View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by JRC Publications Rep

Institute for Energy

A BAYESIAN FRAMEWORK FOR THE QUANTITATIVE MODELLING OF THE ENIQ METHODOLOGY FOR QUALIFICATION OF NON-DESTRUCTIVE TESTING

Authors: Luca Gandossi, Kaisa Simola

DG JRC
Institute for Energy
2007

JRC
EUROPEAN COMMISSION
EUR 22675 EN

Mission of the Institute for Energy
The Institute for Energy provides scientific and technical support for the conception, development, implementation and monitoring of Community policies related to energy. Special emphasis is given to the security of energy supply and to sustainable and safe energy production.

European Commission
Directorate-General Joint Research Centre (DG JRC)
http://www.jrc.ec.europa.eu/

Institute for Energy, Petten (the Netherlands)
http://ie.jrc.ec.europa.eu/

Contact details:
Luca Gandossi
Tel: +31 (0)224 56 5250
E-mail: luca.gandossi@jrc.nl

Legal Notice

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

The use of trademarks in this publication does not constitute an endorsement by the European Commission.

The views expressed in this publication are the sole responsibility of the author(s) and do not necessarily reflect the views of the European Commission.

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server http://europa.eu/

EUR 22675 EN
ISSN 1018-5593
Luxembourg: Office for Official Publications of the European Communities

© European Communities, 2007
Reproduction is authorised provided the source is acknowledged.
Printed in the Netherlands

A BAYESIAN FRAMEWORK FOR THE QUANTITATIVE MODELLING OF THE ENIQ METHODOLOGY FOR QUALIFICATION OF NON-DESTRUCTIVE TESTING

Luca Gandossi & Kaisa Simola

May 2007

FOREWORD

The output from the European inspection qualification process is generally a statement concluding whether or not there is high confidence that the required inspection capability will be achieved in practice, for the specified inspection system, component and defect range. However, this process does not provide a quantitative measure of inspection capability of the type that could be used for instance in the connection of the risk-informed in-service inspection (RI-ISI) process. In a quantitative RI-ISI, a quantitative measure of inspection effectiveness is needed in determining the risk reduction associated with inspection.

The issue of linking the European qualification process and a quantitative measure of inspection capability has been discusses within the ENIQ (European Network for Inspection and Qualification) over several years. In 2005 the ENIQ Task Group on Risk decided to initiate an activity to address this question. A program of work was proposed to investigate and demonstrate an approach to providing some objective measure of the confidence which comes from inspection qualification, and allowing risk reduction associated with a qualified inspection to be calculated. The work plan focuses on following issues:

- Investigating sensitivity of risk reduction to POD level and detail;
- Investigating the use of user-defined POD curve as target for qualification;
- Testing a Bayesian approach to quantifying output from qualification;
- Linking qualification outcome, risk reduction and inspection interval;
- Pilot study of overall process, including a pilot qualification board.

The work is organised in a project “Link Between Risk-Informed In-Service Inspection and Inspection Qualification”, coordinated by Doosan Babcock, UK. The project is partly funded by a group of nuclear utilities. In addition, the Joint Research Centre, Institute of Energy is participating in the project with a significant work contribution.

This report contributes to the project by addressing the Bayesian approach to quantify output from qualification. The research work has been carried out at the Joint Research Centre, Institute for Energy during year 2006.

Luca Gandossi
Scientific Officer
DG Joint Research Centre (JRC-IE)

Kaisa Simola
Senior Research Scientist
VTT Technical Research Centre of Finland
(Visiting Scientist at JRC-IE 2004-2006)

FOREWORD **3**

1 INTRODUCTION **5**

2 LINK BETWEEN THE ENIQ QUALIFICATION METHODOLOGY AND PROBABILITY OF DETECTION **5**

3 PRINCIPLES OF PROBABILITY OF DETECTION ESTIMATION **7**

4 REVIEW OF THE BAYESIAN MODEL PROPOSED FOR THE QUANTIFICATION OF THE TJ **9**

4.1 DETERMINATION OF THE POD FROM A QUANTIFIED TJ AND TRIAL RESULTS 10
4.2 QUANTIFICATION OF THE TECHNICAL JUSTIFICATION 11
4.3 GENERAL PRINCIPLES FOR COMBINING THE TECHNICAL JUSTIFICATION AND TEST PIECE TRIAL RESULTS 13
4.3.1 Approach 1 ..... *13*
4.3.2 Approach 2 ..... *14*
4.3.3 Approach 3 ..... *14*
4.4 DETECTION TARGETS: INPUT OR OUTPUT OF THE QUALIFICATION PROCESS? 15
4.5 LINK BETWEEN DETECTION TARGET AND SAMPLE SIZE 17

5 EXAMPLES **20**

5.1 EXAMPLES OF APPROACH 1 21
5.1.1 Example A1 ..... *21*
5.1.2 Example A2 ..... *26*
5.1.3 Example A3 ..... *29*
5.2 EXAMPLES OF APPROACH 2 32
5.2.1 Example B1 ..... *33*
5.2.2 Example B2 ..... *36*
5.3 EXAMPLE OF APPROACH 3 39
5.3.1 Example C1 ..... *39*

6 DISCUSSION OF SOME IMPORTANT ISSUES **41**

6.1 INDEPENDENCE OF TJ AND PRACTICAL TRIALS 41
6.2 DEFINITION OF DEFECT POPULATION 41
6.3 REPRESENTATIVENESS OF TEST BLOCK DEFECTS 42
6.4 CHOICE OF THE PRIOR DISTRIBUTION 44

7 CONCLUSIONS **47**

8 ACKNOWLEDGEMENTS **47**

9 REFERENCES **47**

APPENDIX 1: ESTIMATION OF A POPULATION PARAMETER IN CLASSICAL AND BAYESIAN STATISTICS. **48**

CLASSICAL STATISTICS 48
BAYESIAN STATISTICS 49
REFERENCES 54

1 Introduction

The European methodology for qualification of non-destructive testing, produced by the European Network for Inspection and Qualification (ENIQ), has been adopted as the basis of inspection qualifications for nuclear utilities in many European countries [1]. According to this methodology, the inspection qualification is based on a combination of technical justification (TJ) and practical trials. The methodology is qualitative in nature, and it does not give explicit guidance on how the evidence from the technical justification and results from trials should be weighted.

Recently, we have proposed a quantified approach to combine evidence from technical justifications and practical trials [2]. A Bayesian statistical framework for the quantification process was introduced, and some examples of possibilities to combine technical justification and trial results were given. The underlying idea was to improve transparency in the qualification process, whilst producing at the same time estimates of probability of detection that could for instance be used in structural reliability evaluation and Risk-Informed In-Service Inspection.

In the present work, we attempt to give a more detailed description of the approach and some guidelines regarding how a user (utility, qualification body, etc.) could tackle the problem of quantifying the outcome of a qualification exercise in practical terms.

This report is structured in the following way:

1) We first discuss the link between inspection qualification (ENIQ approach) and a quantitative measure of inspection capability such as the probability of detection (Chapter 2);
2) We review the simple principles and prescriptions that an experimenter would follow when attempting to determine the probability of detection of an NDE system in a rigorous statistical setting (Chapter 3);
3) We then present the main concepts of the Bayesian framework for quantification of the ENIQ qualification methodology (Chapter 4);
4) We give several examples showing how the proposed approach could be put in practice (Chapter 5);
5) Finally, we discuss some important issues, such as the definition of defect population and the representativeness of test block defects (Chapter 6).

2 Link between the ENIQ qualification methodology and probability of detection

The ultimate aim of inspection qualification is to provide assurance that the inspection objectives can be met when applied in practice. The intuitive way to do so would simply involve demonstrating that the inspection system can indeed find the required percentage of defects under controlled experimental conditions closely simulating the reality in which the system in meant to be applied. Practically, this imply the necessity of procuring and inspecting a large number of components similar (in terms of geometry, materials, etc.) to those meant for inspection, containing defects representative (in terms of morphology, size, etc.) of those damage mechanisms the inspection is targeted to, and under conditions similar (in terms of environment, inspectors, etc.) to those that will be found in reality.

However, this is often impossible to achieve even if vast financial resources were available (which is often not the case). For instance, for certain combinations of materials and defect morphologies, it is impossible to create artificial defects which are truly representative of real in-service flaws.

These very considerations led the ENIQ network to state that [3]: "*[…] ENIQ argues that it is normally not possible or practicable to construct a convincing argument for inspection capability using results from test piece trials alone*".

As we will see in more detail in the following, the number of trials to perform in order to obtain a statistically valid result is punishingly high. Again, ENIQ recognised this explicitly: "*For example, if 95% probability of detection of a particular defect type were required, with 95% confidence, this would require the manufacture of test specimens containing 59 examples of this defect type, and the detection of all 59 in the test piece trial. This process would have to be repeated for each defect type of concern.*"

Thus, ENIQ introduced an approach, the European Qualification methodology, based on the sum of two items [1]: practical tests and a Technical Justification. Practical tests are experiments conducted on simplified or representative test pieces resembling the component to be inspected. The Technical Justification (TJ) is a document which assembles together all available evidence on the effectiveness of the test, including previous experience, experimental studies, mathematical modelling, physical reasoning and so forth.

The ENIQ approach to inspection qualification, thus, is inescapably qualitative in nature. Having recognised the difficulty of performing sufficient real practical tests, the ENIQ approach ultimately relies on engineering (expert) judgement to decide on whether an inspection system can indeed be said to have "passed" qualification. These considerations are fully justified, and the ENIQ methodology has rightly become the established philosophical approach to inspection qualification for utilities and regulators alike in many European countries.

Despite its qualitative nature, the ENIQ methodology explicitly mentions detection objectives by requiring the specification of flaw population and detection capabilities. Concerning the latter, Ref. [1] places the following among the essential part of the input information: "*Detection and false calls: The detection rate which the relevant involved parties regard as necessary for the actual test. (This may arise from a regulatory requirement). Qualification will aim to assess whether this detection rate is attainable for the test method chosen. […]*"

Thus the ENIQ methodology requires that some inspection capability in terms of detection rate is specified as an <u>input parameter</u> and that such capability is demonstrated for the NDT system to be qualified. The inspection capability must be demonstrated for specified defect types (flaw populations in the ENIQ terminology, [1]), defined as "*the flaws or conditions which must be detected by the actual NDT in the real components. This information is likely to include size, position, type, orientation, etc.*"

If the detection target is expressed as a detection rate, which is calculated as the proportion of successes over a very limited number of trials, a question is often raised about the confidence with which such a value represents the probability of detection. On the other hand, as the Technical Justification is seen as an essential part of the qualification process adding confidence to the capability assessment, this confidence increase through the TJ should be somehow credited in practice.

In our framework [2], we have proposed a quantitative approach to account for the TJ based on Bayesian statistics. Bayesian statistics offer a formal way of treating expert judgement and combining it with experimental evidence. Our purpose is to produce a quantitative estimate of the probability of detection based on the combined information of the TJ and practical trials.

We have proposed to consider the TJ (or rather, the information "carried" by it) as a set of "equivalent" practical trials (experiments). Thus, assembling and reporting in the TJ some relevant evidence in favour of inspection capability (for instance, results from numerical modelling, or previous experience) is seen as actually performing a number of equivalent trials. This, in turn, reduces the number of real trials necessary to prove the given detection target, when this target is expressed as a probability of detection with a certain confidence.

Such a quantification of the qualification methodology provides a link between the inspection qualification and estimates of probability of detection. The proposed approach is described in Chapter 4 and illustrative examples are given in Chapter 5. Before that we discuss in Chapter 3 the principles of probability of detection estimation.

3 Principles of probability of detection estimation

In this report, when referring to an NDE system, we often use the expression “detection capability”. Unless otherwise explicitly stated, with this statement we refer to the probability of detection, i.e. the probability that the NDE system under consideration will detect a flaw when applied to a defective component. In practice, detection capability is fully described only when considering a second, important feature: the ability of the system to discriminate false calls. The probability of false calls is the probability that the NDE system under consideration will report the presence of a flaw when applied to a non-defective component. An NDE system could have a very high probability of detection but have at the same time a very high probability of false calls, and thus be rather useless.

Our fundamental assumption is that we attempt to determine the probability of detection as a population proportion, by means of a set of independent Bernoulli trials. That is, we assume that the process of measuring the detection capability of a given NDE system consists in its repeated application on a set of representative defective components. A statement on the detection capability is eventually derived from the ratio of successes over the total number of trials.

In the theory of probability and statistics, a Bernoulli trial is an experiment whose outcome is random and can be either of two possible outcomes, called "success" and "failure." In practice it refers to a single experiment which can have one of two possible outcomes. Mathematically, such a trial is modelled by a random variable which can take only two values, 0 and 1.

In Appendix 1 we present a more detailed review of the issues related to the experimental determination of a population proportion in both classical and Bayesian statistics. To avoid complications or misunderstandings, we have there considered the general problem of determining a population proportion. This could be the fraction of red marbles in an urn containing blue and red marbles, the proportion of people suffering a particular disease, or the proportion of voters in a country that vote for party X. Therefore, in the Appendix we have purposely avoided referring to p as a probability of detection. This is in order to avoid a possible source of confusion between the probability of detection itself (the proportion, p, we want to determine experimentally), and the confidence level (also a probability) associated with the interval estimators derived there. We want to stress again the fact that often ENIQ practitioners speak of “high confidence” that a given (qualified) NDE system will be able to reliably identify the required defects. Such a statement of “high confidence” is perfectly understandable in terms of spoken language, but it is in truth a statement of “high probability of detection”.

Summarising, Appendix 1 deals with a commonly encountered setting, when the experimenter cannot measure the proportion, p, directly (for instance because that would involve interviewing too large a number of people, or because each “experiment” would be too costly, etc.). Thus, the experimenter can only carry out a limited number of experiments, and try to draw some conclusions from the outcome of this set of tests. A very common (and often justified) assumption is that the experiments are independent, that is the outcome of one does not affect (and is not affected by) the outcome of the others. If p is the unknown proportion that the experimenter is trying to determine, each experiment in this framework is a Bernoulli trial. The probability of a success (for instance, drawing a red marble) is then p, and the probability of a failure (drawing a blue marble) is $q=1-p$. Clearly, as p is a proportion, it is defined on the interval [0,1]

The principal point of Appendix 1 is to show that, in a rigorous statistical setting, an optimal way to report the outcome of an experiment is by means of an interval estimator. Alternatively, an

interval estimator of required width can be specified as input for the experiment, and the sample size adjusted accordingly. Typically, we are interested in one-sided lower bound interval estimators, because it is naturally most important to set a lower bound for the probability of detection.

We will then often deal with intervals of the form

$$[p_{100\delta\%}, 1] \tag{1}$$

meaning that we are 100xδ% certain that p is in the interval $[p_{100\delta\%}, 1]$ (or, equivalently, that we are 100xδ% certain that p is greater than $p_{\delta\%}$)[1].

As stated above, we attempt to determine the probability of detection as a population proportion, by means of a set of independent Bernoulli trials. Other models could be employed, for instance the well known method proposed by Berens [4]. In the latter, the probability of detection is assumed (a priori) to be some parametric function of crack size. The parameters are then determined with statistical methods (maximizing the likelihood function, for instance) to best fit the available experimental data. Such a model has proved successful because it allows the determination of a POD curve for the whole range of crack sizes ranging between 0 and the wall thickness with a rather limited number of trials, whereas (strictly speaking) the Bernoulli setting (determining a population proportion) would require several tests just for a single crack size.

We here assume that the probability of detection, p, associated with defects belonging to a population as defined above does exist. A value p intrinsic to the NDT system under consideration (seen as the combination of all the above-mentioned variables) must exist if p is seen in the frequentistic interpretation of probability. In this sense, p is seen as the percentage of detection of that given defect type, i.e. the number of detected defects divided by the total number of trials, as the total number of trials approaches infinity.

In Appendix 1, we have described how, in the Bayesian statistical framework, the result (namely, the number of successes) of a set of Bernoulli trials can be treated as a sample from a binomial distribution with parameter p. In turn, the parameter p is modelled as a random variable and the uncertainty related to it is expressed with a probability distribution. We have seen how the natural conjugate distribution of the Binomial distribution is the Beta distribution, and thus how it is a very convenient choice to assume that:

$$p \sim Beta\ (\alpha, \beta) \tag{2}$$

At any given time, the individual experimenter is entitled to choose for the parameters (α, β) the values that he or she feels most appropriate to describe the current knowledge about p. If some new knowledge becomes available, for instance because a new set of experiments is carried out, the experimenter moves (updates) the current (prior) set of parameters (α, β) to a new (posterior) set.

Let us suppose that the current knowledge regarding p is summarised by the set of values (α_{prior}, β_{prior}). We do not discuss now how this state of knowledge was reached. Let us then assume that a set of N trials is carried out (i.e. the NDE system at hand is applied to a set of N defective components) and let us supposed that a number, N_s, of such flaws are detected (successes). The number of failures is then $N_f=N-N_s$ In Appendix 1 we have seen that the new set of parameters of the Beta distribution that describe the knowledge regarding p can very easily be obtained as:

[1] Formally, this language is slightly loose. We will indeed often say that p is in the interval [$p_{\square\%}$, *1*], as this will be sufficient for the purposes of this work, but we remark here that strictly speaking in classical statistics the statement made is that the interval covers the parameter, and not that the parameter is inside the interval.

$$\alpha_{post} = \alpha_{prior} + N_s$$
$$\beta_{post} = \beta_{prior} + N_f = \beta_{prior} + N - N_s \quad (3)$$

Thus the posterior distribution for p becomes

$$p \sim Beta(\alpha + N_s, \beta + N - N_s) \quad (4)$$

In other words, the new parameter α of the Beta posterior is equal to the old parameter α increased by the number of successes, N_s, whereas the new parameter β of the Beta posterior is equal to the old parameter β increased by the number of failures to detect a flaw, (N-N_s).

If no knowledge is currently available, for instance because no experiment has been carried out, the experimenter should choose α and β so that a so-called non-informative distribution is obtained. A reasonable choice could then be α=1 and β=1, so that the prior distribution becomes the uniform distribution. In such a case, the experimenter is assuming that the probability of detection is equally likely to be anywhere in the interval [0,1].

An important property of the updating process is that if new evidence becomes available, it can be readily used to obtain a new posterior. If, for instance, a second set of N_2 trials is carried out, Eq. (3) can be applied again to obtain new parameters α and β. Notably, the order in which the first and subsequent sets of trials are carried out does not affect the outcome.

From the knowledge of the posterior, confidence intervals can be easily derived[2]. For instance, for any desired confidence level δ (say .95) and considering one-sided lower bound intervals, we can find that value $p_{\delta\%}$ ($p_{95\%}$) so that the area of the distribution within the interval [$p_{\delta\%}$, 1] is exactly δ.

It may be a trivial point, but it is worth stressing the fact that confidence levels should be suitably high (for instance, .95 or .99). It clearly does not make much sense reporting a very high probability of detection (say 99%) with a confidence level of say 40%. There would be a very good chance (60%) that the true value of the parameter p is actually outside the reported interval (and therefore somewhere – anywhere – between 0% and 99%).

4 Review of the Bayesian model proposed for the quantification of the TJ

An inherent assumption in the ENIQ methodology is that the assessment of the inspection capability can be partly based on the TJ. This is fully justifiable, since the information collected for the TJ does provide evidence related to the inspection capability. One reason (at least so far) for not using this information to support a quantitative evaluation of the flaw detection probability arises from a frequentistic interpretation of probability. In the Bayesian – or subjective – interpretation of probability, probabilities are interpreted as a measure of our degree of belief. The Bayesian framework allows the quantification of expert judgements in terms of probabilities, and thus enables the utilization of the TJ in quantitative POD assessment.

In our approach, we have introduced a way to combine a judgement of the inspection capability based on the TJ with the results of practical trials. Figure 1 illustrates the step-wise principle of combining the evidence. In the previous chapter we described a convenient way to model the results of practical trials in a Bayesian framework. For convenience, it is natural to use a

[2] We will use the more understandable “confidence interval” when dealing with interval estimators, even if in this Bayesian approach it would rather be technically correct to use “credible set” (see Appendix 1 for a discussion on this difference.

mathematically similar structure to model the evidence obtained from the TJ. Thus the essential idea we propose consists in interpreting the TJ in terms of an equivalent set of practical trials [2]. We suggest that the TJ be quantified using two numbers: an equivalent total number of trials, N_{TJ}, and an equivalent number of successes, $N_{TJ,s}$. These numbers, provided by experts in a documented and transparent manner, are then used in combination with the number of practical trials, N_{trials}, (and associated number of successes, $N_{trials,s}$) to hopefully prove the achievement of the qualification objectives.

In the following we will first summarise the mathematical part of the approach (4.1) and then discuss how the TJ could be quantified in practice (4.2). In section 4.3 we propose alternative ways to combine the TJ and test piece trial results.

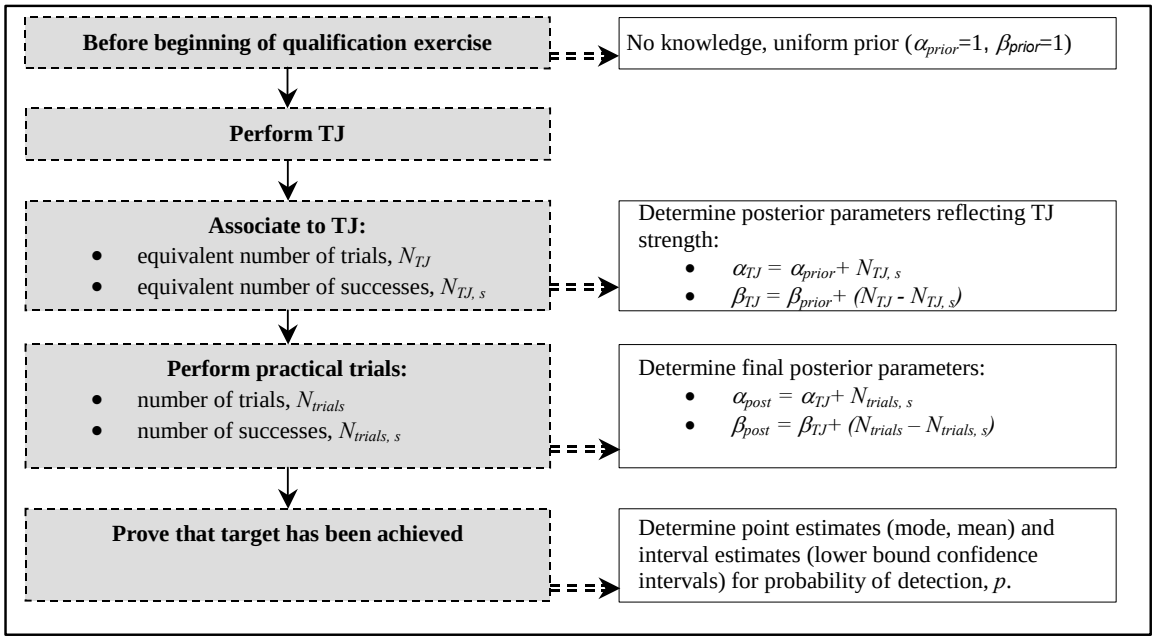

Figure 1 Principle for combining evidence from TJ and practical trials to prove that reliability target is achieved.

4.1 Determination of the POD from a quantified TJ and trial results

In a Bayesian framework we start from a prior distribution, which should reflect our knowledge. In our approach we propose to use a non-informative prior as the starting point. One could argue that some prior information on the inspection capability is available even before the TJ is produced. However, it would be practically impossible to distinguish that prior knowledge from evidence collected during the TJ. So we assume that all qualitative evidence on the inspection capability prior to the practical trials is assembled in the TJ. This is why we propose that the starting point is one of no knowledge (Figure 1), whereas an expert practitioner would be able to make some informed statement even before beginning to work on the TJ.

As described in Chapter 3, we use a Beta distribution to express our uncertainty related to the parameter p, which in practice is the measure of the POD. For mathematical convenience we also use a Beta distribution (as discussed above and in Appendix 1) for the prior distribution. The Beta distribution with parameter values α=1 and β=1 reduces to the uniform distribution, for which the probability of detection is equally likely to be anywhere in the interval [0,1].

The non-informative prior distribution is first updated with the equivalent numbers of trials and successes, N_{TJ} and $N_{TJ,s}$ obtained from the TJ quantification. Finally the numbers of trials and successes from the practical test piece trials are combined to obtain the posterior parameters.

The posterior distribution for p, expressed in Eq. (3), would thus be defined by the following parameters:

$$\alpha_{post} = 1 + N_{TJ,s} + N_{trials,s}$$
$$\beta_{post} = 1 + (N_{TJ} - N_{TJ,s}) + (N_{trials} - N_{trials,s}) \quad (5)$$

The problem is now shifted to determine the equivalent technical justification sample size, N_{TJ}, and an equivalent number of successes, $N_{TJ,s}$. Determining (or better, deciding) the value of N_{TJ} is ultimately related to deciding how much the TJ is "worth" when compared to the practical trials. Determining $N_{TJ,s}$ is connected to giving the TJ a score that measures how close to perfect the TJ really is.

4.2 Quantification of the technical justification

In [2] the decision of how to partition the original sample size between TJ and real experiments was left to the user. As stated, such a decision is a matter of expert judgement, and the goal of this work is ultimately to give some guidelines on how to proceed.

The quantification of the technical justification consists of the following logical steps:

1. Identification of the information sources relevant for the judgement;
2. Definition of suitable scoring: how to value and weight the qualitative pieces of evidence and what kind of scale should be used;
3. Quantification of the evidence contained in the TJ.

All relevant information sources for an inspection qualification are (or at least – for a properly compiled TJ – should be) defined and discussed within the technical justification, and the ENIQ documents provide guidance on this aspect. For example, ENIQ RP2 [5] sets up the recommended contents for a technical justification. Thus, a TJ should comprehensively cover all those factors affecting the capability of the inspection. For the quantification of the inspection capability based on the TJ, it is necessary to extract from the TJ the evidence related to the inspection capability.

In the ENIQ RP 3, the evidence on the effectiveness of the test is broken down into four main elements:

- Theoretical modelling
- Experimental evidence
- Parametric studies
- Equipment and data analysis

This breakdown could be used as a starting point for evaluating the various pieces of evidence.

The next step is to define weighting and scoring principles for the quantification. In general, **Weighting** is associated with the determination of N_{TJ}, i.e. the relative weight that the technical justification has when compared with the practical trials. **Scoring** is associated with the determination of $N_{TJ,s}$, i.e. the judgement of how good the information contained in the TJ really is.

Weighting is needed in two phases of the quantification process: 1) one should determine what is the weight of the TJ in comparison with practical trials, and 2) one should determine the relative importance of various pieces of evidence within the TJ.

The ENIQ Recommended Practice 3 (Strategy document for technical justification, [3]) gives some qualitative guidance on the relative weight that the TJ would probably have in comparison with practical trials for different types of qualifications, see Table 1. This guidance could also be used as a starting point for the weighting in the quantification process.

When determining the relative importance of various pieces of evidence within the TJ, one starting point could be the Table 2 below, also taken from the ENIQ recommended practice 3. This table gives guidance on types of evidence to be included for different types of technical justification. It is worth noting, that the relative weights of these elements would not necessarily be commensurate with the likelihood of having such evidence included in the TJ. However, if some evidence is unlikely to be included, it could be expected that it would not have much weight.

Table 1 Relative weight of the technical justification with respect to test piece trials [3]

Type of TJ	Overall weight of TJ
Justify inspection procedure	Varies
Justify use of test pieces and defect populations	Small
Justify inspection equipment	Varies
Extend qualification to different geometry	Large
Extend qualification to different material structure	Varies
Qualify upgraded equipment or software	Large
Qualify upgraded procedure	Large
Qualify for changed defect descriptions	Large

Table 2 Guidance on types of evidence to be included for different types of technical justification

Type of TJ	Theoretical modelling	Experimental evidence	Parametric studies	Equipment and data analysis
Justify inspection procedure	VL	VL	L	L
Justify use of test pieces and defect population	L	VL	L	U
Justify inspection equipment	U	U	L	VL
Extend qualification to different geometry	VL	VL	L	U
Extend qualification to different material structure	L	VL	L	U
Qualify upgraded equipment or software	U	U	U	VL
Qualify upgraded procedure	VL	L	U	L
Qualify for changed defect descriptions	VL	L	VL	U

Key to Table:
VL=TJ is very likely to contain evidence of this type; **L**=TJ is likely to contain evidence of this type (depending on the specific case); **U**=TJ is unlikely to contain evidence of this type.

4.3 General principles for combining the technical justification and test piece trial results

In [2], we proposed three different ways in which the problem of quantifying a qualification exercise could be addressed (at least in principle).

These were:

Approach 1: Quantifying the technical justification in terms of score and relevance;

Approach 2: Technical Justification representing a number of successful trials;

Approach 3: Use of trial results to achieve target expected value of POD, and use of TJ for increasing confidence.

4.3.1 Approach 1: quantifying the technical justification in terms of score and relevance

Approach 1 represents the most natural way to proceed. Based on common sense, we postulated that the following basic principles should apply:

- If some evidence is missing, this should imply less weight for the TJ posterior. This means that the equivalent TJ sample size, N_{TJ}, should be smaller than in the case of stronger evidence.
- If evidence is present showing that some defects could be missed, this should imply a lower expected value of the TJ posterior, i.e. the ratio of $N_{TJ,s}$ over N_{TJ} should be smaller than in the case where the evidence is more convincing regarding detection capability.

We assume that the TJ can be broken down into a number of elements and that the impact of each element towards demonstrating inspection capability is independent from the other. As seen above, these elements could be for instance: theoretical modelling, experimental evidence, parametric studies and equipment and data analysis.

We should then ask the following questions for each of such elements:

- How convincing is this piece of evidence in support of the inspection capability?
- What relative weight has this evidence in the overall justification?

To tackle the first question, the task is to score each element with a number between (0,1) which expresses the degree of “goodness” of each element. The closer the value is to 1, the better the evidence contained in the TJ element supports the detectability of defects. To tackle the second question, the elements can be weighted to reflect the importance of the element in the overall justification.

In general, we suggest that the following steps can be taken:

1. Decision of the TJ equivalent sample size, N_{TJ};
2. Decision regarding the relative weights of the TJ elements;
3. Decision regarding the scores of the individual TJ elements;
4. Calculation of TJ total weighted score;
5. Calculation of TJ posterior parameters.

This approach is best explained with some examples, see Chapter 5. In practice it may be easier to judge the weight of the TJ compared to practical trials after the relative weights and scores for the TJ elements have been determined, thus the above step 1 would be done only

after steps 2 and 3. Nonetheless, in the examples presented in Chapter 5 we follow the step order suggested here.

An important issue associated with this approach, also identified in [2], is the fact that unless the total TJ score is very high, the TJ actually results in carrying a negative value, in the sense that despite having spent resources to produce supporting evidence, we are actually forced to conclude that additional trials would be necessary to guarantee the same capability target.

This fact will be better elucidated using the examples. Here it suffices to say that if we analyse how the TJ score is converted into the two parameters α_{TJ} and β_{TJ} of the TJ posterior distribution, it is straightforward to see how any number less than unity is bound to yield a β_{TJ} greater than zero, which in turn can be interpreted as a number of "equivalent failures" carried by the TJ. This problem is not a deficiency of the model itself but rather of the interpretation or definition of a "high confidence". If one cannot be sufficiently convinced of the evidence provided by the elements of the TJ, it is unreasonable to set very high requirements for the POD to be achieved. On the other hand, it is very difficult to quantify the TJ score if it is quite high anyway.

4.3.2 Approach 2: Technical Justification representing a number of successful trials

The above considerations lead us to conclude that another possible way forward would be to quantify the technical justification in terms of **equivalent successes only**. Approaches 2 and 3 are thus based on this idea. Indeed, unless the TJ has highlighted some intrinsic limitations in the NDE system under qualification (which would necessarily force us either to give up qualification or to undertake major changes), it would be difficult to argue that the TJ does not carry some (possibly small) positive value. This would even be the case for a very basic TJ, assembled investing only few resources.

Under the assumption of a TJ carrying only equivalent successes, one only needs to choose the weight carried by the TJ, so as to be proportional to the "strength" of the information and supporting evidence contained in it.

Under Approach 2, the TJ is straightforwardly modelled in terms of equivalent successes only. Thus:

$$\begin{aligned} N_{TJ} &= N_{TJ,s} \\ N_{TJ,f} &= 0 \end{aligned} \tag{6}$$

The problem is reduced to decide how to partition the sample size between TJ and practical trials. This is of course not a trivial matter, and it is ultimately a matter of expert judgement. We show two examples in Chapter 5.

4.3.3 Approach 3: use of trial results to achieve target expected value of POD, and use of TJ for increasing confidence

Approach 3 is based on the same basic idea as before, i.e. that the technical justification is modelled in terms of equivalent successes only. As the main purpose of the TJ, even in broadly accepted qualitative terms, is to provide "confidence" in the inspection, we can translate this idea in a more soundly defined statistical framework, assuming that a certain expected detection probability has to be achieved with practical trials (real experiments), and the TJ is used to reach a predetermined lower confidence bound.

Thus, we introduce the following two steps [2]:

- Fix a required expected value for p, and prove it using a reduced number of practical trials;
- Fix a required lower bound for p (e.g. 95 % or 99 %), and prove it with the aid of the TJ.

The idea is based on the fact that to obtain a posterior, the order in which the evidence become available is not relevant. Thus, even if this is not done in reality, it can be thought that the practical trials are carried out first, and the TJ is assembled later.

If we use a uniform prior ($\alpha_{prior} = \beta_{prior} = 1$), and form a first posterior after the evidence $N_{trials,s}$ in N_{trials} is gathered, the expected value of such distribution is

$$E(p) = \frac{1 + N_{trials,s}}{(1 + N_{trials,s}) + (1 + N_{trials} - N_{trials,s})} = \\ = \frac{1 + N_{trials,s}}{2 + N_{trials}} \tag{7}$$

Let us assume that we find all the flaws in the trials, then $N_{trials,s} = N_{trials}$ and thus

$$E(p) = \frac{1 + N_{trials}}{2 + N_{trials}} \tag{8}$$

From Equation 8, we can then derive an expression to obtain the required number of trials needed to obtain a given expected value, say μ.

$$N_{trials} = \frac{2\mu - 1}{1 - \mu} \tag{9}$$

In other words, if we now inspect and detect all the flaws in N_{trials} practical trials, where N_{trials} is obtained from equation (9), we can prove an expected value μ.

Equation 9 has been plotted in Figure 2, and some values are reported in Table 3 for clarity It can be seen, for example, that only 8 out of 8 flaws need to be found to guarantee an expected value E(p) = 0.9, and 18 out of 18 flaws to guarantee an expected value E(p) = 0.95.

The Technical Justification is then converted into an equivalent number of successes, which is then used to prove a second, distinct target: some given lower bound value of the probability of detection. This approach is best explained with an example, see section 5.3.

4.4 Detection targets: input or output of the qualification process?

In [2], we proposed a logical process that began with fixing a detection target. As introduced above, we proposed that such a target should be expressed rigorously, by means of a confidence interval and associated confidence level.

As discussed, a particularly useful detection target would be a one-sided interval, one in the form expressed by equation (1). For instance, we could have specified our target to be the following: prove with 95% confidence that the probability of detection is at least 90% (thus, δ=0.95 and $p_{95\%}$=0.90.) Fixing such a detection target automatically dictates the required sample size needed in order to prove it, as we will discuss shortly. This represents a situation in which a detection target is fixed as the input of the qualification exercise, and the qualification itself is seen as the process that proves (or disproves) the feasibility of achieving such a target in practice.

Table 3 Required number of trials to guarantee a given expect value, μ.

Expected value, μ	Required number of trials, N_{trials}	Expected value, μ	Required number of trials, N_{trials}
0.8	3	0.935	13
0.81	3	0.94	14
0.82	3	0.945	16
0.83	3	0.95	18
0.84	4	0.955	20
0.85	4	0.96	23
0.86	5	0.965	26
0.87	5	0.97	31
0.88	6	0.975	38
0.89	7	0.98	48
0.9	8	0.985	64
0.91	9	0.99	98
0.92	10	0.995	198
0.93	12	0.999	998

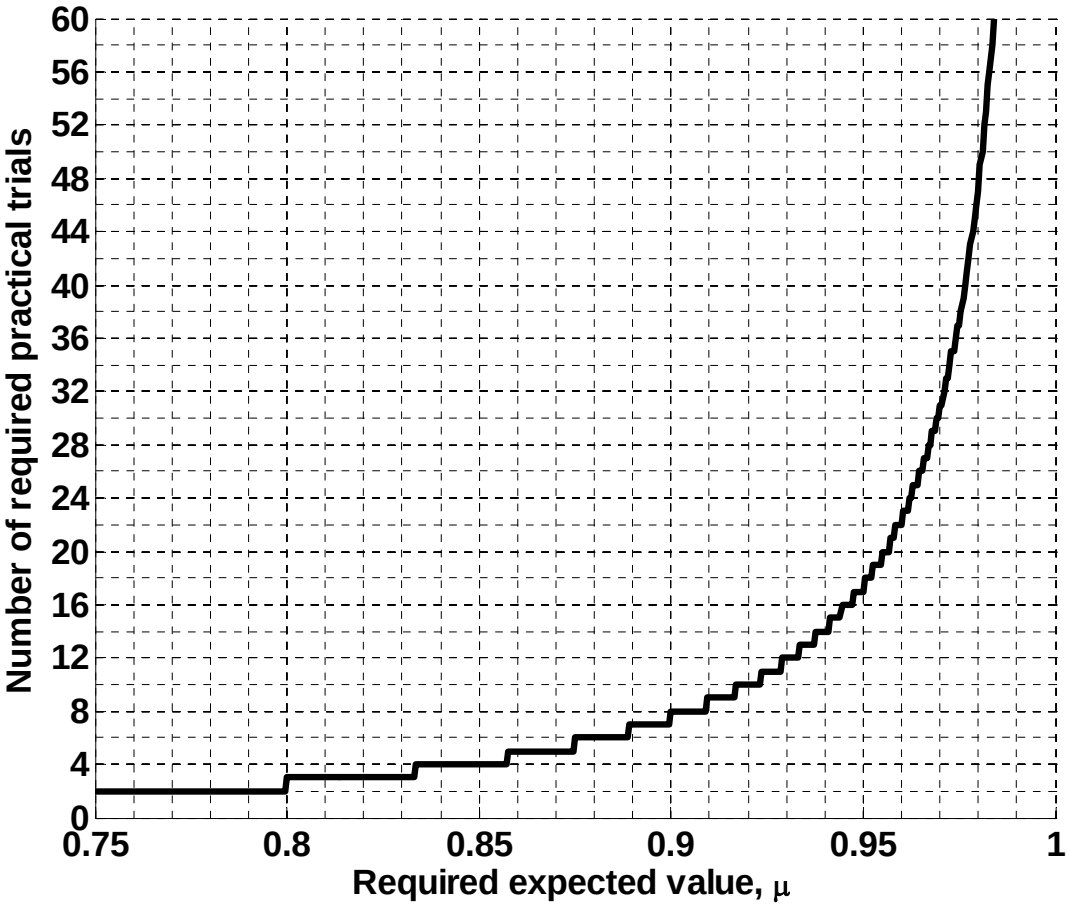

Figure 2 Number of practical trials required to guarantee a given expected value, μ, for the probability of detection

This is not the only way in which the model could be used. For instance, we envisage a situation where someone (for instance, a qualification body) is asked to examine an ENIQ-qualified NDE system. This particular user could decide to use our model as an aid towards making a final decision on whether the NDE system can indeed be deemed to have passed qualification. Thus, the setting would not be to fix a detection target as an input of the process, but to obtain it as an output, after having assessed and scored all the available evidence (TJ and results from trials, open and blind).

Let us call these two different approaches as “detection target as input” and “detection target as output”. Summarising, “detection target as input” implies a setting where the detection target is explicitly given before the start of the qualification work (compiling the TJ, performing practical trials, etc.). The detection target mathematically determines the required sample size (which includes the TJ equivalent sample size) and, therefore, explicitly influences the qualification work (for instance, affecting the decision regarding the number of practical trials to be carried out). “Detection target as output” implies a setting where the qualification work has been carried out without this explicit target in mind. Our model is only applied in a second stage, for instance as an aid to decide whether the evidence accumulated (TJ then practical trials) is enough to judge the NDE system at hand as having “passed” qualification. The way our Bayesian model is applied is entirely left to the user and his needs.

In the following section we discuss the link between detection target and sample size.

4.5 Link between detection target and sample size

We have argued that a reasonable choice to express the target detection capability is by means of a confidence interval and associated confidence level. Naturally, we want to prove that the inspection capability, in terms of probability of detection, is at least better than some given value. This was expressed by equation (1) as

$$[p_{100\delta\%}, 1]$$

Let us assume that we wish to prove with 95% confidence that the probability of detection is at least 90%. Then δ is equal to 0.95 and $p_{95\%}$ to 0.90. A similar example is treated in Appendix 1, where we have shown (equation A17) that it is sufficient to carry out an experiment of N trials with N_s successes so that the following inequality is satisfied:

$$\delta > 1 - F(x, 1 + N_s, 1 + N - N_s) \tag{10}$$

where F is the cumulative Beta distribution function with parameters $\alpha=1+N_s$ and $\beta=1+N-N_s$. This function is usually found in all mathematical software, including spreadsheets such as Excel (function BETADIST).

Figure 3 has been plotted for the example at hand. The curves plotted are the functions $1-F(x,\alpha,\beta)$, for different combinations of α and β. The abscissa, x, represents the probability of detection, p, whereas the ordinate represents the associated confidence level, δ. If we want that our experiment $\{N, N_s\}$ proves the target set above, i.e. verify the inequality of equation (10), the values N and N_s must be such that the curve $1-F(x,1+N_s,1+N-N_s)$ passes through or above the point (0.90, 0.95), i.e. must intercept the red rectangle of Figure 3.

It can be shown that the smallest sample size required to prove the example target is N=28. In this case, α=29 and β=1. We thus need 28 out of 28 successes to prove the target. Allowing for a single failure, it is easy to calculate that a sample size N=45 is required. In this second case α=45 and β=2, and we thus need 44 successes out of 45 trials. More combinations can be obtained by allowing an increasing numbers of failures.

As we are usually more interested in making the sample size as small as possible, we usually focus our attention on the special situation where the number of failures is zero, and therefore N_s=N. This is done without loss of generality. The inequality of equation (10) can easily be used to determine any combination of N and N_s, including those cases where N_s<N.

The curves plotted in Figure 4 have been obtained in the special case N_s=N. Such curves easily allow determining the required sample size for any given target ($p_{100\delta\%}$, δ). It is sufficient to identify on the graph the point whose Cartesian coordinates (x,y) are x=$p_{100\delta\%}$ and y=δ, and find out the value of N for the first curve above the point. In Figure 4, for clarity, only discrete values of N are represented.

In Figure 5 the required information has been extracted for three confidence levels typically encountered in engineering applications: .90, .95 and .99. This figure shows the lower bound probability of detection versus the sample size (assuming again that N_s=N). For instance, at a confidence level of 99% and after carrying out N=20 trials, all of which are successes, we can only conclude that the lower bound probability of detection is slightly above 0.80.

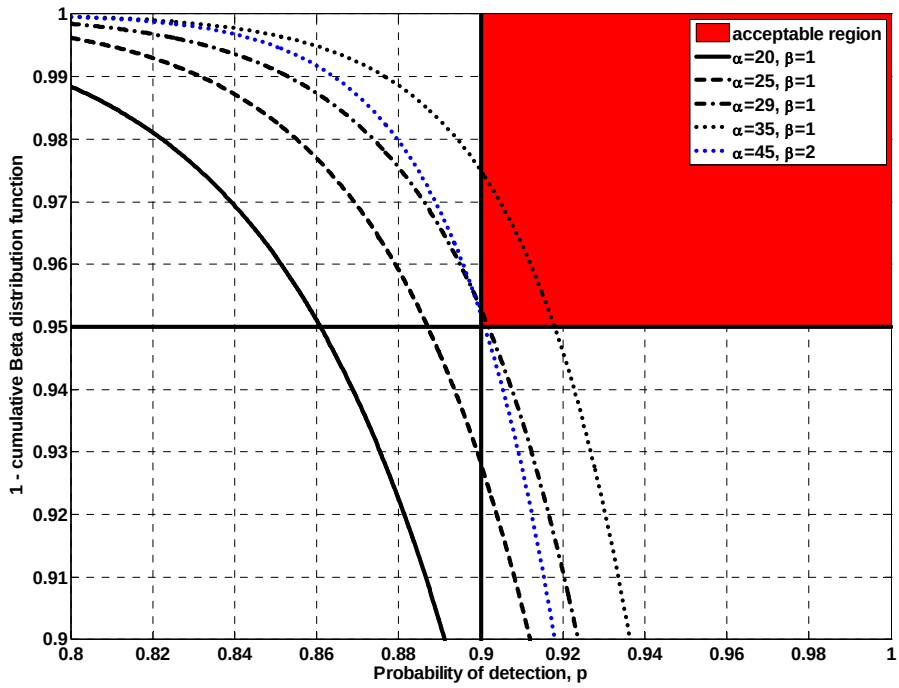

Figure 3 Some examples of 1-*F* curves, with *F* the cumulative Beta distribution function

Figure 5 can be used to determine the required sample size for a given detection target. Let us suppose for instance that we want to prove that the probability of detection is at least 0.95. We enter the plot from the y-axis, drawing a horizontal line at y=0.95. This line encounters the data set for δ=0.90 between N=43 and N=44. Therefore, the smallest sample size able to guarantee the target with 90% confidence would be N=44 (if all successes). The line then encounters the data set for δ=.95 between N=57 and N=58. Therefore, the smallest sample size able to guarantee the target with 95% confidence would be N=58 (if all successes). Finally, the line encounters the data set for δ=.99 between N=88 and N=89. Therefore, the smallest sample size able to guarantee the target with 99% confidence would be N=89 (if all successes). The data extracted in such a way is summarised in Table 4 (along with the same information obtained for the situation where a single failure is recorded, N_f=1).

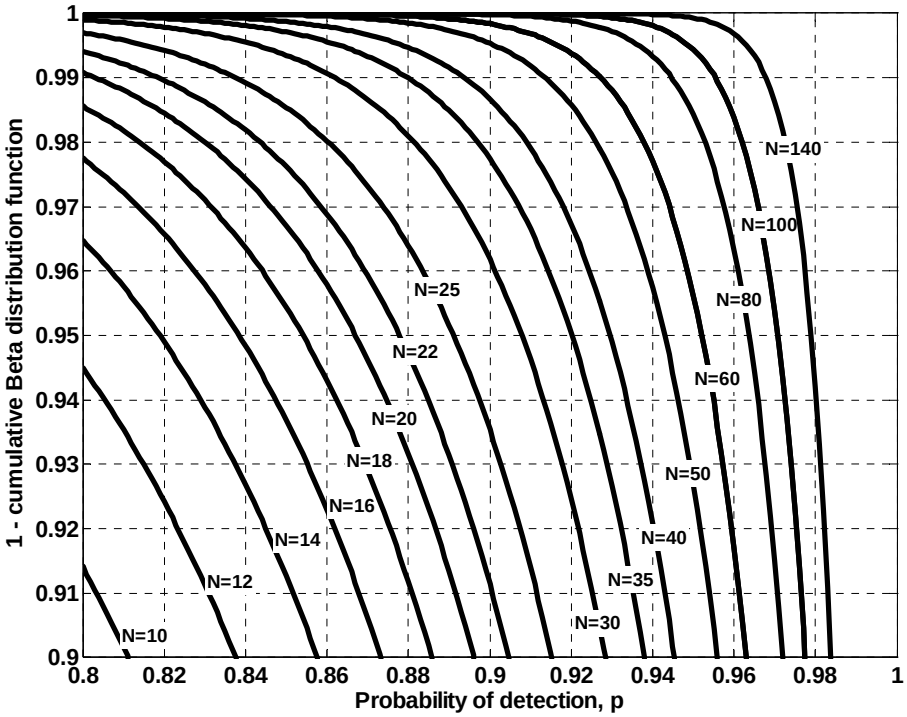

Figure 4 **Examples of 1-*F* curves, with *F* the cumulative Beta distribution function, obtained for the special case N_s=*N***

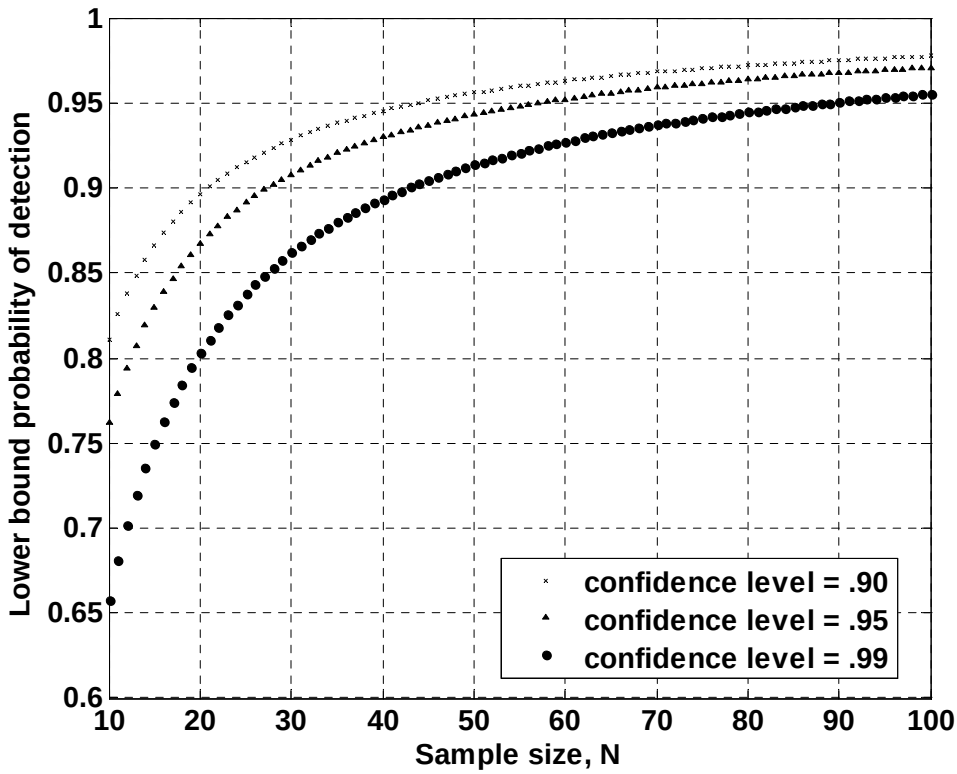

Figure 5 **Lower bound probability of detection, $p_{100\delta\%}$ versus required number of trials**

Table 4 Minimum required number of trials as function of the specified lower bound probability of detection and given confidence level

Lower bound probability of detection, $p_{100\delta\%}$	Number of trials required					
	No failures (N_s=N, N_f=0)			**One failure** (N_s=N-1, N_f=1)		
	δ=0.90	δ=0.95	δ=0.99	δ=0.90	δ=0.95	δ=0.99
0.80	10	13	20	16	20	29
0.81	10	14	21	17	22	31
0.82	11	15	23	19	23	32
0.83	12	16	24	20	24	35
0.84	13	17	26	21	26	37
0.85	14	18	28	23	28	40
0.86	15	19	30	25	30	43
0.87	16	21	33	27	33	47
0.88	18	23	36	29	36	51
0.89	19	25	39	32	40	56
0.90	21	28	43	36	44	62
0.91	24	31	48	40	49	69
0.92	27	35	55	46	56	79
0.93	31	41	63	53	64	90
0.94	37	48	74	62	76	106
0.95	44	58	89	75	91	128
0.96	56	73	112	94	115	162
0.97	75	98	151	127	155	217
0.98	113	148	227	192	234	328
0.99	229	298	458	386	471	>500
0.995	459	>500	>500	>500	>500	>500

5 Examples

Table 5 gives an overview of the various examples covered in this Chapter. We choose to gives three examples covering Approach 1, two examples covering Approach 2 and one example covering Approach 3.

Table 5 Summary of examples

Example	Type of approach	Detection Target	Notes
A1	Approach 1	As output	A TJ is available, with very good scores. Number of trials already fixed. The trials are carried out and all successes are registered. The TJ is weighted (using the number of trials as anchor).
A2	Approach 1	As input	A TJ is available, with very good scores. Detection target is used to decide the required number of trials. Trials are performed and (if all successes) the target is proved.
A3	Approach 1	As input	As Example A2, but TJ has slightly lower scores. Shows that number of practical trials required to prove given target can actually increase.
B1	Approach 2	As input	TJ relative weight decided. TJ still assessed for "degree of satisfaction".
B2	Approach 2	As output	A TJ with very good scores. Number of trials already fixed. The trials are carried out and all successes are registered. The TJ is weighted (using the number of trials as anchor).
C1	Approach 3	As input	Practical trials prove expected value target, TJ used to increase lower bound confidence.

5.1 Examples of Approach 1

5.1.1 Example A1

In the first example, we are asked to make a judgement on the detection capability of an NDE system which is being qualified. The technical justification has already been prepared, assembling the evidence in three main areas: (1) theoretical modelling, (2) experimental evidence and (3) parametric studies. Further, 10 practical trials have been carried out. The NDE system has correctly found all 10 flaws.

Step A1-1 – Decision of the TJ equivalent sample size.
The first decision we are asked to make concerns the TJ equivalent sample size, N_{TJ}. In other words, we need to decide how much we weigh the technical justification as a whole. A possible way forward is to weight the TJ directly against the number of practical trials.

We have performed 10 practical trials. How many equivalent trials do we think the TJ is worth in comparison with this? In other words, if we were to give up completely the TJ, how many additional practical trials would we like to add to the original 10 in order to gain the same confidence that we had in the system capability of detecting defects? Five, ten, or maybe twenty?

We are specifically asked to make an expert judgement and quantify it. We thus need to turn to the TJ and examine it with a critical eye. In the example, we analyse the TJ and we realise that

a lot of resources have been invested in assembling several pieces of evidence, all of which seem to contribute in building a very high confidence that all the required defects will be detected.

We eventually decide that just about 20 additional practical trials would give us the same confidence if we were asked to give up the evidence contained in the TJ. Thus, we decide that:

$$N_{TJ} = 20 \tag{11}$$

Step A1-2 – Decision regarding the relative weights of the TJ elements.
The second decision we are asked to make concerns the relative weights of the TJ elements. We have a Technical Justification in which three main elements have been identified. After examination, we decide that Element 1 contributes towards 30% of the total evidence contained in the TJ, Element 2 is judged to carry 50% of the evidence, and Element 3 20%. These values are summarised in the first column of Table 6, as fractions of 1. The sum of these contributions must necessarily be 1.

Step A1-3 – Decision regarding the score of the TJ elements.
The third step consists of scoring the TJ elements. Crucially, we reiterate the notion that this score must reflect how well the evidence contained in the TJ element supports the detectability of the prescribed defects.

In the example, Elements 1 and 3 are judged to fully support the detectability of all defects in the specified population, and are thus both assigned a score of 100%. Considerations in Element 2 indicate that some limiting defects (such as worst case combinations of size, tilt and skew) could very occasionally be missed. Element 2 is thus scored with a 95%, expressing an intuitive notion that roughly 1 defect in 20 could be missed (purely according to the evidence contained in this Element).

These values are reported in the second column of Table 6, again as fraction of 1.

Step A1-4 – Calculation of TJ total weighted score.
The score of the TJ as a whole is easily obtained. For each element, the score and weight are multiplied and an element weighted score is obtained (third column of Table 6). The elements' weighted scores are finally added together to determine the TJ total weighted score, w_{TJ}.

In the example of Table 6, according to the weight and evaluation given to the elements, the total TJ score is estimated to be 0.975.

Table 6 Hypothetical data used in example A1.

	Relative weight	Score	Weighted score
Element 1	0.3	1	0.3
Element 2	0.5	0.95	0.475
Element 3	0.2	1	.2
	$\Sigma = 1$		**$\Sigma = 0.975 = w_{TJ}$**

Step A1-5 – Calculation of TJ posterior parameters (TJ updating)
We suggest that the TJ total score is straightforwardly used to determine the equivalent number of successes, N_{TJs}, in the following way:

$$N_{TJs} = w_{TJ} \cdot N_{TJ} \tag{12}$$

In other words, the TJ total score is simply interpreted as the fraction of equivalent successes over the equivalent number of trials.

If we start the Bayesian updating process with a uniform prior, as we have suggested before, we have:

$$\alpha_{prior} = 1 \qquad \beta_{prior} = 1 \tag{13}$$

The parameters of the Beta posterior distribution (after updating with the evidence provided by the TJ) are thus obtained as follows:

$$\alpha_{TJ} = 1 + N_{TJs} \qquad \beta_{TJ} = 1 + N_{TJf} = 1 + N_{TJ} - N_{TJs} \tag{14}$$

The expected value and mode of the posterior distribution are (see Appendix 1):

$$\begin{aligned} E(p) &= \frac{\alpha}{\alpha+\beta} = \frac{1+N_{TJ,s}}{2+N_{TJ}} \\ Mode(p) &= \frac{\alpha-1}{\alpha+\beta-2} = \frac{N_{TJ,s}}{N_{TJ}} = (w_{TJ}) \end{aligned} \tag{15}$$

Thus, the mode of the posterior distribution (i.e. the location where the distribution function attains its maximum) is equal to the TJ total score. This is appealing from an intuitive point of view. In the example at hand:

$$N_{TJs} = w_{TJ} \cdot N_{TJs} = 0.975 \times 20 = 19.5 \tag{16}$$

$$\alpha_{TJ} = 20.5 \qquad \beta_{TJ} = 1.5 \tag{17}$$

$$\begin{aligned} E(p) &= 0.932 \\ Mode(p) &= 0.975 \end{aligned} \tag{18}$$

The probability density function of the TJ posterior is plotted in figure 6 with a black dashed line.

Step A1-6 – Updating with evidence from practical trials

The evidence obtained from practical (open) trials can now be taken into account. Since:

$$N_{trials} = 10 \qquad N_{trials,s} = 10 \tag{19}$$

a second posterior is thus easily obtained.

$$\alpha_{trials} = \alpha_{TJ} + N_{trials,s} \qquad \beta_{trials} = \beta_{TJ} + N_{trials,f} \tag{20}$$

In the example at hand:

$$\alpha_{trials} = 30.5 \qquad \beta_{trials} = 1.5 \tag{21}$$

The expected value and mode of the second posterior are:

$$\begin{aligned} E(p) &= 0.953 \\ Mode(p) &= 0.983 \end{aligned} \tag{22}$$

The probability density function of the second posterior is plotted in figure 6 with a black dash-dot line.

Step A1-7 – (optional) Updating with evidence from blind trials
If further evidence, obtained for instance from blind trials, was available, a third posterior could be obtained.

Let us for instance assume that the NDE system in the example at hand has been applied to a set of 15 blind trials. We suppose that a single defect was missed. Then:

$$N_{blind\ trials} = 15 \qquad N_{blind\ trials,s} = 14 \tag{23}$$

A third posterior is again obtained:

$$\alpha_{blind\ trials} = \alpha_{trials} + N_{blind\ trials,s} \qquad \beta_{trials} = \beta_{trials} + N_{blind\ trials,f} \tag{24}$$

Thus

$$\alpha_{blind\ trials} = 44.5 \qquad \beta_{trials} = 2.5 \tag{25}$$

Expected value and mode of the third posterior are:

$$\begin{aligned} E(p) &= 0.947 \\ Mode(p) &= 0.967 \end{aligned} \tag{26}$$

The probability density function of this posterior is plotted in figure 6 with a red dash-dot line.

Step A1-8 – Reporting
As discussed above, a convenient way to offer a complete summary of how the information available is described by the posteriors obtained in the updating process is by means of 1-F curves, with F the cumulative Beta distribution function. In these curves, plotted in figure 7 for the example at hand, the abscissa *x* represents the lower bound probability of detection, *p*, and the ordinate y represents the associated confidence level, δ. Therefore, the curves of figure 7 allows us to obtain directly, for any given required confidence level, the correspondent probability of detection.

Let us for instance assume that no blind trials have been carried out, thus the final posterior is the second one, represented in figure 7 by the black dash-dot line. At a confidence level of 90%, the lower bound probability of detection is found to be just above 0.9.

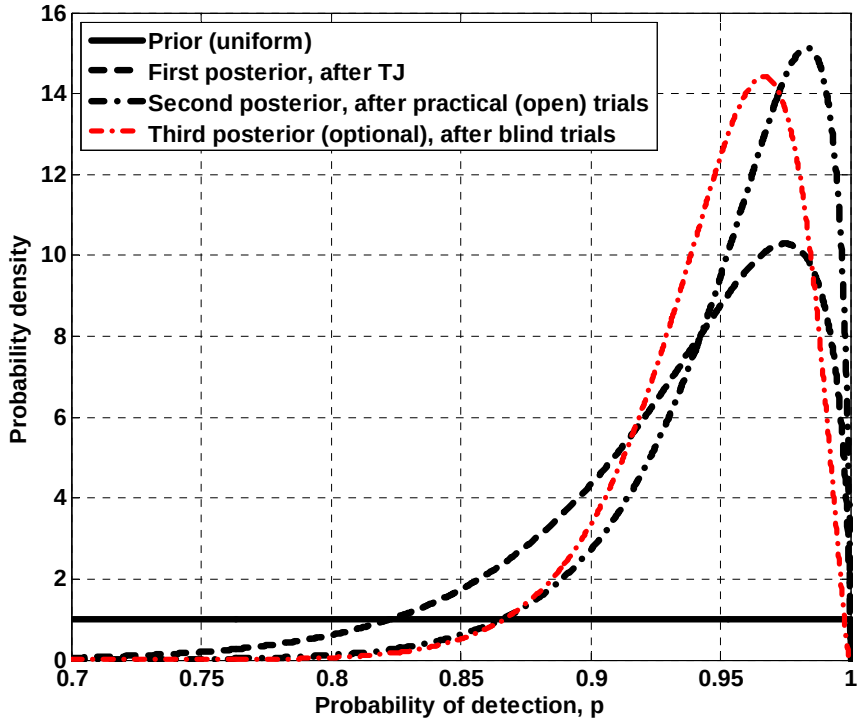

Figure 6 **Probability densities for example A1**

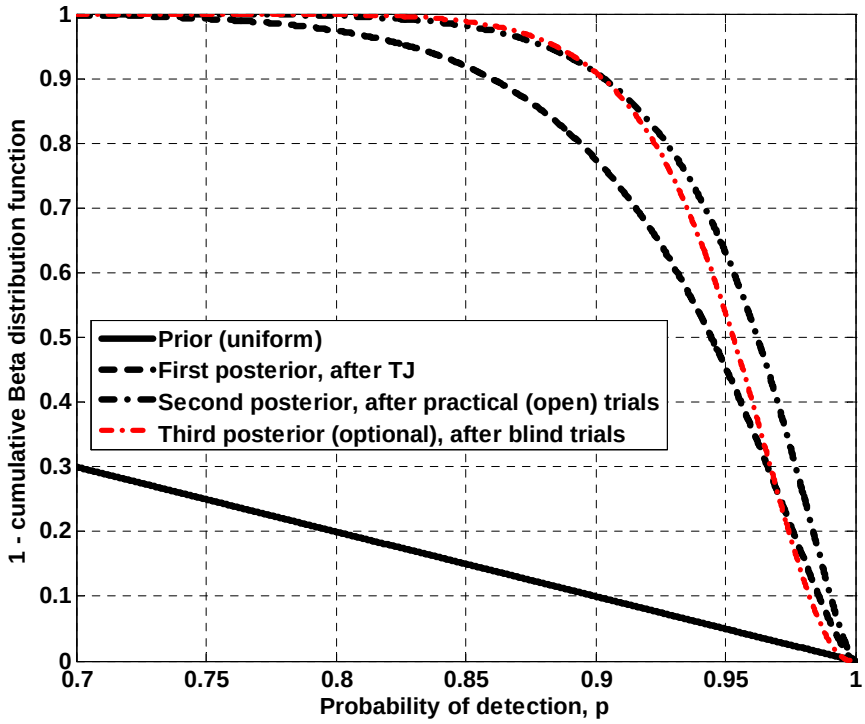

Figure 7 **1-F curves (with F the cumulative Beta distribution function) for example A1**

5.1.2 Example A2

In the second example the detection target is given as input. A technical justification has already been prepared, assembling evidence in two main areas: (1) experimental evidence and (2) parametric studies. We are asked to determine the number (sample size) of practical trials that is required in order to prove the input target.

The input target is the following: show that, with 95% confidence, the lower bound detection probability is at least 80%, or

$$p_{95\%} = 0.8 \tag{27}$$

Step A2-1 – Determination of required total sample size.
As discussed above (section 4.5), the total sample size can be easily derived from a detection target determining which 1-F curves (i.e. which combinations of parameters α and β) are such that the curves pass through or above the point (0.80, 0.95).

Figure 8 has been plotted for the example at hand. The curves plotted are the functions 1-$F(x,\alpha,\beta)$, for different combinations of α and β, that pass as close as possible (and above) the target point (0.80, 0.95).

For zero failures (β=1), the minimum required sample size would be N=13 (α=14, N_s=13). In the case of one failure (β=2), N=21 (α=21, N_s=20). In the case of two failures (β=3), N=29 (α=28, N_s=27), etc.

As we have not yet carried out the experiments, we keep these different sample sizes in our mind, without committing for the time being to any of them in particular.

Step A2-2 – Decision on the TJ equivalent sample size.
We are asked to determine the TJ equivalent sample size, N_{TJ}, i.e. how much we weigh the technical justification as a whole.

We may feel that in the TJ at hand a reasonable amount of evidence has been assembled supporting detectability. Taking also into account the considerations from Step A2-1 (from which, unless a higher number of failures is obtained, we expect the total sample size to be in the range 15-30) we eventually decide that the TJ at hand is worth about 10 practical trials. Thus

$$N_{TJ} = 10 \tag{28}$$

Step A2-3 – Decision regarding the relative weights of the TJ elements.
We must now assess the relative weights of the TJ elements. We have a Technical Justification in which two main elements have been identified. After examination, we decide that Element 1 contributes towards 70% of the total evidence contained in the TJ and Element 2 the remaining 30%. These values are summarised in the first column of Table 7, as fractions of 1. The sum of these contributions must necessarily be 1.

Step A2-4 – Decision regarding the score of the TJ elements.
The fourth step consists of scoring the TJ elements. This score must reflect how well the evidence contained in the TJ element supports the detectability of the prescribed defects.

In the example, Element 2 is judged to fully support the detectability of all defects in the specified population, and is thus assigned a score of 100%. On the other hand, Element 1 (experimental evidence) is slightly more problematic, as data from old experiments seems to

indicate that some defects could be missed. Element 2 is thus scored at 90%, expressing an intuitive notion that roughly 1 defect in 10 could be missed. These values are reported in the second column of Table 7, again as fraction of 1.

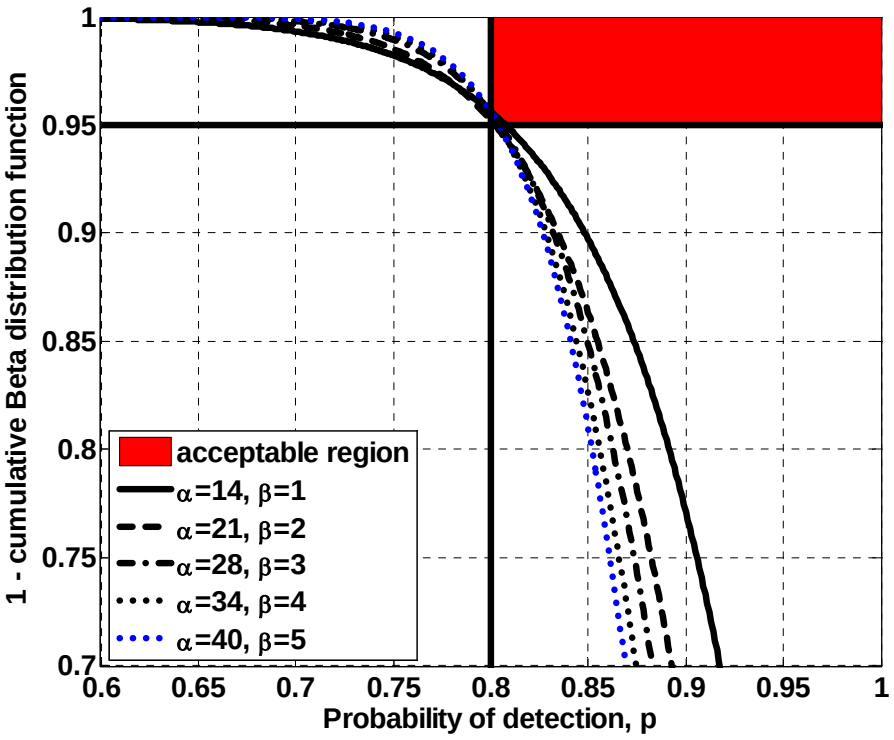

Figure 8 1-F curves (with F the cumulative Beta distribution function) for example A2

Step A2-5 – Calculation of TJ total weighted score.
The score of the TJ as a whole is easily obtained. For each element, the score and weight are multiplied and an element weighted score is obtained (third column of Table 7). The elements' weighted scores are finally added together to determine the TJ total weighted score, w_{TJ}. In this example, according to the weight and evaluation given to the elements, the total TJ score is estimated to be 0.93.

Table 7 Hypothetical data used in example A2.

	Relative weight	Score	Weighted score
Element 1	0.7	0.90	0.63
Element 2	0.3	1	0.30
	Σ = 1		**Σ = 0.93= w_{TJ}**

Step A2-6 – Calculation of TJ posterior parameters (TJ updating)
As before, we suggest that the TJ total score is used to determine the equivalent number of successes, N_{TJs}, in the following way:

$$N_{TJs} = w_{TJ} \cdot N_{TJ} \tag{29}$$

Starting as usual the Bayesian updating process with a uniform prior, we have:

$$\alpha_{prior} = 1 \qquad \beta_{prior} = 1 \tag{30}$$

The parameters of the Beta posterior distribution (after updating with the evidence provided by the TJ) are thus obtained:

$$N_{TJs} = w_{TJ} \cdot N_{TJs} = 0.93 \times 10 = 9.3 \tag{31}$$

$$\alpha_{TJ} = 10.3 \qquad \beta_{TJ} = 1.7 \tag{32}$$

Step A2-7 – Determination of minimum sample size of practical trials

Let us consider again Figure 8. We can see that achieving the target by means of the first curve, which was obtained in the case of zero failures (α=14, β=1), is now impossible. The TJ has been evaluated as carrying 9.3 equivalent successes and therefore 0.7 equivalent failures. The next "best" possibility (in terms of minimizing the number of required trials) is now offered by the second curve plotted in Figure 8. Such a curve was obtained for a single failure, with a total sample size N=21 (α=21, β=2, N_s=20). Starting with N_{TJs}=9.3, we can achieve the input detection target if we can demonstrate that

$$N_{trials} = 11 \qquad N_{trials,s} = 11 \qquad N_{trials,f} = 0 \tag{33}$$

because the second posterior would then be characterised by

$$\alpha_{trials} = \alpha_{TJ} + N_{trials,s} = 21.3 \qquad \beta_{trials} = \beta_{TJ} + N_{trials,f} = 1.7 \tag{34}$$

which is better than (α=21, β=2), thus bettering the required target. The target input would then be demonstrated if 11 out of 11 defects were detected in practical trials.

Allowing for failures in practical trials would necessarily increase the sample size. For instance, the input detection target could be also demonstrated if

$$N_{trials} = 19 \qquad N_{trials,s} = 18 \qquad N_{trials,f} = 1 \tag{35}$$

because the second posterior would then be characterised by

$$\alpha_{trials} = \alpha_{TJ} + N_{trials,s} = 28.3 \qquad \beta_{trials} = \beta_{TJ} + N_{trials,f} = 2.7 \tag{36}$$

which is better than (α=28, β=3, third curve of Figure 8), also bettering the required target. The target input would then be demonstrated if 18 out of 19 defects were detected in practical trials, etcetera.

Step A2-8 – Performance of practical trials

We can use the discussion above to decide to manufacture 11 defects. The NDE system at hand is then applied to these defects. If all 11 are detected, the input target can be judged to be demonstrated (we note again that this conclusion is an expert judgement based on the earlier assumptions of TJ weight and scores).

Discussion of example A2

The situation presented in this example could easily be reversed. For instance, there could be a situation in which the number of trials is decided beforehand, and the given (input) detection target is used to determine the required TJ equivalent sample size. It would then be up to the user to convert this number into the "amount of evidence" required to compile an appropriate TJ.

In general, it could be argued that very often the evidence to be compiled into the TJ and the number of practical trials are decided together. This example only explores a possible way of applying our methodology.

An interesting point, worthy of note, is the fact the in the example, having a TJ only results in an improvement of two practical trials at best. Proving the target without the TJ would have required finding 13 out of 13 defects in practical trials. With the TJ having the properties described above, the target can be proved by finding 11 out of 11 practical trials. This fact, discussed in greater detail at the end of Example A3, is due to the equivalent number of failures carried by the TJ, which is not zero.

5.1.3 Example A3

The third example is very similar to the previous one: the detection target is given as input, and again we are asked to determine the number (sample size) of practical trials that is required in order to prove the input target. A technical justification has already been prepared, assembling evidence in four main areas: (1) Theoretical modelling, (2) Experimental evidence, (3) Parametric studies and (4) Equipment and data analysis.

The main difference from Example A2 will be that the TJ is given lower scores and the detection target is set higher.

The input target is the following: show that, with 95% confidence, the lower bound detection probability is at least 90%, or

$$p_{95\%}=0.9 \tag{37}$$

Step A3-1 – Determination of required total sample size.
Again, the total sample size is derived from the input detection target determining which 1-F curves (i.e. which combinations of parameters α and β) are such that the curves pass through or above the point (0.90, 0.95). Figure 9 has been plotted for the example at hand. The curves plotted are the functions 1-$F(x,\alpha,\beta)$, for different combinations of α and β, that pass as close as possible (and above) the target point (0.90, 0.95).

For zero failures (β=1), the minimum required sample size would be N=28 (α=29, N_s=28). In the case of one failure (β=2), N=45 (α=45, N_s=44). In the case of two failures (β=3), N=60 (α=59, N_s=58), etc.

As we have not yet carried out the experiments, we keep these different sample sizes in our mind, without committing for the time being to any of them in particular.

Step A3-2 – Decision on the TJ equivalent sample size.
We are asked to consider the TJ equivalent sample size, N_{TJ}, i.e. how much we weigh the technical justification as a whole.

We may feel that in the TJ at hand quite a large amount of evidence has been assembled. We eventually decide that the TJ at hand is worth about 30 practical trials. Thus

$$N_{TJ} = 30 \tag{38}$$

Step A3-3 – Decision regarding the relative weights of the TJ elements.
We must now assess the relative weights of the TJ elements. We have a Technical Justification in which four main elements have been identified. After examination, we decide that Element 1 contributes towards 20% of the total evidence contained in the TJ, Element 2 40%, Element 3 30% and Element 4 the remaining 10%. These values are summarised in the first column of Table 8, as fractions of 1. The sum of these contributions must necessarily be 1.

Step A3-4 – Decision regarding the score of the TJ elements.
The fourth step consists of scoring the TJ elements. This score must reflect how well the evidence contained in the TJ element supports the detectability of the prescribed defects.

As stated above in this example the TJ is scored slightly lower than before, but still with scores equal to or above 90%. For instance, Element 1 and Element 4 are scored 95%, whereas Element 2 and Element 3 are scored 90%. These values are reported in the second column of Table 8, again as fraction of 1.

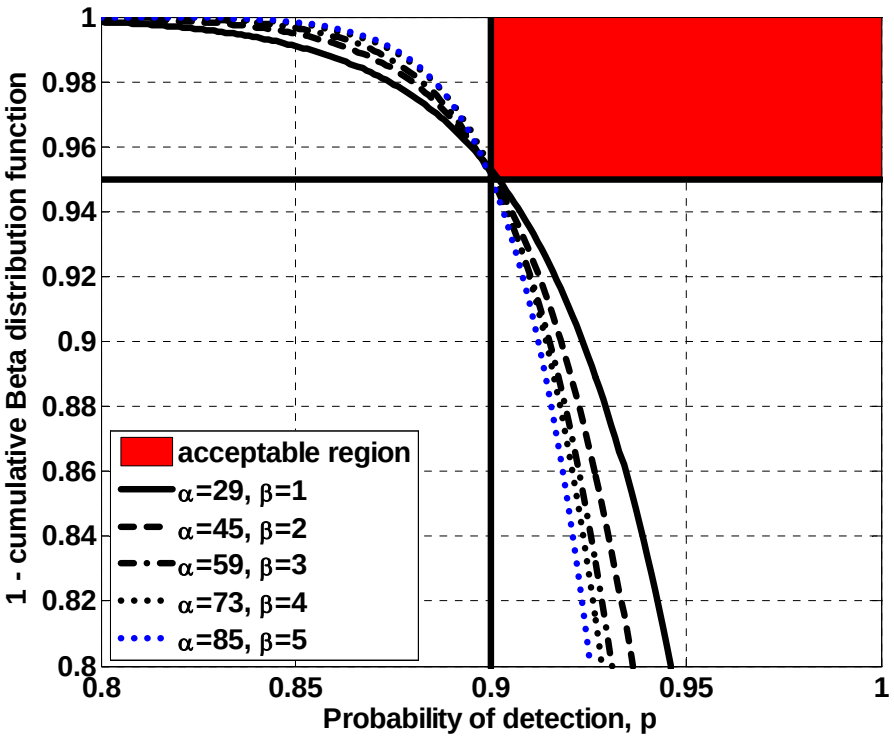

Figure 9 1-F curves (with F the cumulative Beta distribution function) for example A3

Step A3-5 – Calculation of TJ total weighted score.
The score of the TJ as a whole is easily obtained. For each element, the score and weight are multiplied and an element weighted score is obtained (third column of Table 8). The elements' weighted scores are finally added together to determine the TJ total weighted score, w_{TJ}. In this example, according to the weight and evaluation given to the elements, the total TJ score is estimated to be 0.915.

Table 8 Hypothetical data used in example A3.

	Relative weight	Score	Weighted score
Element 1	0.2	0.95	0.19
Element 2	0.4	0.90	0.36
Element 3	0.3	0.90	0.27
Element 4	0.1	0.95	0.095
	$\Sigma = 1$		**$\Sigma = 0.915 = w_{TJ}$**

Step A3-6 – Calculation of TJ posterior parameters (TJ updating)
As before, we suggest that the TJ total score is used to determine the equivalent number of successes, N_{TJs}, in the following way:

$$N_{TJs} = \mathrm{w_{TJ}} \cdot N_{TJ} \tag{39}$$

Starting as usual the Bayesian updating process with a uniform prior, we have:

$$\alpha_{prior} = 1 \qquad \beta_{prior} = 1 \tag{40}$$

The parameters of the Beta posterior distribution (after updating with the evidence provided by the TJ) are thus obtained:

$$\begin{aligned} N_{TJs} &= \mathrm{w_{TJ}} \cdot N_{TJs} = 0.915 \times 30 = 27.45 \\ N_{TJf} &= N_{TJs} - N_{TJs} = 30 - 27.45 = 2.55 \end{aligned} \tag{41}$$

$$\alpha_{TJ} = 28.45 \qquad \beta_{TJ} = 3.55 \tag{42}$$

Step A3-7 – Determination of minimum sample size of practical trials
Let us now consider Figure 9. We can see that it is now actually impossible to achieve the target by means of the first three curves, which were obtained in the case of zero, one and two failures respectively. The TJ has been evaluated as carrying 2.55 equivalent failures. The next "best" possibility (in terms of minimizing the number of required trials) is now offered by the fourth curve plotted in Figure 9. Such a curve was obtained for three failures, with a total sample size N=75 (α=73, β=4, N_s=72). Starting with N_{TJs}=27.45, we can achieve the input detection target if we can demonstrate that

$$N_{trials} = 45 \qquad N_{trials,s} = 45 \qquad N_{trials,f} = 0 \tag{43}$$

because the second posterior would now be characterised by

$$\alpha_{trials} = \alpha_{TJ} + N_{trials,s} = 73.45 \qquad \beta_{trials} = \beta_{TJ} + N_{trials,f} = 3.55 \tag{44}$$

which is better than (α=73, β=4), thus bettering the required target. The target input would then be demonstrated if 45 out of 45 defects were detected in practical trials. Allowing for failures in the practical trials would further increase the sample size.

Step A3-8 – Performance of practical trials
As in Example A2, we can use the discussion above to decide to manufacture 45 defects. The NDE system at hand is then applied to these defects. If all 45 are detected, the input target can be judged demonstrated.

Discussion of Example A3
We have now an apparent paradox, which is worth discussing. We have assembled a TJ which, according to our own judgement, contained a lot of evidence. We probably spent a great deal of resources (time, money, etc.) piecing it together. The TJ was scored rather highly, with each individual Element receiving at least 90%. Yet the conclusion is that in order to prove the input detection target, 45 practical trials on defective components are needed, and all defects must be found.

Consider again Figure 9. The first curve (α=29, β=1) shows that the target could be achieved simply inspecting and detecting flaws in 28 out of 28 defective components. Now, after compiling the TJ and factoring in the evidence carried by it, we are forced to conclude that we actually need <u>more</u> practical trials.

From a mathematical point of view, this fact is easily explained. If we interpret how the TJ score is converted into the two parameters α_{TJ} and β_{TJ} of the TJ posterior distribution, it is straightforward to see how any number less than unity is bound to yield a β_{TJ} greater than zero, which in turn can be interpreted as a number of "equivalent failures" carried by the TJ.

From a more practical point of view, it is clear that the problem lies with the way the TJ score is interpreted. A TJ total score of 90% is converted into the quantitative statement: 90% of the defects of our "equivalent" set of experiment were successfully detected; the remaining 10% were missed. This, according to the total sample size that is chosen to represent the TJ (say for instance 10), has in the end the same influence that finding 9 out of 10 real defects (and therefore missing one) would have when it comes to drawing statistical conclusions.

In conclusion, the above-illustrated problem is not a deficiency of the model itself but rather of the interpretation or definition of a "high confidence". If one cannot be sufficiently convinced of the evidence provided by the elements of the TJ, it is unreasonable to set very high requirements for the POD to be achieved. On the other hand, it is very difficult to quantify the TJ score if it is quite high anyway. How for instance do we distinguish a score of 95% from one of 99%?

5.2 Examples of Approach 2

As discussed above, Approach 1 is sound from a conceptual point of view, but its application can lead to problems.

In our alternative Approach 2, we made the following case. As resources have been spent towards the compilation of the TJ, it can be argued that a TJ could be scored 100%, unless the TJ could explicitly point out some intrinsic limitations that would prevent the achievement of the qualification targets.

If we are in a situation where the evidence contained in the TJ does not explicitly point out intrinsic limitations, we can then assume that each Element of the TJ is automatically scored 100%. The strength or relative weakness of the TJ is then simply reflected through its equivalent sample size. Thus, a "small" TJ, compiled with little resources and for instance covering only one Element, could be represented by a smaller number of equivalent trials. A "large" TJ, where a great deal of resources is invested and for instance covering several Elements are extensively covered, could be represented by a larger number of equivalent trials.

As a starting point we could pre-define the maximum relative weight that a "perfect" TJ would get, and the relative weakness of the TJ would lower this weight. We have already seen that some general guidance for ENIQ recommended practice 3 [3] gives some qualitative guidance on the relative weight that the TJ would likely have in comparison with practical trials for different types of qualifications, see Table 1.

In Approach 2, instead of assigning to the elements the scores that reflected the how the TJ supports the detectability of the defects, we define a measure of satisfaction to reflect how extensive the evidence assembled in the TJ is. Note that one can use a similar identification of elements and their weights as described in Approach 1, in order to break the analysis of the TJ in smaller entities, and thus improve the transparency of the quantification process.

In the following two examples we explore how the quantification of a qualification process could be achieved under Approach 2.

5.2.1 Example B1

In the first example the detection target is given as input. The qualification is taking place with the aim of justifying the inspection procedure. A technical justification has already been prepared, assembling evidence in three main areas: (1) experimental evidence, (2) theoretical modelling and (3) parametric studies. We are asked to determine the number (sample size) of practical trials that is required in order to prove the input target.

The input target is the following: show that, with 95% confidence, the lower bound detection probability is at least 95%:

$$p_{95\%}=0.95 \tag{45}$$

In this example we follow an approach base on the following steps:

1. Decide, for the type of qualification at hand, the maximum relative weight of the TJ (versus practical trials);
2. Identify the principal elements of the TJ. Decide the relative weight of each element and determine the "satisfaction" of each;
3. Determine the TJ equivalent number of trials, N_{TJ};
4. Determine the number of practical trials necessary to achieve the desired target.

The maximum relative weight of the TJ (versus practical trials) is the relative number of equivalent trials that a "perfect" TJ would carry. (We mean here with "perfect" TJ not that the TJ corresponds to an infinite number of trials, but rather that all the evidence – which can be limited – contained in the TJ points to a 100% detection).

Step B1-1 – Determination of required total sample size.
The total sample size is derived from the detection target determining which 1-F curves (i.e. which combinations of parameters α and β) are such that the curves pass through or above the point (0.95, 0.95).

Figure 10 has been plotted for the example at hand. The curves plotted are the functions 1-$F(x,\alpha,\beta)$, for three combinations of α and β, that pass as close as possible (and above) the target point (0.95, 0.95). For zero failures (β=1), the minimum required sample size would be N=58 (α=59, N_s=59). In the case of one failure (β=2), N=92 (α=92, N_s=91). In the case of two failures (β=3), N=123 (α=122, N_s=121).

Step B1-2 – Decision on the maximum relative weight of the TJ.
In this example, we first decide the maximum relative weight of the TJ (versus practical trials). This decision could have even been made, for instance by the regulators or the qualification body, for the type of qualification at hand long before the start of the particular qualification exercise. We have seen that ENIQ recommended practice 3, [3], gives some qualitative guidance on the relative weight that the TJ would likely have in comparison with practical trials for different types of qualifications (Table 1). In principle, a similar table could be agreed upon between the parties involved where quantified values (or more likely, ranges), are given. An example is given in Table 9, with the caveat that the quantitative ranges given there are purely illustrative and should by no means be taken as guidance.

One possible objection would be that agreeing on the values for such a table would be near to impossible, considering the every qualification exercise is somehow unique. Indeed, ENIQ recognises that for certain types of qualifications, such as one aimed at justifying the inspection procedure, the relative weight of the TJ varies. It is thus, and again, a matter of expert judgement as to which choice the user should make for a sound value of the relative weight that the TJ should carry.

Let us suppose that, for the example at hand, it is decided that the TJ should have a maximum relative weight, r_{TJ}, of 0.5 (i.e. 50% of the total, TJ + trials) when compared to the practical trials.

$$r_{TJ} = 0.5 \tag{46}$$

Table 9 Quantified relative weights of technical justification with respect to test piece trials (the values given are purely illustrative)

Type of TJ	Overall weight of TJ (from ENIQ RP3, [3])	Range of relative TJ weights
Justify inspection procedure	Varies	0.1-0.8
Justify use of test pieces and defect populations	Small	0.1-0.3
Justify inspection equipment	Varies	0.1-0.8
Extend qualification to different geometry	Large	0.8-1
Extend qualification to different material structure	Varies	0.1-0.8
Qualify upgraded equipment or software	Large	0.8-1
Qualify upgraded procedure	Large	0.8-1
Qualify for changed defect descriptions	Large	0.8-1

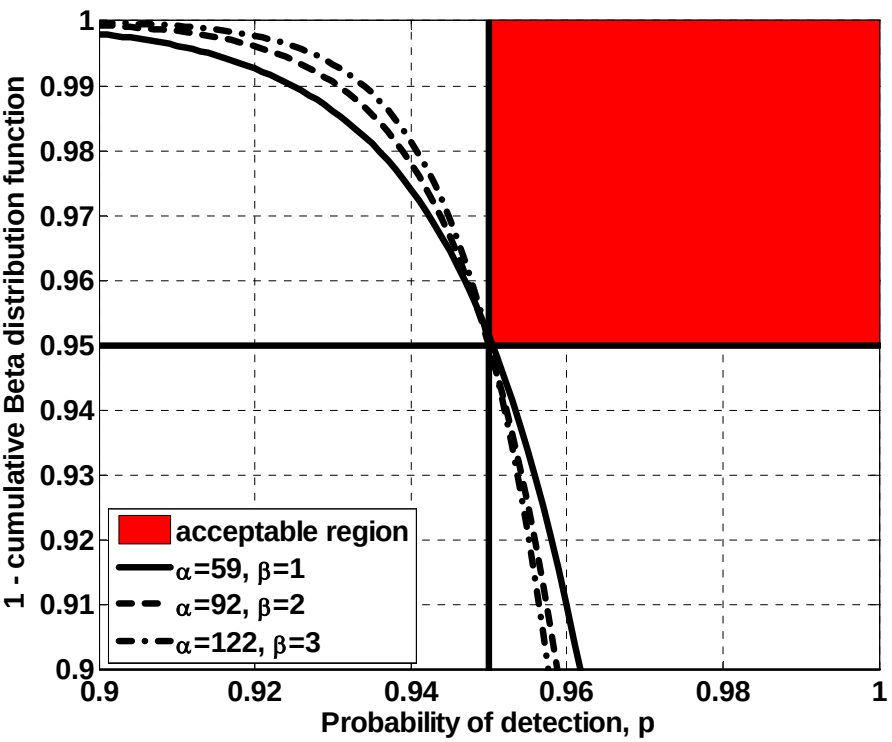

Figure 10 1-F curves (with F the cumulative Beta distribution function) for example B1

Step B1-3 – Decision of the relative weight of each element
We then turn to the TJ, and assess how much weight each element should carry. Using our judgement, we decide that: Element 1 carries 30% of the TJ weight, Element 2 50% of the TJ weight and Element 3 the remaining 20% of the TJ weight. These values are summarised in the first column of Table 10.

Step B1-4 – Decision on the degree of satisfaction of each element
Again using our judgement, we analyse the TJ to see how "satisfactorily" the various elements are covered. We then assign a degree of satisfaction to the various elements. For instance, Element 1 receives a 90% score, Element 2 receives a 100% score and Element 3 receives a 95% score. These values are summarised in the second column of Table 10.

Step B1-5 – Calculation of TJ equivalent sample size

The total "degree of satisfaction" of the TJ, s_{TJ}, is thus obtained adding the individual element contributions (third column of Table 10).

Table 10 Hypothetical data used in example B1.

	Relative weight	Degree of satisfaction	Weighted degree of satisfaction
Element 1	0.30	0.90	0.27
Element 2	0.50	1	0.5
Element 3	0.20	0.95	0.19
	$\Sigma = 1$		**$\Sigma = 0.96 = s_{TJ}$**

The relative weight of the TJ is then obtained scaling down the maximum relative weight, decided at Step B1-2, by the factor s_{TJ}.

$$W_{TJ}=s_{TJ}\times r_{TJ}=0.5\times 0.96=0.48 \tag{47}$$

We have seen (step B1-1) that for zero failures, the minimum required sample size is N=58. In this approach, the TJ is modelled as equivalent successes only. Therefore:

$$N_{TJ}=N_{TJs}=N\times W_{TJ}=0.48\times 58 \qquad = 27 \tag{48}$$

Note that the numbers obtained multiplying N and W_{TJ} are obtained rounding to the next lower integer, as we cannot have fractions of practical trials.

Step B1-6 – Determination of sample size of practical trials
The number of practical trials required to prove the input target follows:

$$N_{trials}=N-N_{TJs} \qquad = 31 \qquad \text{(allowing no failures)} \tag{49}$$

In conclusion, we would need to inspect 31 defective components and find all flaws to prove the input target.

5.2.2 Example B2

In the second example, we are asked to make a judgement on the detection capability of a NDE system which is being qualified. The technical justification has already been prepared, assembling the evidence in three main areas: (1) theoretical modelling, (2) experimental evidence and (3) parametric studies. Further, 8 practical trials have been carried out. The NDE system has correctly found all 8 flaws. Further, a set of blind trials is carried out. This information is known to the qualification body who is performing the quantification exercise. The result from the blind trials is the following: 11 out of 12 defects were correctly identified, one was missed.

Step B2-1 – Decision on the TJ equivalent sample size.
The first decision we are asked to make regards the TJ equivalent sample size, N_{TJ}. In other words, we need to decide how much we weigh the technical justification as a whole, <u>in terms of successes only</u>. Again, a possible way forward is to weight the TJ directly against the number of practical trials. We must bear in mind that in this case, this decision will fully represent the TJ, as no further scoring will take place. We may for instance decide that the TJ is equivalent to 15 successful practical trials.

$$N_{TJ} = 15 \tag{50}$$

Step B2-2 – Calculation of TJ posterior parameters (TJ updating)
As suggested, the TJ is quantified in terms of equivalent successes only:

$$N_{TJs} = N_{TJ} \tag{51}$$

If we start the Bayesian updating process with a uniform prior, as usual, we have:

$$\alpha_{prior} = 1 \qquad \beta_{prior} = 1 \tag{52}$$

The parameters of the Beta posterior distribution (after updating with the evidence provided by the TJ) are thus obtained as follows:

$$\alpha_{TJ} = 1 + N_{TJs} \qquad \beta_{TJ} = 1 \tag{53}$$

In the example at hand:

$$N_{TJs} = 15 \tag{54}$$

$$\alpha_{TJ} = 16 \qquad \beta_{TJ} = 1 \tag{55}$$

Step B2-3 – Updating with evidence from practical trials
The evidence obtained from practical (open) trials can now be taken into account. Since:

$$N_{trials} = 8 \qquad N_{trials,s} = 8 \tag{56}$$

A second posterior is thus obtained.

$$\alpha_{trials} = \alpha_{TJ} + N_{trials,s} \qquad \beta_{trials} = \beta_{TJ} + N_{trials,f} \tag{57}$$

In the example at hand:

$$\alpha_{trials} = 24 \qquad \beta_{trials} = 1 \tag{58}$$

Step B2-4 – Updating with evidence from blind trials
As assumed above, we have

$$N_{blind\ trials} = 12 \qquad N_{blind\ trials,s} = 11 \tag{59}$$

A third posterior is obtained:

$$\alpha_{blind\ trials} = \alpha_{trials} + N_{blind\ trials,s} \qquad \beta_{trials} = \beta_{trials} + N_{blind\ trials,f} \tag{60}$$

Thus:

$$\alpha_{blind\ trials} = 35 \qquad \beta_{trials} = 2 \tag{61}$$

The expected value and mode of the third posterior are:

$$\begin{aligned} E(p) &= 0.946 \\ Mode(p) &= 0.971 \end{aligned} \tag{62}$$

All three posterior probability densities are plotted in Figure 11.

Step B2-5 – Reporting
As in example A1, which we discussed above, we offer a complete summary of how the information available is described by the posteriors obtained in the updating process by means of 1-F curves, with F the cumulative Beta distribution function. In these curves, plotted in Figure 12 for the example at hand, the abscissa *x* represents the lower bound probability of detection, *p*, and the ordinate y represents the associated confidence level, δ. Therefore, the curves of Figure 12 allow us to obtain directly, for any given required confidence level, the correspondent probability of detection.

For instance, we can draw a horizontal line corresponding to δ=0.9. This line intercepts the 1-F curve corresponding to the TJ posterior at x=0.865, the open trials posterior at x=0.908 and the blind trials posterior at x=0.896. Therefore, the 90% confidence lower bound probability of detection is

$$\begin{aligned} p_{90\%} &= 0.865 && \text{(after the TJ updating)} \\ p_{90\%} &= 0\ 908 && \text{(after the open trials updating)} \\ p_{90\%} &= 0.896 && \text{(after the blind trials updating)} \end{aligned} \tag{63}$$

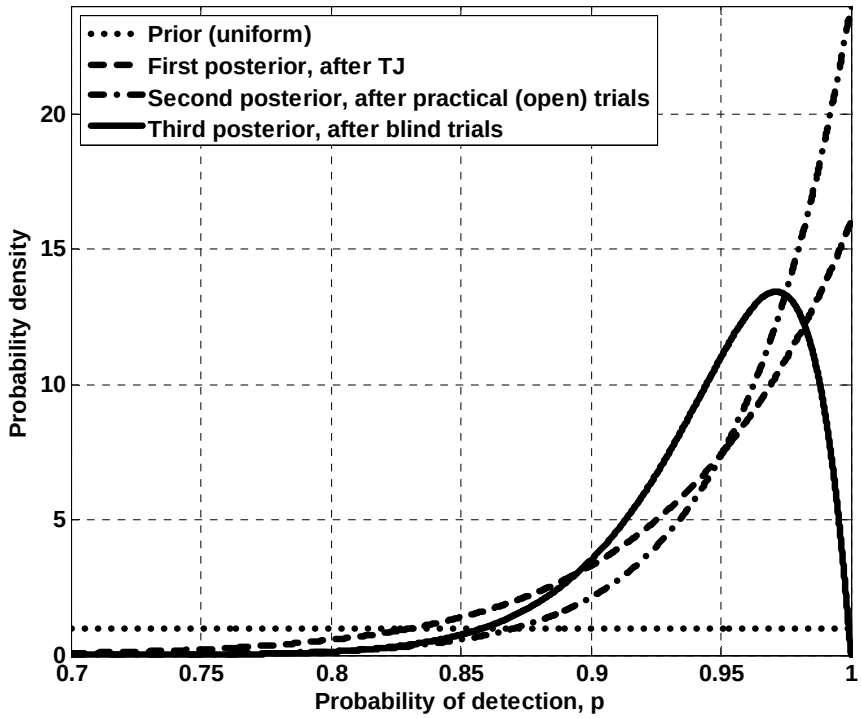

Figure 11 Probability densities for example B2

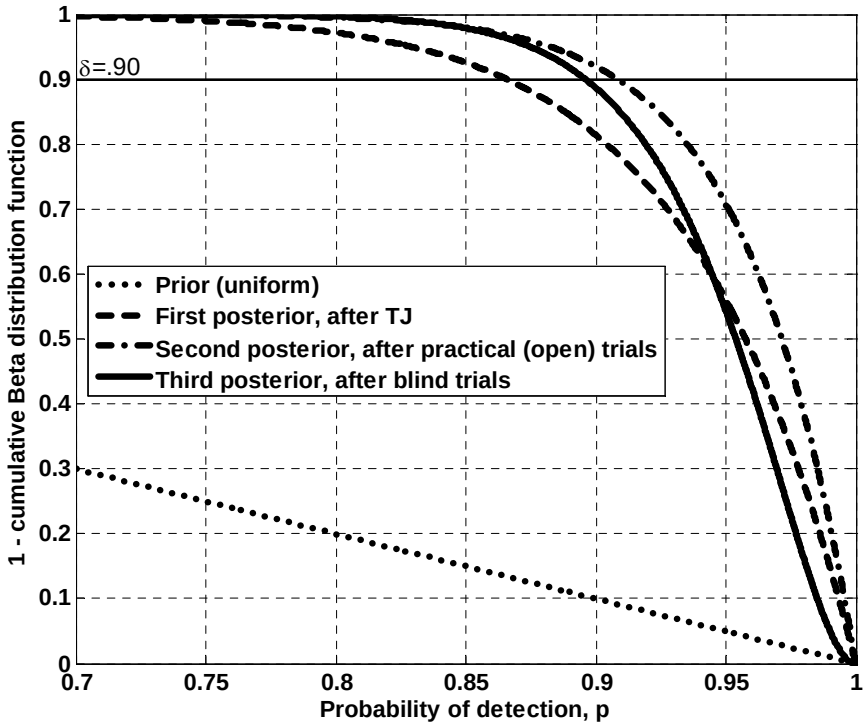

Figure 12 1-F curves (with F the cumulative Beta distribution function) for example B2

5.3 Example of Approach 3

As described above, in Approach 3 we suggest that two detection targets are set. The first is a given expected value for *p*, the second a given lower bound for *p*. The first is then demonstrated using a number of practical trials, the second is proved with the aid of the TJ.

5.3.1 Example C1

Step C1-1 – Fixing the detection targets.
In this example, we want to prove with practical trials that the expected value of the detection probability is 0.95.

$$\mu = 0.95 \tag{64}$$

Further, we would like to use the TJ to prove that the 95% lower bound detection probability is 0.9:

$$p_{95\%} = 0.90 \tag{65}$$

Step C1-2 – Determination of practical trials sample size.
Equation (9) gives the required number of trials needed to obtain a given expected value $\mu = E(p)$, assuming no failures:

$$N_{trials} = \frac{2\mu - 1}{1 - \mu} \tag{66}$$

We can see from Figure 2 and Table 3 that the required sample size (in the case of no failures) is 18. Thus:

$$\begin{aligned} N_{trials} &= 18 \\ N_{trials,s} &= 18 \\ N_{trials,f} &= 0 \end{aligned} \tag{67}$$

Clearly, this will actually need to be achieved in practice. If one failure is recorded, the input targets cannot be proven. Note that this reduction in the number of trials comes of course with a "price": in case of 18 hits out of 18 trials, $p_{.95}$ = 0.854, i.e. the 95% confidence lower bound value is "only" just above 85%.

Step C1-3 – Determination of practical trials sample size.
At this point, the TJ comes into play. The dashed line of Figure 13 represents the posterior knowledge (in terms of 1 minus cumulative Beta distribution) after the practical trials updating. This line is such that the distribution expected value is precisely 0.95, as required by equation (64).

The second target, equation (65), now requires that the curve is pushed (by means of the TJ updating) to intercept or be above the point (0.9, 0.95). The solid line represents the first case in which the target could be achieved, and it is obtained for N_{TJ}=10. We thus need, to prove the second target:

$$\begin{aligned} N_{TJ} &= 10 \\ N_{TJ,s} &= 10 \\ N_{TJ,f} &= 0 \end{aligned} \tag{68}$$

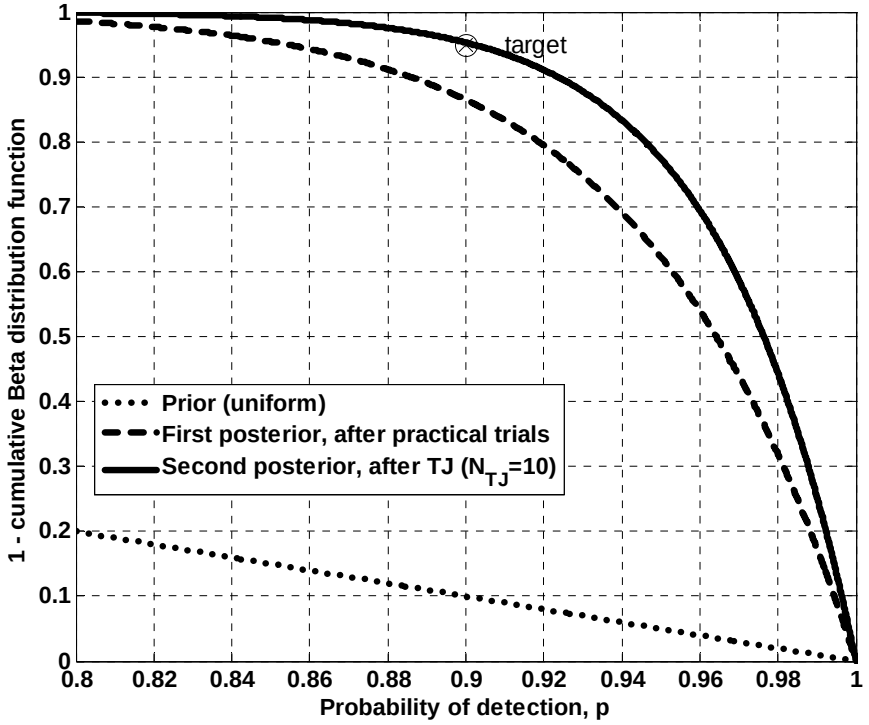

Figure 13 **1-F curves (with F the cumulative Beta distribution function) for example C1**

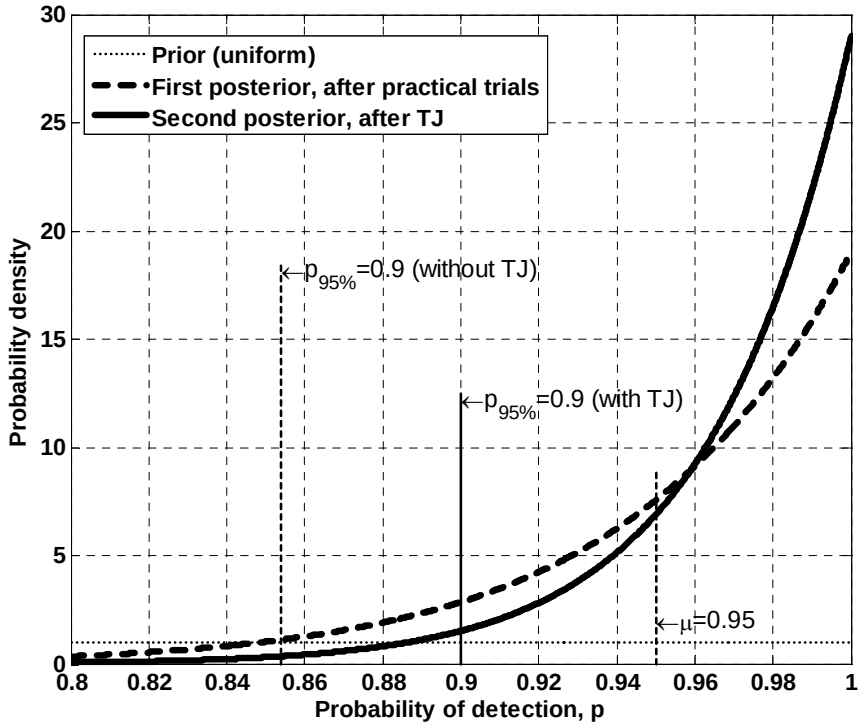

Figure 14 **Probability densities for example C1**

The probability densities are plotted in Figure 14. Again, the dashed line represents the posterior after updating with the practical trials only. The expected value is as high as required. The effect of the TJ updating is seen as “pushing” the 95% confidence lower bound value as high as required by the second target.

6 Discussion of some important issues

The Bayesian framework presented above is intended to be in principle a very straightforward method and one easy to use, although we fully recognise the difficulty of quantifying the TJ. The approach we have proposed is based on several simplifying assumptions, of which the user should be well aware. In this chapter we discuss and analyse some issues that may be of importance.

6.1 Independence of TJ and practical trials

The basic idea we have introduced relies on the assumption that the technical justification is seen as equivalent to a certain number of practical trials (and therefore, it is “translated” into an equivalent sample size and equivalent number of successes). This number of equivalent successes is then straightforwardly added to the successes in the practical trials (the real experiments) to draw conclusions regarding the probability of detection (using appropriate statistical tools of data reduction such as interval estimation, etc.).

One possible criticism is the issue of the independence of the TJ and the practical trials. In some circumstances, evidence gathered in the TJ is used to decide the number and type of defective components to be included in the practical trials. In other circumstances, evidence gathered carrying out a certain number of practical trials could influence the decision of which elements should be covered in the technical justification. It could then be argued that, in these situations, the TJ and practical trials are not truly independent. For instance, a conceptual error in the TJ could lead to a wrong choice of defective components for the practical trials. The error would thus “propagate” and be counted twice in the proposed modelling approach. By the same token, the individual equivalent trials constituting the TJ are clearly not independent Bernoulli trials.

The user of the proposed Bayesian approach must keep these considerations in mind. We repeat, as we have done throughout this report, that our method offers a simple way of breaking down and quantifying a series of expert decisions. The user must bear in mind that the final responsibility for such decisions ultimately lies in his or her hands.

6.2 Definition of defect population

In the model we proposed in [2], we considered a single, fixed flaw size and we assumed that all the input parameters (such as component type, material and acting damage mechanism) were defined and fixed. We did not specify any particular crack size, and we only generically stated that the procedure could be repeated for any crack size.

However, this definition may be too limited for practical application, and in the following we discuss the possibility of defining the population of defects whose probability of detection is under investigation in a more convenient way. In principle, our method offers much flexibility in this.

Usually, in a qualification exercise of an NDE system, all the attributes defining the problems are set as input parameters. For instance, the type and material of the components to be inspected (say butt welds in austenitic main steam lines), the acting damage mechanism(s) (say thermal fatigue), the defect attributes (say transgranular, inner-surface breaking cracks) are well

defined. The ENIQ methodology indeed requires that such input be specified prior to beginning the qualification.

It is quite natural to see all defects that (may) exist in such a setting as belonging to one more or less homogenous population. Let us now consider the defects parameters.

Traditionally, the probability of detection is modelled as being critically depended on crack size. Therefore:

$$POD=POD(a) \tag{69}$$

where a is one relevant dimension of the defect (for instance, its through-wall extent). This is a desirable concept, because fracture mechanics assessments of defect tolerance place a fundamental importance on such a relevant dimension. A structural engineer, for instance, may assess that a given structure (e.g. a butt weld) can tolerate (with good safety margins) a defect whose relative through-wall extent is 30% of wall thickness. The burden is then on the NDE engineer to demonstrate that the system chosen for inspection can indeed detect defects equal to or larger in size than this.

Whilst this is perfectly understandable, caution should be exercised in assuming that crack size, a, is indeed the most relevant parameter affecting crack detection. It could be that variations in other factors, such as crack tilt, or skew, or crack surface roughness, etc., are more important to crack detection than crack size. This should be taken into account when designing and manufacturing artificial defects meant to be representative.

Also, it is important to highlight the fact that the relationship expressed by Eq. (69) is (nearly) always modelled as a monotonic increasing function of crack size a. Whereas this is another very appealing assumption, it could not always be a justified one. When using an ultrasonic technique, for instance, it is quite intuitive that a bigger crack will reflect more acoustic energy than a smaller one, and therefore the transducers will, all other conditions being fixed, be more likely to detect it. Time of Flight Diffraction (TOFD) techniques, on the other end, rely on detecting the diffraction patterns of acoustic waves at crack tips. A bigger embedded crack (with a tip closer to the surface) could be less likely to be detected than a smaller crack.

In general, we propose that a working definition of defect population could include "*all defects of given attributes (i.e. the set of component, material, damage mechanism, morphology, etc. under investigation) whose size equals or exceeds a given size* $\bar{a}$". Within the ENIQ framework, it would then be natural to define $\bar{a}$ as the qualification size, a_c.

6.3 Representativeness of test block defects

One important issue is what conclusions can be drawn from the trial results. When test blocks are designed, the aim is often to manufacture defects that represent the most difficult defects to be found. In our framework, we have originally assumed the trial defects to be a representative sample from "real" defects in the components to be inspected.

Intuitively one would expect that actually the probability of detection (e.g. of defects of a certain size) is higher in the case where the defect population includes not only the most difficult defects (due to their tilt, or skew etc.), but also the whole range of possible defects. Thus, we could claim that finding 10 out of 10 test piece "worst case" defects actually results in higher confidence than what a statistical analysis (with the assumption of identical defect populations) indicates.

This intuitive notion can be formalised mathematically. Let us suppose that the following assumptions are made:

1. We subdivide the flaw population into two classes. Class 1 contains the flaws which are most difficult to detect, class 2 the remaining flaws.
2. The probability of detecting flaws belonging to each class is same for all cracks in the class. We call these two quantities p_1 and p_2.
3. The following test is carried out on flaws from class 1: the NDE system is applied to n defective components, and k are detected.
4. $p_2 \geq p_1$.

We again assume a Bayesian perspective. p_1 and p_2 are not unknown, fixed quantities, but random variables described by probability distributions. We therefore (1) assume that prior distributions can be chosen to represent the current knowledge about p_1 and p_2; (2) carry out an experiment to gather information and (3) use Bayes' theorem to update the prior distributions to obtain posterior distributions for p_1 and p_2.

We chose again a Beta distribution to represent p_1:

$$p_1 \sim \pi(p_1) = Be(\alpha, \beta) \tag{70}$$

We assumed that $p_2 \geq p_1$. If no other information is given, it is natural to specify the conditional distribution of p_2, given p_1, as a uniform distribution:

$$p_2 \mid p_1 \sim \pi(p_2 \mid p_1) = U(p_1, 1) \tag{71}$$

Now, the joint prior density is

$$\pi(p_1, p_2) = \frac{1}{B(\alpha, \beta)} p_1^{\alpha-1}(1-p_1)^{\beta-1} \times I_{\{p_2 \geq p_1\}} \times \frac{1}{1-p_1}, 0 \leq p_1 \leq 1, 0 \leq p_2 \leq 1 \tag{72}$$

where $B(\cdot,\cdot)$ is the beta function and $I_{\{\cdot\}}$ is the indicator function. If the test is modelled as a Bernoulli trial, the number of observed flaws follows the binomial distribution. Thus the probability of finding k flaws in n trials is:

$$P(K = k \mid n, p_1) = \binom{n}{k} p_1^k (1-p_1)^{n-k} \tag{73}$$

which is now the likelihood function. Applying Bayes theorem (the details are skipped for clarity), the joint posterior distribution is

$$\pi(p_1, p_2 \mid K = k) = \frac{1}{B(\alpha + k, \beta + n - k)} p_1^{\alpha+k-1}(1-p_1)^{\beta+n-k-1} \times I_{\{p_2 \geq p_1\}} \times \frac{1}{1-p_1} \tag{74}$$

It is possible to determine first the posterior for p_1 (since the likelihood function does not depend on p_2), and then simply use the conditional distribution of p_2, given by equation (71). The marginal posterior density of p_2 can be obtained by integrating the joint posterior density with respect to p_1:

$$\pi(p_2 \mid k, n) = \frac{\alpha + \beta + n - 1}{\beta + n - k - 1} \times F_B(p_2 \mid \alpha + k, \beta + n - k - 1) \tag{75}$$

in which $F_B(p_2 \mid \alpha + k, \beta + n - k - 1)$ is the cumulative $B(\alpha + k, \beta + n - k - 1)$-distribution at point p_2.

Let us illustrate this model with an example. Let us suppose that we carry out n=10 practical trials, on a set of defective components that we judge are truly representative of worst-case conditions. p_1 represents the probability of detection of defects in this "more difficult" class. p_2 represents the probability of detection of the remaining defects. As stated above, we chose a

uniform prior distribution for p_1, equation (70), and we assume that, conditional on p_1, p_2 is uniformly distributed between 0 and p_1, equation (71). These prior distributions are shown in Figure 15 with black lines. Note that the unconditional distribution for p_2 is not uniform (dotted black line).

The NDE system is then applied to the defective components, and the outcome is k=10 successes, that is all flaws are found. The equations above permit us to calculate the posterior distributions for p_1 and p_2. These are also plotted in Figure 15 (red lines). The prior and posterior cumulative distributions (as usual, in terms of 1-F curves) are plotted in Figure 16.

It is interesting to analyse these latter curves in more details. In Figure 17, the prior and posterior 1-F curves for p_1 and p_2 are plotted for values of p ranging between 0.8 and 1. To this graph, a 1-F curve for a "traditional" case (i.e. a case in which only one class of defects is identified, as has been assumed throughout this report) has been added. The (blue dotted) curve was obtained using a uniform prior updated to reflect the outcome N_s=N=20. Let us call p_t the probability of detection of such case. The blue dotted curve thus represents 1 minus the cumulative Beta distribution function obtained as a posterior.

It is clear to see that the posteriors for p_t and for p_2 are rather similar. In other words, under the assumptions above, it can be argued that finding 10 out of 10 of the "more difficult" defects gives the same confidence in the NDE system under examination that the finding of 20 out of 20 "average" defects.

The example above was based on a minimal assumption regarding the two defect populations, i.e. that the probability of detection is higher for a population consisting of "all kind of defects" as compared with the population representing difficult test piece defects. This assumption didn't say anything about how much higher the POD would be. Such assumptions would increase the complexity of the model and, above all, increase the need to credibly justify the use of any numbers. At this stage, we exclude the development of such models from our studies.

It should also be borne in mind that the inspections performed in the field might include human factors (such as stress, fatigue, etc.) which may have an adverse effect on the POD, thus compensating to some extent for the above considered conservatism.

6.4 Choice of the prior distribution

In our approach and throughout all our examples we have used a uniform distribution as our prior distribution. We have justified its use on the one hand by its mathematical convenience – the uniform distribution is a special case of the Beta distribution, which in turn is the conjugate distribution of the binomial distribution (see Appendix 1). On the other hand, the uniform distribution is a so-called non-informative prior for the parameter of the binomial distribution. A non-informative prior expresses only vague or general information about a variable – in our case that the variable is equally likely to have a value anywhere between 0 and 1. The use of the uniform distribution implies that we want to include as little information as possible in the prior, and give all the weight to the documented information in the TJ and trial results.

The choice of a prior distribution is often a subject of debate, and it is worth discussing here alternative choices of the prior distribution.

It could be justified to use an informative prior, since some inspections are more difficult than others, and the NDT qualification experts do have some information about the detectability of flaws, given for instance the material to be inspected or the particular technique being employed.

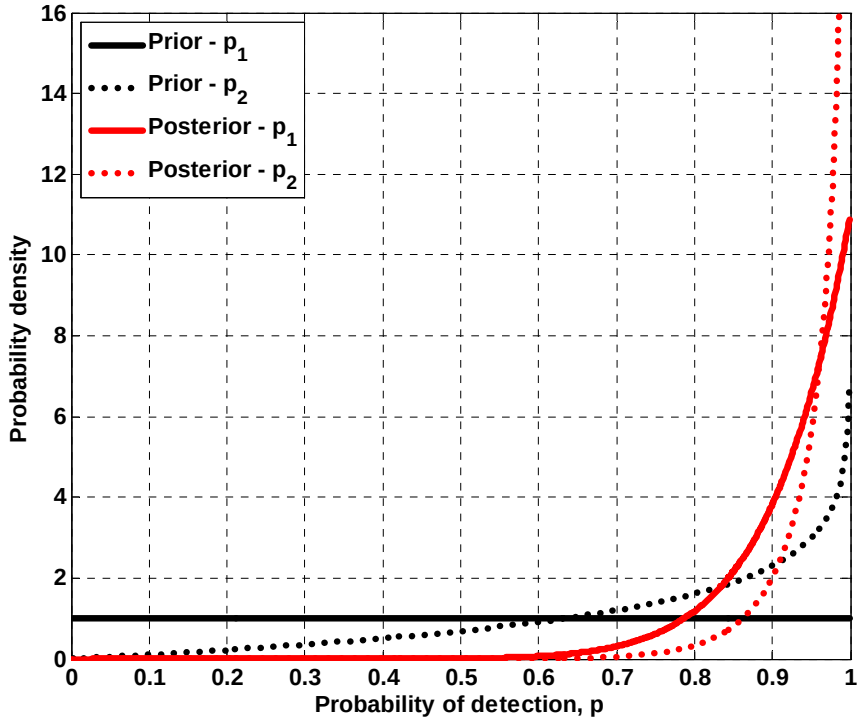

Figure 15 **Probability densities for p_1 and p_2**

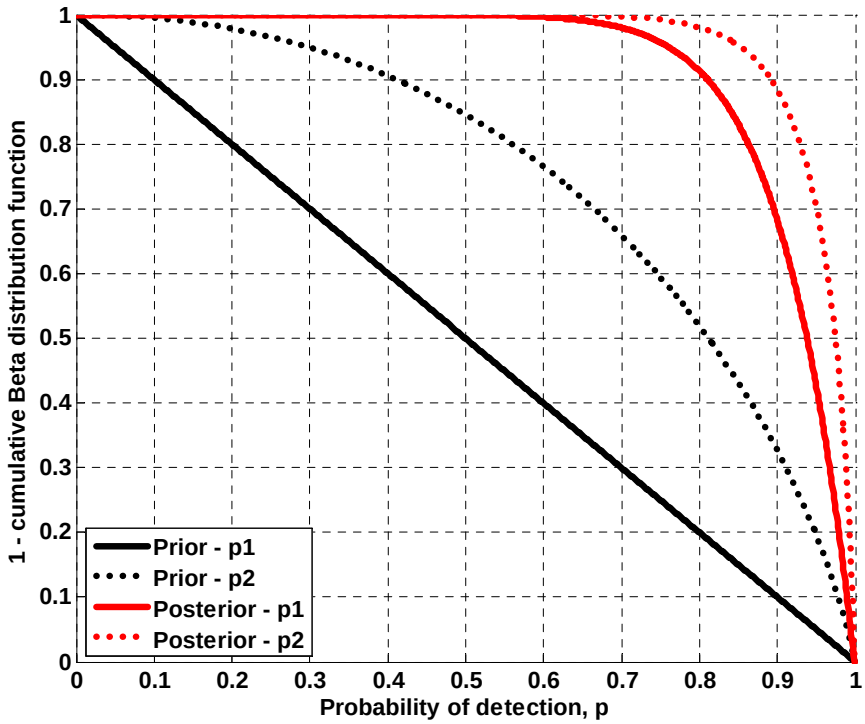

Figure 16 **1-F curves (with F the cumulative Beta distribution function) for p_1 and p_2**

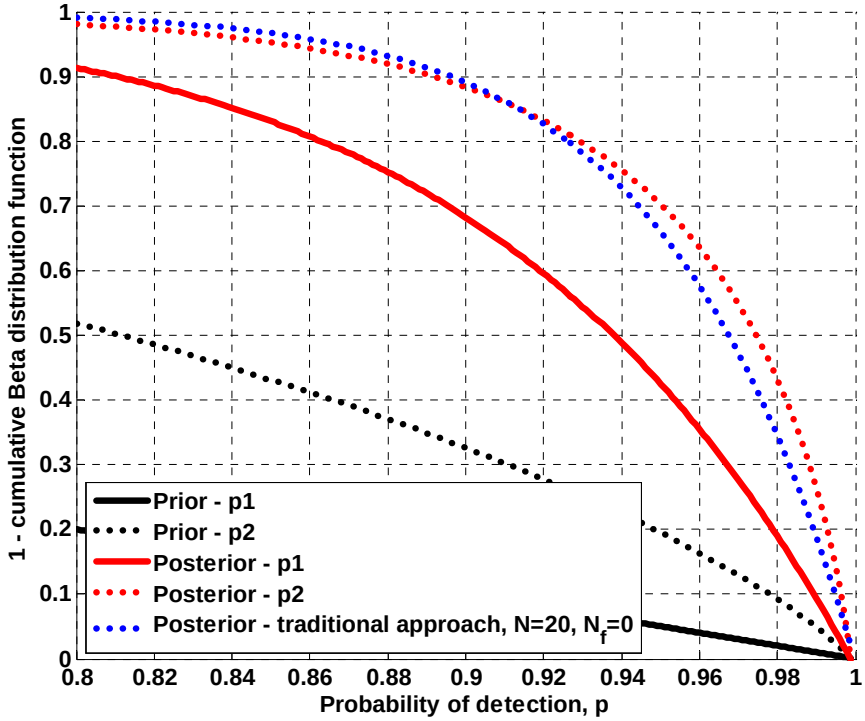

Figure 17 **1-F curves (with F the cumulative Beta distribution function) for p_1 and p_2. The 1-F curve for a simple case (only one class of defects, like**

There is however a problem related to the use of an informative prior. In the Bayesian updating, the posterior distribution is obtained by updating the prior distribution with new knowledge. For the qualification expert it may be impossible to identify exactly on which information this prior knowledge is based, and thus it may be impossible to extract and exclude the same information from the evaluation of the TJ. For instance, the TJ often contains a section on experimental evidence, providing information on results from other relevant qualifications, experimental studies and field experience. It is very likely that the prior knowledge of the qualification expert is based on this experience. If such information has influenced the choice of a prior distribution, and is then evaluated again in the TJ scoring, it will result in being counted twice.

Our starting point is that the TJ is a document where all evidence on the effectiveness of the test is assembled [RP2]. Thus we choose to use the non-informative prior and assume that all relevant information will be taken into account through the TJ. We believe that systematically analysing and evaluating the TJ's various elements is the most transparent and accountable approach to quantify the TJ. Using the quantified TJ to update the non-informative prior provides the most plausible posterior distribution, which is further updated with the results from practical trials.

7 Conclusions

In this report, we have discussed at length the Bayesian framework for quantifying the ENIQ inspection qualification methodology we proposed in [2]. Using this approach, we have shown that it is possible to combine the evidence gathered in the technical justification with the data from trials. The approach allows structuring and quantifying the qualification process, so that the combination of experimental evidence and capability judgement has a more solid and transparent basis. Further, it provides a quantitative measure of the inspection capability even with a limited number of practical trials.

The approach is based on the principle that the evidence from a technical justification is quantified and expressed as parameters of a Beta-binomial distribution. This quantified judgement of the confidence in the TJ is combined with results from practical trials, and the resulting confidence in the NDT system qualification is compared to the preset performance target. The performance target should not be expressed simply as a probability of detection, but the uncertainties should be accounted for by defining a lower bound probability of detection and associated confidence level.

We have illustrated, with the aid of several examples, how the approach could be used in practice. At this stage we have not developed more detailed guidelines for practical application. The next step planned is to apply the model to a real case, where a dummy qualification body will be set up and the feasibility of the approach investigated. The ultimate aim is to develop more detailed rules for the quantification of the TJ in order to define the equivalence between the overall quality of an individual TJ and the corresponding number of equivalent trials.

8 Acknowledgements

We would like to thank Prof. Urho Pulkkinen of VTT for developing the Bayesian model presented in section 6.3 and Mr. Vic Chapman for fruitful discussions. We would also like to acknowledge Dr. A. Eriksson of JRC, Dr. B. Shepherd of Mitsui Babcock and Dr. R. Chapman of British Energy for many constructive comments. Finally, we acknowledge the comments we received from Forsmark Kraftgrupp.

9 References

[1] European methodology for qualification of non-destructive testing: second issue. EUR 17299 EN. 1997.

[2] Gandossi, L. and Simola, K., Framework for the quantitative modelling of the European methodology for qualification of non-destructive testing, International Journal of Pressure Vessels and piping 82 (2005) 814-824.

[3] ENIQ recommended practice 3: strategy document for technical justification. ENIQ Report No. 5, JRC-Petten, EUR 18100/EN; 1998.

[4] Berens, A.P., NDE reliability data analysis. In Metals Handbook, 9th edn, Vol. 17, ASM Int., 1989, pp. 689-701.

[5] ENIQ recommended practice 2: recommended contents for a technical justification. ENIQ Report No. 4, JRC-Petten, EUR 18099/EN; 1998.

Appendix 1 Estimation of a population parameter in classical and Bayesian statistics.

Classical statistics

According to the classical (frequentistic) theory of probability, the probability of an event is the limit of the percentage of times that the event occurs in repeated, independent trials under essentially the same circumstances. *p* is a parameter whose value is unknown and cannot be measured directly, unless an infinite number of trials can be arranged.

An approximation of *p* can be determined by carrying out trials on a reduced number of defects introduced in test pieces. In statistical terms, this means taking a sample of the relevant quantity and inferring some information about the true characteristic of the whole population. If this is done, and the experiment yields N_s successes over a total number *N* trials, it is very natural to approximate *p* with the ratio N_s/N.

$$\hat{p} = \frac{N_s}{N} \tag{A1}$$

It is important to highlight the fact that $\hat{p}$ (p-hat) is only an approximation of *p*. For this reason, it is called a *point estimate of p*. $\hat{p}$ is binomially distributed. This is because the number of successes, N_s, is binomially distributed. It is easy to see that if we have *N* experiments, each one with success probability *p*, the probability of having exactly N_s successes will be:

$$p(successes = N_s) = \left(\frac{N_s}{N}\right) p^{N_s} (1-p)^{N-N_s} \tag{A2}$$

$\hat{p}$ is the best estimate of *p* we have after carrying out the set of trials, but alone it does not tell us much about the true population proportion, unless *N* is large. To see this with an example, let us consider the limit case in which *N*=1. Then we will have either $\hat{p}$=0 (the single trial was a failure) or $\hat{p}$=1 (the single trial was a success), but this does not tell us much about *p*.

This is reason why the classical statistical framework envisages the use of interval estimators. The purpose of using an interval estimator, rather then a point estimator such as $\hat{p}$, is to have some quantitative guarantee of capturing the parameter of interest. The sacrifice of some precision in the estimate, moving from a point estimator to an interval estimator, results in an increased confidence that the assertion is correct.

Formally, an interval estimate of a real-valued parameter p is any pair of functions L(p), U(p) that satisfy $L(p) \leq U(p)$ for all *p*. If $\hat{p}$ is observed, the inference $L(\hat{p}) \leq p \leq U(\hat{p})$ is made. The interval [*L(p), U(p)*] is called an *interval estimator*. For an interval estimator [*L(p), U(p)*] of a parameter p, the *coverage probability* of [*L(p), U(p)*] is the probability that the random interval [*L(p), U(p)*] covers the true parameter p. For an interval estimator [*L(p), U(p)*] of a parameter p, the *confidence coefficient* of [*L(p), U(p)*]is the infimum of the coverage probabilities.

It is often of interest to find a lower bound confidence interval. For a proportion, *p*, this means determining an interval of the form [*L(p), 1*] for the problem at hand. When $\hat{p}$ is observed, the inference [*L(p), 1*] is made.

It is interesting to note that in classical statistics, it is said that the interval ***covers*** the parameter, not that the parameter is ***inside*** the interval. This is because, strictly speaking, the random quantity is the interval, NOT the parameter. It is tempting to say, and many experimenters actually do, that “the probability is 95% that *p* is in the interval [.9, 1]. Within classical statistics

this statement is invalid since the parameter is assumed fixed. Formally, the interval [0.9, 1] is one of the realised values of the random interval [*L(p), 1*]. The realisation occurred upon observing $\hat{p}$=1. Since the true unknown parameter *p* does not move, p is in the realised interval [.9, 1] with probability either 0 or 1. When we say that the realised interval [.9, 1] has a 95% chance of coverage, we only mean that we know that 90% of the sample points of the random interval cover the true parameter.

We will see in the following section that in the Bayesian framework the construction and interpretation of confidence intervals is more straightforward and intuitive than in the classical framework. Confidence intervals are actually called *credible sets* in the Bayesian framework, to underline the fact that their interpretation is conceptually different from classical confidence intervals.

Bayesian statistics

The Bayesian approach to statistics is fundamentally different from the classical one [ref. A1]. We have seen how in the classical approach, the parameter *p* is thought to be an unknown, but fixed, quantity. A random sample is drawn from a population and, based on the observed sample, knowledge about the value of *p* is obtained (for instance, $\hat{p}$).

In the Bayesian approach, *p* is considered to be a quantity whose variation can be described by a probability distribution. This is a subjective distribution, based on the experimenter's knowledge. Before carrying out any experiment, the experimenter can attribute a *prior distribution* to *p*. This expresses all the knowledge (or lack of it) of the experimenter before undertaking the experiment. The experiment (a set of trials) is then performed and the prior distribution is updated using this information. A new distribution (the *posterior distribution*) is thus obtained. This updating is done using Bayes' rule, and hence the name Bayesian statistics.

Bayes' rule for discrete probabilities takes the following form:

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)} \tag{A3}$$

whereas Bayes' rule for probability distributions can be written as

$$\pi(\theta \mid x) = \frac{f(x \mid \theta) \cdot \pi(\theta)}{\int f(x \mid \theta) \cdot \pi(\theta) d\theta} \tag{A4}$$

where $\pi(\theta)$ denotes the prior distribution, $\pi(\theta|x)$ is the posterior distribution (the distribution of θ given the data x) and f(x|θ) is the sampling distribution.

After the experimental evidence has been gathered, the posterior distribution is calculated using Equation A4.

In general, the integral in the denominator of the right hand side of Equation A4 must be calculated numerically. In some special cases, the posterior can be obtained in closed form (i.e. the integral can be solved analytically). This happens for particular choices of the sampling and prior distribution. Such distributions families are called *conjugate*. For instance, the Beta family is conjugate for the Binomial family. In other words, if the sampling distribution is binomial and we choose a beta prior, we will obtain a beta posterior in closed form.

When adopting the Bayesian framework, we still assume that that the results of test piece trials are seen as a sample from a Binomial distribution. Now the parameter *p* of the Binomial distribution is considered as an unknown variable, and the uncertainty related to this variable is

expressed with a probability distribution. As we have seen, the natural conjugate distribution of the Binomial distribution is the Beta distribution, and thus it is a natural choice for the form of the distribution of *p*.

Thus, we assume a Beta prior distribution for p:

$$p \sim Beta\ (\alpha,\beta) \tag{A5}$$

where

$$Beta(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1-p)^{\beta-1}\,p^{\alpha-1} \tag{A6}$$

and Γ is the gamma function. The parameters α and β determine the shape of the Beta distribution. The distribution is defined for α>0 and β>0. If α>1 and β>1, the Beta distribution is unimodal. In this case, the expected value, variance, and mode of the distribution are given by

$$\begin{aligned} E(p) &= \frac{\alpha}{\alpha+\beta} \\ Var(p) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ Mode(p) &= \frac{\alpha-1}{\alpha+\beta-2} \end{aligned} \tag{A7}$$

In Figure A1 some distributions from the Beta family are plotted. The uniform distribution is a special case of the Beta distribution when $\alpha = \beta = 1$.

We are now required to choose some values for α and β that reflect our prior knowledge about *p*. Very likely, we may wish to express a prior "ignorance". We do not have any idea about the possible values *p* may take, so a reasonable choice may be to take $\alpha = \beta = 1$ and thus work with a uniform prior. We may have instead some reasons to believe that *p* is in the whereabouts of 0.5, and therefore we may chose $\alpha = \beta = 5$, or even $\alpha = \beta = 50$. Observe that the more we increase the values of α and β, the more the distribution becomes narrow, reflecting a reduced uncertainty about *p*.

After forming the prior, we proceed to carry out practical trials as in the classical framework. Again, we record N_s successes out of *N* trials.

It can be easily shown that the posterior distribution for *p* is now

$$p \sim Beta\ (\alpha+N_s,\beta+N-N_s). \tag{A8}$$

In other words, the new parameter α of the Beta posterior is equal to the old parameter α increased by the number of hits, N_s, whereas the new parameter β of the Beta posterior is equal to the old parameter β increased by the number of failures, (N-N_s).

$$\begin{aligned} \alpha_{posterior} &= \alpha_{prior} + N_s \\ \beta_{posterior} &= \beta_{prior} + N_f = \beta_{prior} + N - N_s \end{aligned} \tag{A9}$$

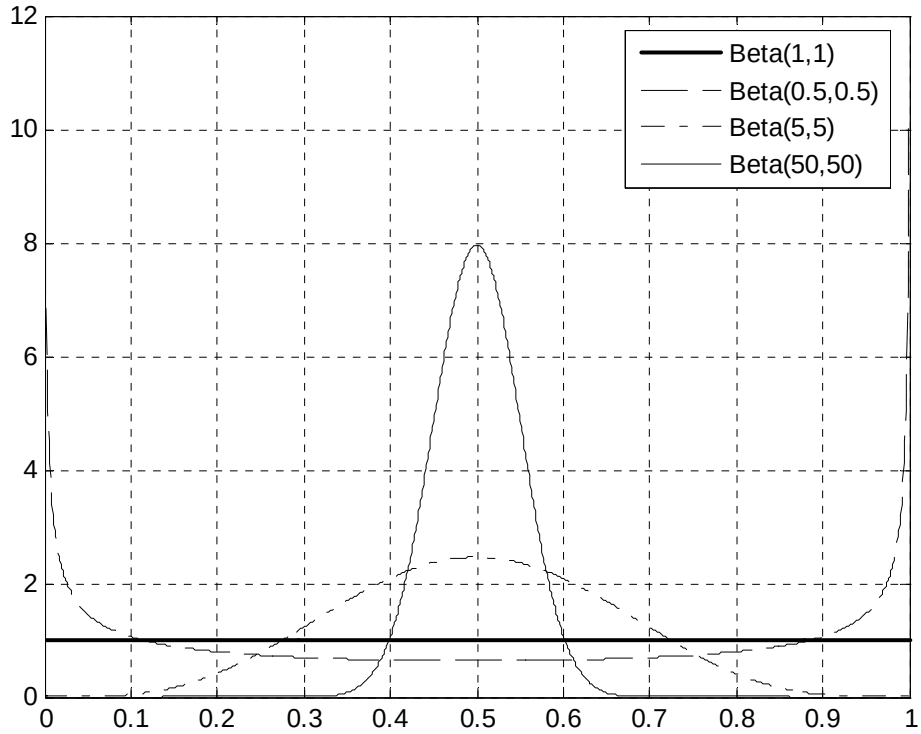

Figure A1 Distributions from the Beta family ($\alpha = \beta$)

The posterior distribution for *p* fully describes what we now know about *p*. Different Bayesian point estimates for *p* can simply be obtained considering the mean, the mode or the median of the posterior distribution. We are more interested in considering interval estimators, which in the Bayesian framework take the name of *credible sets*. The knowledge of the posterior distribution of *p* makes straightforward the calculation of such intervals.

A one-sided 100δ% (with 0<δ<1) lower bound credible set is simply defined by that number k for which the following relationship holds:

$$\int_{k}^{1} Beta(x,\alpha,\beta)dx = \delta \qquad \text{(A10)}$$

In Figure A2, the meaning of this equation is illustrated. The probability distribution function for Beta(5,1) is represented. The one-sided lower bound set is defined by the shaded region. Thus, k is found so that the area under the distribution function between k and 1 is exactly δ. We can thus say that we are 100δ% certain that the parameter of interest, *p*, is greater than k, and in this sense k provides a 100δ%-lower bound estimate for *p*.

We can indicate such a one-sided interval as

$$[p_{100\delta\%}, 1] \qquad \text{(A11)}$$

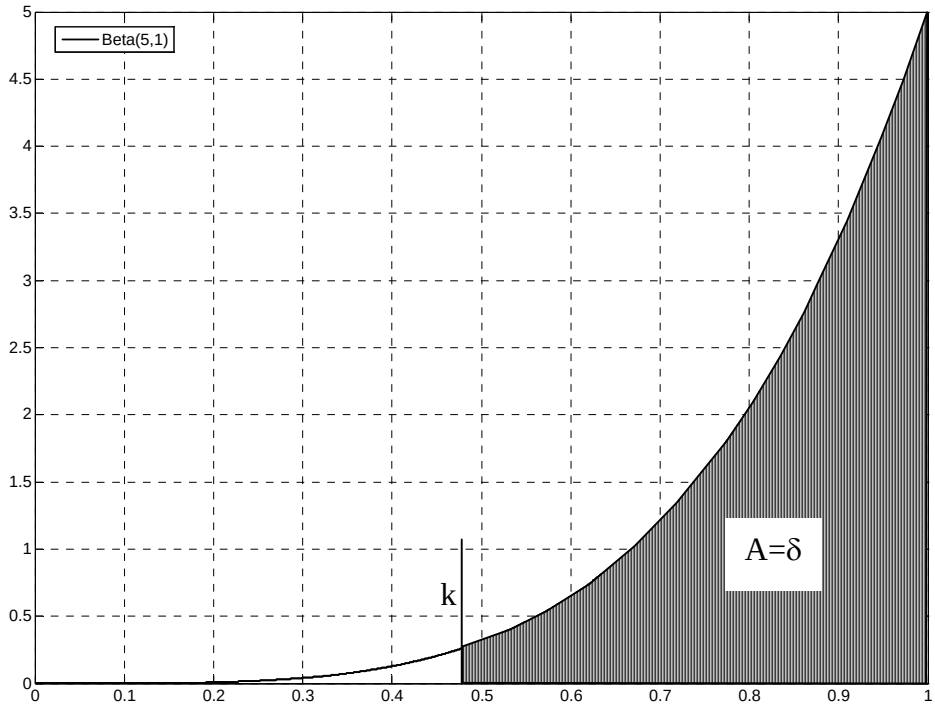

Figure A2 Beta(5,1) distribution and example of credible set

It can be easily shown that classical and Bayesian statistics yield very similar (but not equal) results (say for instance when calculating a confidence interval and the equivalent credible set) when a uniform prior ($p \sim$ Beta(1,1)) is chosen.

δ can also be expressed in the following way:

$$\delta = \int_{k}^{1} Beta(x,\alpha,\beta)dx = 1 - \int_{0}^{k} Beta(x,\alpha,\beta)dx = 1 - F(x,\alpha,\beta) \qquad \text{(A12)}$$

where $F(x,\alpha,\beta)$ is the cumulative Beta distribution function. This equation can be used in different ways. Straightforwardly, for a given x=*p* and given parameters α and β, the associated confidence level δ can be found. Alternatively, both x and δ can be specified and different combinations of the parameters α and β can be determined so that equation (A12) is verified. In Figure A3 are plotted some examples of the function 1-$F(x,\alpha,\beta)$.

Let us suppose for instance that somehow we know intuitively that the population proportion, *p*, is very high, but we want to prove this in a rigorous statistical way. For instance, let us suppose that we have a machine that manufactures a tool. The machine is very sophisticated, but we do not exclude the possibility that the manufactured tool could be defective. We call *p* the long-term proportion of tools that are satisfactorily manufactured by the machine, i.e. tools that are not defective. We want to prove that, with 99% confidence, p is greater than 95%. To be scrupulous, even if we have been operating the machine before and know something about its capabilities, we now do not make any assumptions regarding p. A reasonable – and conservative – starting point is to assume that p is uniformly distributed in the interval [0, 1].

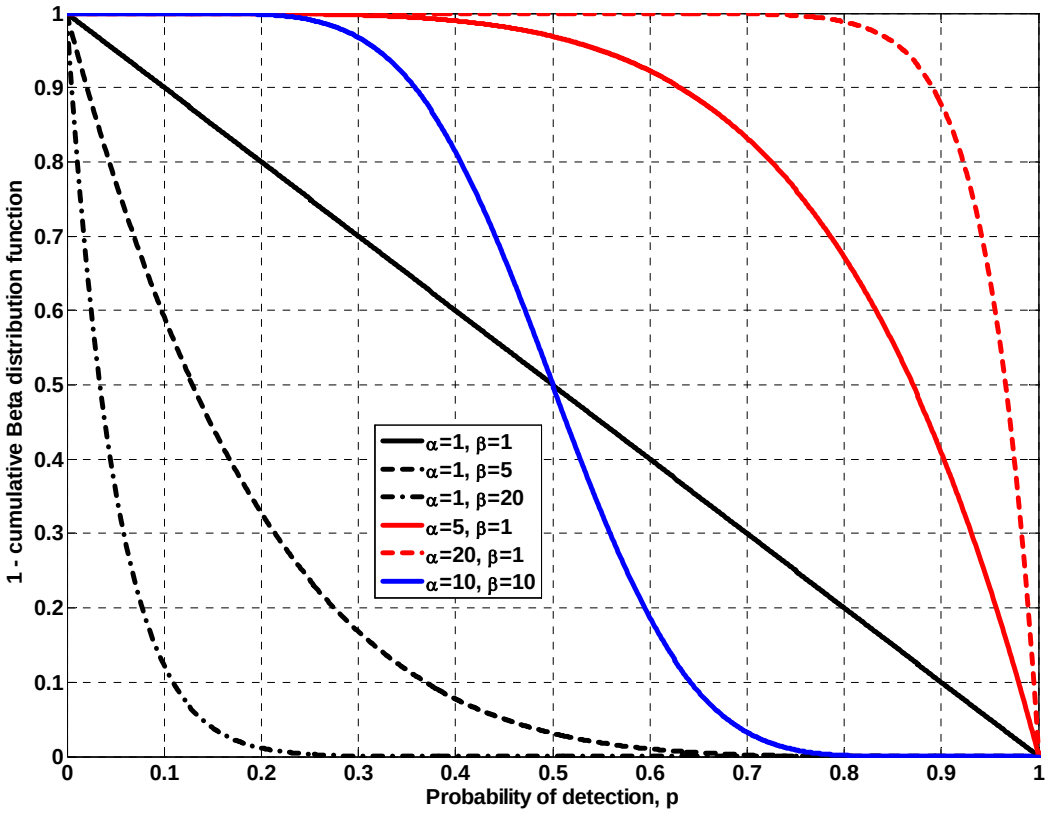

Figure A3 Some examples of the function 1-*F(x,α,β)* for various choices of the parameters *α* and *β*.

Thus, the prior distribution for p is assumed to be

$$p \sim Beta\ (1,1). \tag{A13}$$

If we were now to carry out this experiment: "manufacture *N* tools, and count the number of non-defective tools, N_s", we would easily obtain a posterior distribution such as

$$p \sim Beta(1+N_s, 1+N-N_s). \tag{A14}$$

Equation (A12) would then be rewritten as

$$\delta = 1 - F(x, 1+N_s, 1+N-N_s) \tag{A15}$$

And in our example, this would equate to finding those values of *N* and N_s satisfying

$$0.99 = 1 - F(0.95, 1+N_s, 1+N-N_s) \tag{A16}$$

Actually, all those values of *N* and N_s satisfying the following inequality would also be acceptable, because the confidence interval [$p_{100\delta\%}$, 1] would be narrower (i.e. $p_{100\delta\%}$ closer to 1) than what is prescribed:

$$\delta > 1 - F(x, 1+N_s, 1+N-N_s) \tag{A17}$$

In other words, all choices of (N, N_s) yielding curves 1-$F(x,1+N_s,1+N-N_s)$ passing through or above the point (0.95, 0.99) would be acceptable, see Figure A4. Clearly, in a real situation, one would be able to choose N, but not N_s.

Figure A4 has been plotted for the example at hand. The curves plotted are the functions 1-$F(x,\alpha,\beta)$, where F is the cumulative Beta distribution function. The abscissa, x, therefore represents the population proportion, p, whereas the ordinate represents the associated confidence level, δ. If we want our experiment {N, N_s} to prove the target ($p_{100\delta\%}$, δ), the values N, N_s must be such that the curve 1-F$(x,1+N_s,1+N-N_s)$ passes through or above the point ($p_{100\delta\%}$, δ), i.e. must intercept the red rectangle of Figure A4.

For the example at hand, it turns out that the smallest sample size required to prove the target is N=89. In this case, α=90 and β=1 (remember, we are starting from a uniform prior). We thus need 89 out of 89 successes to prove the target. Next, allowing for a single failure, we determine that we require a sample size N=129. In this second case α=129 and β=2, and we thus need 128 successes out of 129 trials. It is clear that every single failure requires a considerable number of successes to "re-establish" proof of the target.

If we are reasonably certain that p is very high, we can decide to conduct the experiment using the smallest sample size, N=89 in this case. Obviously, if a single failure is recorded the target will not be proven (but then again, p was maybe not as high as we previously thought).

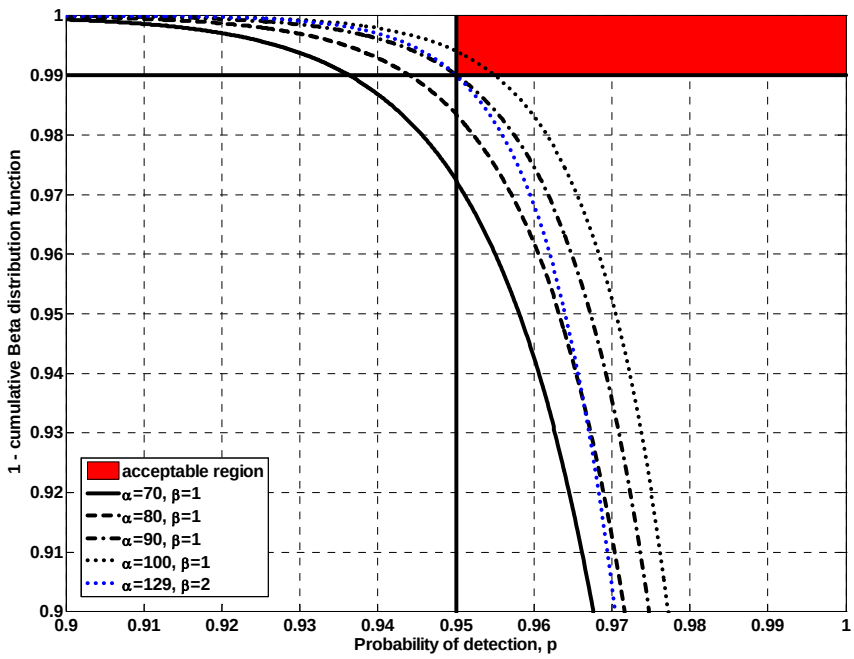

Figure A4 Some examples of the function 1-$F(x,\alpha,\beta)$

References

[A1] Bernardo, J.M., Smith, A.F.M., Bayesian Theory, John Wiley & Sons, 2000.

European Commission

EUR 22675 EN – DG JRC – Institute for Energy

A BAYESIAN FRAMEWORK FOR THE QUANTITATIVE MODELLING OF THE ENIQ METHODOLOGY FOR QUALIFICATION OF NON-DESTRUCTIVE TESTING

Authors

Luca Gandossi	DG-JRC-IE
Kaisa Simola	VTT Technical Research Centre of Finland

Luxembourg: Office for Official Publications of the European Communities
2007 – 22 pp. – 21 x 29.7 cm
EUR - Scientific and Technical Research Series; ISSN 1018-5593

Abstract

The European methodology for qualification of non-destructive testing, produced by the European Network for Inspection and Qualification (ENIQ), has been adopted as the basis of inspection qualifications for nuclear utilities in many European countries. According to this methodology, the inspection qualification is based on a combination of technical justification (TJ) and practical trials. The methodology is qualitative in nature, and it does not give explicit guidance on how the evidence from the technical justification and results from trials should be weighted. Recently, we have proposed a quantified approach to combine evidence from technical justifications and practical trials. A Bayesian statistical framework for the quantification process was introduced, and some examples of possibilities to combine technical justification and trial results were given. The underlying idea was to improve transparency in the qualification process, whilst producing at the same time estimates of probability of detection that could for instance be used in structural reliability evaluation and Risk-Informed In-Service Inspection. In this report, we attempt to give a more detailed description of the approach and some guidelines regarding how a user (utility, qualification body, etc.) could tackle the problem of quantifying the outcome of a qualification exercise in practical terms.

The mission of the Joint Research Centre is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

JRC
EUROPEAN COMMISSION

Publications Office
Publications.eu.int