NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Technical Report 32-1165

Reliability-Confidence Combinations for Small-Sample Tests of Aerospace Ordnance Items

A. G. Benedict

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

October 15, 1967

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Technical Report 32-1165

Reliability-Confidence Combinations for Small-Sample Tests of Aerospace Ordnance Items

A. G. Benedict

Approved by:

W. Gin, *Manager*
Solid Propellant Engineering Section

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

October 15, 1967

TECHNICAL REPORT 32-1165

Prepared Under Contract No. NAS 7-100
National Aeronautics & Space Administration

Contents

Tables

Abstract

Analyses epitomized by the widely-used equation

$$\gamma = 1 - R^n$$

are shown to provide unexpectedly reasonable approximations for small-sample tests of aerospace ordnance items; high-reliability, low-confidence combinations are discussed.

Reliability-Confidence Combinations for Small-Sample Tests of Aerospace Ordnance Items

I. Introduction

For a complex aerospace system to have even a 50% chance of performing its intended function, the chances of failure in any one critical subsystem component must be extremely low.

In the evaluation of the reliability of aerospace ordnance components, it is common to find that

(1) The items are expensive, limiting the number of samples available.

(2) The items are of a special design, with no past history to suggest their reliability.

(3) Because the target reliability approaches unity, evaluation should be directed at fail rates rather than reliability.[1]

(4) The items to be tested are "single-shot" and, unlike solenoid relays for example, cannot be operated nondestructively; items operated in test cannot be used in flight.

(5) Tests by both attributes and variables may be appropriate, but the frequency distribution for tests by variables is unknown.[2]

(6) There will be some uncertainty (if not confusion) in the selection of a reliability-confidence combination that best expresses the results of the evaluation.

Today, evaluation of small-sample aerospace ordnance tests follows along lines developed in the midtwenties by

[1]A decrease of less than 1% in reliability, from 99.9% to 99%, would represent a tenfold increase in fail rate (i.e., from 1/1000 to 1/100).

[2]Log-normal or comparable distributions are often assumed simply as a convenient economy. Although actual test results involving only small samples may prove to be consistent with some arbitrary distribution, use of such a preselected distribution for extrapolation to extreme percentiles may be grossly misleading.

Western Electric and Bell Telephone engineers[3] as characterized by the equation

$$\sum_{j=0}^{j=F} \binom{n}{j} f^j (1-f)^{n-j} = \sum_{j=0}^{j=F} \binom{n}{j} (1-R)^j R^{(n-j)} = (1-\gamma) \tag{1}$$

where

n = number of pass/fail tests

F = number of fails detected

$$\binom{n}{j} = \frac{n!}{(n-j)!\,j!}$$

R = assumed reliability

f = assumed fail rate

γ = lower confidence limit on R (or upper confidence limit on f)

Equation (1) reduces, for a series of n tests with no failures,[4] to

$$(1-f)^n = R^n = 1-\gamma \tag{2}$$

Both equations allow the results of any particular test program to be expressed as indicating any of an infinite number of reliability-confidence combinations. For example, a five-sample, no-fail test could be interpreted as indicating any one of as many combinations as one wished, including those shown in Table 1, as calculated from Eq. (2).

The question naturally arises as to which combination, if any, is the most descriptive. For example, if the five-sample test were used to judge the quality of a further six items drawn from the same lot, would the 80% reliability-67% confidence combination imply that, of the six samples, at least five (80%), or four (67%), or three (80% of 67%), should be expected to pass, or should none be expected to pass because none are 100% reliable?[5]

Table 1. Some reliability-confidence combinations for a five-sample no-fail test[a]

Reliability R, %	Confidence γ, %	Maximum fail rate f, %
99	5	1
90	41	10
87	50	13
80	67	20
70	83	30
60	92	40
50	97	50
40	99	60

[a]Note the rapid rise in confidence for a decrease in reliability from 99% to 90%. It may be shown that, for $f \approx 0$, $\log_{10}(1-\gamma) \approx -0.434\,nf$.

Before proceeding, it is important to note the original problem faced by the Western Electric and Bell Telephone engineers: how to use small samples to determine whether a particular incoming shipment of components met the level of quality shown by prior shipments from the manufacturer. If earlier shipments had exhibited a reliability of 98% or better, with an occasional bad lot having a reliability as low, for example, as 80%, the customer might be content with a small-sample receiving inspection scheme that gave him a good chance of detecting lots with a reliability of less than 98%. By comparison, many aerospace ordnance items involve short-run, one-time production offering no prior history on which to base an expected reliability; this raises the further question of whether Eq. (2) is grossly inappropriate in such cases.

II. Inherent Reliability of "No-History" Lots

At first glance, it might seem that evaluation of a lot for which all fail rates between 0% and 100% were equally likely would be a much more pessimistic process than evaluation of a lot for which fail rates below, say, 70% were assumed to be somewhat unlikely. This is not so, however, because small-sample, no-fail tests quickly cull (or "screen") high fail-rate lots.

[3]Dodge, H. F., and Romig, H. G., *Sampling Inspection Tables, Single and Double Sampling*, Second Edition, p. 1. John Wiley & Sons, Inc., New York, 1959.

[4]For simplicity, subsequent discussion is limited to no-fail tests; these are of particular interest in aerospace ordnance because the number of samples available for test is usually so small that even one failure may imply an intolerably high fail rate.

[5]Although Eqs. (1) and (2) relate to tests by attributes, reliability-confidence combinations are also used to express the outcome of tests by variables, giving rise to the same question.

If an n-sample test is made on a large number K of lots of size L for which all proportions p of passes P's between 0 and 1 are equally likely, the chances C of any lot yielding a sample of n P's is given by[6]

$$C = p^n \tag{3}$$

and the number of lots M that will yield such samples is given by

$$M = \int_0^1 KC\,dp = \int_0^1 Kp^n\,dp$$

from which

$$M = \frac{K}{n+1} \tag{4}$$

The proportion R_i of P's in these M lots representing their average quality or "inherent reliability" is given by

$$R_i = \frac{1}{ML}\int_0^1 KL\,p^n\,p\,dp = \frac{KL}{ML} \times \frac{1}{n+2} = \frac{n+1}{n+2} \tag{5}$$

and the proportion γ of the M lots that will have at least any proportion R of P's is given by

$$\gamma = (n+1)\int_R^1 p^n\,dp$$

from which

$$\gamma = 1 - R^{n+1} \tag{6}$$

Note the similarity between Eq. 6 and Eq. 2. Equation 6 shows that an isolated lot yielding no fails in a sample of size n will have (with confidence γ) a comparatively high reliability even if, prior to sampling, all reliabilities between 0 and 1 were considered equally likely. It can be shown that although the inherent reliability approaches 1 as n approaches ∞, the corresponding γ approaches only the limit $1 - e^{-1}$ (=63.22%).

III. Confidence and Unexpected Fail Rates

Although a proportion $\gamma = 1 - R^{n+1}$ of lots yielding no fails in samples of size n will contain a proportion $(1 - R) = 1/(n+2)$ or less of faulty items, the remaining fraction R^{n+1} of the lots will contain a proportion of faulty items larger than $1/(n+2)$; from these lots will come unexpectedly high and possibly disappointing failure rates. For example, from an eight-sample, no-fail draw, the probability of various fail rates can be illustrated as in Table 2.

Table 2. Approximate chances of various fail rates for a lot yielding no fails in eight samples

Reliability R[a]	Confidence γ[b]	Most likely maximum fail rate $(1 - R)$	Approximate chances of indicated fail rate[c]
1	0	0	0
0.9	0.613	0.1	(0.613 − 0) = 0.61
0.8	0.864	0.2	(0.864 − 0.613) = 0.25
0.7	0.959	0.3	(0.959 − 0.864) = 0.10
0.6	0.9897	0.4	(0.9897 − 0.959) = 0.03
0.5	0.9980	0.5	(0.9980 − 0.9897) = 0.008
0.4	0.9997	0.6	(0.9997 − 0.9980) = 0.002

[a]The "inherent" reliability of Eq. 5 is 0.9.
[b]Calculated from Eq. 6.
[c]These chances relate strictly to ranges of fail rates, but are more easily visualized as relating to fails in a lot of size 10. The significance of this column can be illustrated by considering 10 rafts, each supported by 10 empty oil drums; if the rafts are made from drums for which a reliability of 90% has been established at a confidence of 60%, and if, with more than 1 drum leaking, any of the rafts sink, it would be optimistic to expect more than 6 rafts to float.

From Table 2, an expectation that the fail rate would be 1/10 or less would result in a 25% chance of being disappointed by a factor of 2, and a 10% chance of being disappointed by a factor of 3.

It is obvious that large discrepancies in fail rates will be common unless the chances of excessive fail rates $(1 - \gamma)$ are of the same order as the expected fail rate $(1 - R)$.[7] Thus, for a nonmisleading confidence,

$$\gamma \approx R$$

but

$$\gamma = 1 - R^{n+1}$$

$$\therefore R \approx 1 - R^{n+1}$$

[6]Equation (3) is true only if samples are returned to the lot as drawn, or if the sample size n is so small by comparison with the lot size L that the removal of the sample has no significant effect on the proportion p of P's remaining.

[7]Two comments may be of use at this point: first, although "confidence" is frequently interpreted as a measure of the number of samples tested, this is not necessarily more true of confidence than it is of reliability; second, although large increases in otherwise low confidences can be made at the expense of relatively small decreases in high reliabilities, this situation is reversed if the reliability originally assumed is significantly lower than the corresponding confidence.

or

$$n \approx \frac{\log(1 - R)}{\log R} - 1 \tag{7}$$

The following is a tabulation of n's from Eq. 7 for various R's:

R	n
0.9999	92,000
0.999	6,900
0.995	1,060
0.99	455
0.95	57
0.90	21
0.85	10

The quantities indicated by Eq. 7 are so large for high reliabilities that proof of high reliability of aerospace ordnance components by attribute testing is generally impractical (or unrealistic).

It would seem, however, that tests by variables might allow the use of much smaller samples, even if results were evaluated to have matching confidence-reliability combinations, but this would be true only if a frequency distribution is assumed.

Thus the usual qualification program (involving, typically, no more than one or two hundred test samples) is an anachronism, a carry-over from applications where a limited number of successful tests was considered to demonstrate adequate reliability; a program involving 57 samples, for example, could demonstrate a reliability of only 0.95.[8] At best, small-sample qualification programs of the formalized type will do little more than detect gross fail rates; if test-sample quantities are limited, detection of low potential fail rates depends not on such qualification programs, but rather on skillful use of "off-limits" or other comparatively uncommon test techniques, directed toward revealing failure modes rather than toward demonstrating good qualities.

IV. Conclusions

1. Although strictly not applicable to items with no reliability history, test result analysis techniques epitomized by

$$\gamma = 1 - R^n$$

can continue to be used without introducing serious error.

2. Low-confidence, high-reliability combinations have little practical meaning, but a combination involving equal confidence and reliability will be indicative of the maximum proportion of fails that may be expected.

3. Samples of size n that yield no failures belong to "families" of lots having an inherent reliability

$$R_i = \frac{n+1}{n+2}$$

with a confidence $\gamma = 1 - R^{n+1}$.

4. Qualification tests are practically useless for demonstrating low failure rates.

5. Small-sample tests by variables are useful for comparative but not absolute evaluation of high reliability.

[8]See the text tabulation of n's for various values of R. This quantity would be required for each parameter; for example, firings at low, medium, and high temperature would call for a total of 171 if a 0.95 reliability were to be demonstrated at each temperature.