternative) |
| Evidence (glove, DNA test, etc.) | Data collected (for the test) |
| Does Evidence (data) overwhelmingly contradict the assumed null hypothesis beyond doubt? | |
| Decision: acquit or convict | Decision: Reject or not Reject the null hypothesis |
| Possible errors (misjudgement) | Error Types (I and II) |
| Convict an Innocent Defendant | Type I: Reject the null when it is true |
| Acquit a Guilty Defendant | 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 |

value and the Power. They both serve to assess our test decision, when taken on a specific sample with a specific test. The p value is 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. The Power of the test is the probability of rejecting $H_0$, with the test statistic value that we have obtained from our sample.

All of the above situations, regarding hypothesis testing, can only be guaranteed if all test assumptions (i.e. statistic distribution under the null, independence and 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 are still acceptable. In these cases we say 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 $\sigma^2$ still yields an approximately Normal Standard distribution for statistic z 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. By transforming the raw data we may obtain a better fit to a more suitable distribution that fulfills the test assumptions. By implementing a distribution free test, we are no longer bound to distribution assumptions (e.g. Normality) which are sometimes difficult to obtain from our data, even after transformation. However, distribution free tests 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 a trade-off involved in test selection, and care must be exercised.

Finally, there are many more types of tests than we have discussed here. Since our objective is to overview the fundamentals of hypothesis testing, only the simple case of the two sided, z-test for a single mean was presented. As with the other topics, the reader is pointed to the references [4, 5] for further reading and examples.

## Summary and Conclusions.

In our first article we overviewed some problems associated with the distribution of a R.V. We also said that, once the R.V. distribution and its associated parameters were known, we could answer all necessary questions and define the best strategy in dealing with such R.V. (or in other words, with taking the best decisions under uncertainty).

In practice, however, the distribution of the R.V. and its parameters are usually unknown. Hence, to achieve our objective (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 the best educated guess regarding its unknown distribution and associated parameters. If, due to previous experience we already have some idea regarding the distribution and its parameters, we test. If we have no idea and want to start constructing a framework of reference, we estimate. This is what sampling, estimation and testing are about.

In the next and final installment of this series of articles, we will apply the (theoretical) concepts discussed in our first two articles to several specific statistical procedures described in MIL-HDBK-5 and MIL-HDBK-17.

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

## Bibliography

1. Metallic Materials and Elements for Aerospace Vehicle Structures MIL HANDBOOK 5G. November 1994.
2. Composite Materials Handbook MIL HANDBOOK 17. 1D.
3. Sampling Theory of Surveys Applications. Sukhatme, P. V. et al. Iowa State University Press. IA (1984).
4. Testing Statistical Hypothesis. Lehman, E. L. Wiley, NY (1959).
5. Statistics: Concepts and Applications. Anderson, D. R., D. J. Sweeney and T. A. Williams. Third Edition. WEST. Saint Paul, MN. (1993).