CHAPTER 7

TEST RESULTS ANALYSIS, PARAMETER ESTIMATION, CONFIDENCE INTERVALS AND HYPOTHESIS TESTING

CONTENTS

1 INTRODUCTION

1.1 Part 4, Chapter 3 (Statistical Distributions) addresses the probability of a specific result given the knowledge of the relevant parameter including its distribution. There is however a second way of approaching the situation. That is, given the results of one or more trials, what can be said about the relevant parameters ?

1.2 Stochasitic parameters such a Reliability cannot be measured directly and repeatably. They can only be measured to a level of statistical confidence. This level of confidence will increase with more data but it can never reach 100% (i.e. you can never be certain of the true reliability). A level of confidence can be calculated, from the practical results, that the real MTBF lies between two bounds or is to one side of a single bound.

1.3 This chapter addresses:

a) Weibull analysis of failure data to assess whether the data is consistent with a constant failure rate regime;

b) the estimation of parameters of a distribution (like the MTBF in the previous paragraph);

c) the level of confidence that a given value or range of a parameter is correct; and

d) the generation and testing of hypotheses about the value of such parameters.

1.4 It may surprise many R&M practitioners to discover that the subject of this leaflet is of great conceptual and philosophical difficulty for statisticians. In fact, the problems of estimation, confidence intervals and the inferences which can be drawn from sample data, have been the subject of fierce debate for 70 years, and are at the core of the controversies between 'Bayesian' and classical statisticians. Since there is no way the subject could possibly be covered in a few paragraphs here, the following discussion is inevitably superficial.

1.5 In general this chapter follows the approach of classical statistics but notes where significant differences occur with the Bayesian approach.

2 WEIBULL ANALYSIS

This section is not required to support the chapters being written at the time of issue. It will be completed later.

3 PARAMETER ESTIMATION

3.1 A population measure, say η, is estimated by a sample measure, say η_s. If the sample size is large, one may expect η_s to lie very close to η; conversely, if the sample size is small, one would be less sure that η_s would lie near to η. η_s is known as a point estimate of η.

3.2 Suppose η_s is measured from a sample of size n. If the sample is returned to the population and another random sample of n drawn, η_s will in general be different for the two samples. If η_s is measured for many such samples, the different values of η_s and their relative frequency form a distribution with PDF $f(\eta_s)$ say.

Figure 1: PDF of sample measure $\square_s$

3.3 It is generally considered desirable in classical statistics to calculate η_S in such a way that the mean, or expected value, of the distribution $f(\eta_S)$ equals the population parameter η. η_S is then called an UNBIASED ESTIMATE of η. For example, when estimating the standard deviation (σ) of a *population* by measuring the standard deviation (s) of a *sample*, the reason that the denominator of the expression for s is made (n-1) rather than n (see para. 3) is in order that s *is* unbiased.

3.4 Although in this chpater, unbiased estimators are generally given, it should be noted that it is not self-evident that unbiased estimators are 'best'[1], or even possible in many applications. In general, it is often difficult to obtain an unbiased estimate of some parameter η. Further, unbiased estimates have certain undesirable properties, perhaps the main one being that if η_s is an unbiased estimate of η, it is *not* generally true that $g(\eta_s)$ is an unbiased estimate of $g(\eta)$, (where g indicates 'function'). It is not proposed to pursue the discussion further here; its main purpose was to show that estimation is not always as obvious and straightforward as it seems, a point which comes over even more strongly in the following paragraphs on Confidence Intervals.

4 CONFIDENCE INTERVALS

4.1 In para. 3.1 it was stated that one makes an estimate (η_s), based on a sample, because it is likely that η_S is near to η. It is a natural progression from this to associate an interval with η_S within which it is likely that η lies. Suppose such an interval is denoted by η_L and η_U, such that $\eta_L < \eta_S < \eta_U$. The classical statistician defines the interval η_L to η_U as the γ% Confidence Interval (CI) on the point estimate η_S, when η_L and η_U are calculated in such a way that if a large number of equal samples were drawn from the population and η_L and η_U calculated for each sample, then about γ% of the intervals would contain the population parameter η. In order to 'locate' η_L and η_U unambiguously, it is normally assumed that in

$\frac{100-\gamma}{2}$% of cases, η would be less than η_L, and in a further $\frac{100-\gamma}{2}$% η would be greater than η_U. (It is not discussed at this stage how to calculate η_L and η_U in general, except to note that it depends on η_S, γ, the sample size and the nature of the parameter being estimated.)

4.2 It is a popular misconception that the above definition of a CI means that there is a γ% probability that η lies in the interval η_L to η_U. Statisticians do not admit that this inference is permissible, one reason being that η is not a random variable. Classical statisticians only make statements about probability of occurrence of a random variable. Having calculated an interval η_L to η_U based on an observed η_S, then η, which is a fixed albeit unknown number, either does or does not lie in the interval. In classical statistics there is no probability about it.

4.3 The growth in the popularity of Bayesian statistics is partly due to a dissatisfaction with the above approach. As indicated in PtDCh2, 'Bayesians' claim that it *is* reasonable to make probability statements about possible values for η, and indeed they formulate PDFs for η, in which a prior PDF for η is modified by sample results using Bayes Theorem to yield a posterior PDF for η. In Bayesian statistics this posterior distribution is used to provide Bayesian probability intervals and the parameter estimate. A Bayesian γ% probability interval *is* a statement that the probability is γ% that η lies in the interval. It should be noted, however, that a 'Bayesian' would not make such a statement using a CI calculated in the classical manner. A Bayesian calculation requires a prior distribution, even of 'prior ignorance'. It is shown in PtDCh2 that even if prior ignorance is expressed as 'all possible values of η are equally likely' (the uniform prior), then Bayesian CIs differ from classical ones. (The classical approach assumes that nothing is known other than what is observed from the sample.)

4.4 In this chapter the CIs offered are calculated in the classical manner.

4.5 In the previous paragraphs two-sided CIs have been discussed; i.e. both η_L and η_U are of interest. Sometimes only one of these is of interest. The CI is then referred to as a ONE-SIDED CI. Suppose a one-sided γ% upper confidence limit is defined by η_U. This means that if η_U is calculated for a large number of equal samples, then η is less than η_U for about γ% of the samples. Thus if η_U is the upper limit of a γ% *two*-sided CI, then $(-\infty,\eta_U)$ would be a $\left(\gamma+\frac{100-\gamma}{2}\right)$% one-sided CI. Similarly, one can think of one-sided lower limits.

5 CONFIDENCE LEVELS

5.1 It has become common in reliability work to quote confidence intervals with low values of γ, of say 50%, 60%, etc. The main justification for this practice appears to be that reliability data is often so bad that using these values for γ is the only way to obtain reasonably narrow CIs. This practice is deprecated for *two-sided* intervals. Many readers of reports look only at the *interval*, without regard to the confidence level, γ, used. They therefore receive a false impression of the accuracy of a measurement if low γ values are used. It is recommended that γ should be at least 80%, and preferably 90% or more when presenting *two*-sided results. If this results in wide CIs, then so be it. At least then it accurately reflects the quality of the parameter estimate.

5.2 However, the 50% *one-sided* limit is of interest, since it provides the MEDIAN UNBIASED ESTIMATOR of η; i.e. if the 50% upper (or lower) confidence bound is calculated for a large number of samples then it will be greater than η in about half the cases and smaller than η in the rest. (Note that the upper and lower 50% *one-sided* limits coincide).

5.3 In practice, for parameters such as MTBF, the 60% one-sided lower confidence limit is normally quoted. This provides a value which errs to the 'safe' side of the median unbiased estimator without producing an unrealistically low value.

6 SIGNIFICANCE OR HYPOTHESIS TESTING

6.1 Hypothesis testing links theoretically derived values for performance parameters to the real world and practical results. It enables a comparison to be made between theoretical estimates or specification values and practical experience. For example, an analyst may have a hypothesis for the value of η. This may arise from:

a) previous work or theoretical assessment having led to the expectation that that η should have the value ξ say; or

b) a specification requirement that needs to be tested.

There is then the wish to examine whether the sample under analysis is consistent with, i.e. supports, the hypothesis, or not. To do so requires a **significance test** or **hypothesis test** to be conducted.

6.2 To perform a significance test the analyst simply examines whether ξ lies in the γ% CI η_L to η_U. If it does, he can say that "η_S is not statistically significantly different from ξ at the (100-γ)% level of significance". This means that the sample data are not inconsistent with the hypothesis that they were drawn from a population in which the parameter of interest had the value ξ. (Note that this should *not* be interpreted as a 'proof' that the original hypothesis is true.) If ξ lies outside the range η_L to η_U then η_S *is* significantly difference from ξ and one can say that if the population parameter were ξ, then the probability of drawing the observed sample is less than (100-γ)%.

6.3 A common error in using significance tests is to use them in situations where one has no prior hypothesis. Analysts are particularly prone to test whether η_S is statistically significantly different from zero, and if this is not so, to simplify the analysis by assuming that they have shown that it is reasonable to take η as zero and proceed on that basis. If there are no *a priori* grounds for believing η to be zero, then there is no reason to take η as zero in preference to any other value in the CI. 'Zero' may only be in the CI because the CI is wide. In fact, the best value to take is η_S. The point to remember is that 'not statistically significantly different from' does *not* mean 'negligibly different from' or 'very near'. A statistically non-significant difference may have considerable practical significance.

6.4 Where significance tests are used repeatedly, for example in lot sampling in a production process, it must be recognised that failures to pass the test will occur from time to time simply due to statistical variation, and that this can occur surprisingly quickly in repeated tests. For example, if an event has a probability of occurrence of 1 in 20, there is a probability of 30% of it occurring at least once in 7 opportunities.

6.5 There are two types of error that can occur in hypothesis testing. First the hypothesis can be rejected when it is true. The probability of this error is usually denoted by α. Second a false hypothesis can be accepted. It is usual to denote this probability as β. Clearly it is desirable to minimise such errors but reducing their probability requires more data (lengthening the test). A compromise has to be achieved where acceptable error probabilities are obtained by an acceptable test.

6.6 It is also common in R&M work to base the risks of error on different hypotheses. A more satisfactory test is produced if two values are tested: an upper test value θ_0 and a lower test value θ_0. The risk of reaching a reject conclusion when η is $> \theta_0$ is α and that of reaching an accept decision when η is $< \theta_1$ is β. The ratio of θ_1 to θ_0 is then known as the discrimination ratio. These concepts are more thoroughly explained in PtDCh10.

6.7 Hypothesis tests can be planned before the data capture takes place (the above text is written on the basis of analysis after capture). This is particularly relevant in R&M work for demonstrations. The principles expressed above still apply but instead of producing an engineering conclusion of statistical correlation, or otherwise, the analyst will produce a test plan. This plan will state:

a) the form of data that is required to be captured and any relevant 'sentencing' rules;

b) the hypotheses that are being tested and the associated risks of error in the result;

c) the criteria under which the test will be terminated with an accept result; and

d) the criteria under which the test will be terminated with a reject result.

7 ESTIMATION OF THE MEAN AND VARIANCE OF ANY DISTRIBUTION

7.1 Notation

7.1.1 In this Chapter the following notation is adopted:

Population Mean	$= \mu$	
Sample Mean	$= \bar{x}$	
Population Standard Deviation	$= \sigma$	(variance σ^2)
Estimate of σ from sample	$= s$	(variance estimate s^2)
Sample size	$= n$	
Random variable	$= x$	
i^{th} x value	$= x_I$	

$f(\eta_S)$ is the PDF associated with the estimator.

7.1.2 Estimating the mean (μ)

$$\bar{x} = \frac{1}{n}.\sum_{i=1}^{n} x_i$$

7.1.3 Estimating the variance (σ^2) with μ not known

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

An equivalent expression which is computationally simpler is:

$$s^2 = \frac{1}{n.(n-1)}.\left(n.\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \right)$$

7.1.4 Estimating the variance (σ^2) μ known.

$$s^2 = \frac{1}{n}.\sum_{i=1}^{n}(x_i - \mu)^2$$

or

$$s^2 = \frac{1}{n}.\sum_{i=1}^{n} x_i^2 - \frac{2.\mu}{n}.\sum_{i=1}^{n} x_i + \mu^2 = \frac{1}{n}.\sum_{i=1}^{n} x_i^2 - 2.\mu.\bar{x} + \mu^2$$

8 ESTIMATION OF THE CONFIDENCE INTERVAL

8.1 CI About $\bar{\mathbf{x}}$

8.1.1 It is not possible to give a general expression for the CI about $\bar{x}$ for any distribution. However, due to the Central Limit Theorem, it is possible to give approximate expressions in cases where n is sufficiently large. The Central Limit Theorem will not be discussed here, except to say that, regardless of the distribution of x, it justifies approximating the distribution of $\bar{x}$ by a Normal distribution with mean μ and variance σ^2/n, where n is sample size and is fairly large.

8.1.2 Thus $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is the Standard Normal Deviate 'z' when n is large; i.e. it is Normally distributed with mean zero and standard deviation 1.

This section is not essential to support demonstrations and will be completed later.

9 ESTIMATING PROBABILITY OF SUCCESS

This section is not essential to support demonstrations and will be completed later.

10 ESTIMATING MTTF AND MTBF

(A more detailed discussion of this topic is given in Ref. 2).

10.1 Estimating MTTF (non-repairable systems)

10.1.1 If a sample of n items are placed on test and the time to failure of each measured, then the estimated MTTF is:

$$\frac{1}{n}\sum_{i=1}^{n} t_i$$

where t_i is the time to failure of the i^{th} item.

10.1.2 Sometimes not all items in such a test fail, for example, if the test is time truncated. In these circumstances it is recommended that some attempt be made to estimate the parameters of the *distribution* of times to failure, then calculate the mean from this. For example, Weibull analysis could be used. The mean could then be estimated from the parameters of the distribution. In general, however, an MTTF cannot be estimated for time truncated tests since the distribution of times to failure beyond the truncation time (T) is unknown. Only if a distribution beyond T is *assumed*, can an MTTF be estimated.

10.1.3 If the times to failure are known to be distributed in accordance with the negative exponential distribution, the MTTF can be estimated rigorously by summing all item operating time and dividing by the number of failures. However, generally in such a test this cannot be assumed, and a Weibull analysis is recommended as preferable.

10.2 CI About MTTF Estimate

10.2.1 If times to failure are distributed exponentially the CI on the MTTF can be obtained as in para 10.4.1.

10.2.2 Where all items are tested to failure, but the distribution is unknown, the Normal approximation can be used if the number of failures exceeds where the standard deviation 's' has to be estimated from the sample and $\bar{x}$ is MTTF as calculated in para. 10.1.1.

10.2.3 If no all items fail, and the procedures suggested in para. 10.1.2 are adopted, the CI cannot easily be obtained.

10.3 Estimating MTBF (repairable systems)

10.3.1 Where trials are conducted, or service usage accumulated, on items which are repaired or replaced on failure, the MTBF of the item is assessed, as for MTTF, by dividing the accumulated active or operational time by the total number of failures. Any time when the item is not active (e.g. due to repair, servicing, etc.) should not be included in the MTBF calculation.

10.3.2 Thus if t_i is the accumulated active time of the ith item, and there were n such items having experienced between them a total of r failures, then:

$$\text{Estimated MTBF } = \frac{1}{r} . \sum_{i=1}^{n} t_i$$

10.3.3 The above estimate has most relevance when MTBF is not changing with time. (This does not necessarily mean times to failure should be negative exponentially distributed.) When the MTBF is changing, other estimation methods may be preferable, e.g. CUSUMs or moving average plots, or plotting on the basis of some growth model such as Duane.

10.3.4 When MTBF is being estimated from special trials, the trials may be completed either when a certain number of test hours are accumulated or when a certain number of failures are accumulated. The above estimation is used in either case.

10.4 CI on MTBF Estimate

(The following is only valid when failures are occurring negative exponentially in time.)

10.4.1 Failure Truncated Trials

It is assumed that in this case the trials cease when the rth failure occurs, and that the accumulated active time is then T hours. Thus:

the point estimate of the MTBF ($\hat{\theta}$) = T/r

The two-sided confidence limits for the 100(1-α)% Confidence Interval are given by:

$$\frac{2T}{\chi^2_{\frac{\alpha}{2},2r}} \leq \hat{\theta} \leq \frac{2T}{\chi^2_{\left(1-\frac{\alpha}{2}\right),2r}}$$

where $\chi^2_{x,y}$ is the value of Chi-square with y degrees of freedom, exceeded with probability x. Values of Chi-square are tabulated in PtDCh3. (These limits may be inverted to obtain the 100(1-α)% CI on failure rate.)

The one-sided confidence limit for 100(1-α)% is given by:

$$\hat{\theta} \geq \frac{2T}{\chi^2_{(1-\alpha),2r}}$$

10.4.2 Time Truncated Tests

It is assumed that in this case trials are terminated when the accumulated equipment running time reaches T. Using the same notation as in para. 10.4.1, the 100(1-α)% CI on MTBF for a time truncated test is:

$$\frac{2T}{\chi^2_{\frac{\alpha}{2},(2r+2)}} \leq \hat{\theta} \leq \frac{2T}{\chi^2_{\left(1-\frac{\alpha}{2}\right),2r}}$$

Calculations proceed in the same way as in para. 10.4.1 except that the lower confidence limit will be slightly different, since χ^2 is computed using 2r+2 degrees of freedom, not 2r as previously.

Note that in this case the lower confidence limit can be computed even if no failures occur.

However, in the case of zero failures the *one-sided* 100(1-α)% lower confidence limit for the MTBF can be calculated exactly, without the use of χ^2 tables, as:

MTBF lower confidence limit = $\dfrac{T}{-\log_e(\alpha)}$ (r = 0)

Conversely, the upper confidence limit for the failure rate is the inverse of this, i.e.:

Failure Rate upper confidence limit = $\dfrac{-\log_e(\alpha)}{T}$ (r = 0).

In general, the one-sided confidence limit for 100(1-α)% is given by:

$$\hat{\theta} \geq \frac{2T}{\chi^2_{(1-\alpha),(2r+2)}}$$

Intentional blank page

LEAFLET 7/0

REFERENCES

1. B Effron and C Morris: *Stein's Paradox in Statistics*. Scientific American, May 1977 Vol 236 No. 5

2. I Basovski: R*eliability Theory and Practice*. Prentice-Hall International

Intentional blank page