System Design and Assessment Notes

Note 26

A Qualitative Discussion of the Effect of Investment in Various Kinds of Technology on System C(R)

by

Paul Ryl

16 February 1984

R & D Associates
2305 Renard, SE
Albuquerque, New Mexico
87106

Abstract

Let C(R) denote the quantitative confidence justified by the available data for the statement that the probability of success P_s of a system is at least R. Then as the system is improved by investment in hardening technology, or as our knowledge of the system is improved by investment in assessment technology, the shape of a graph of C(R) will change. This note presents a brief qualitative discussion of such changes.

Contents

Figures

A qualitative discussion of the effect of investment in various kinds of technology on system C(R) (confidence in system reliability).

The technology which we will discuss is of at least three kinds, viz.,

A. Hardening technology

B. Reliability assessment technology

  1. Prediction technology
  2. Measurement technology

For any <u>level</u> of technology, we can always be 100% confident that the probability of survival of a system is at least 0 (since the probability of <u>any</u>thing is at least 0). And we always have 0% confidence that that probability is at least 1 (since <u>no</u> probability can exceed 1). Thus, for any level of technology the confidence C that the probability is at least R can be represented (qualitatively) by a figure in the RC plane like this:

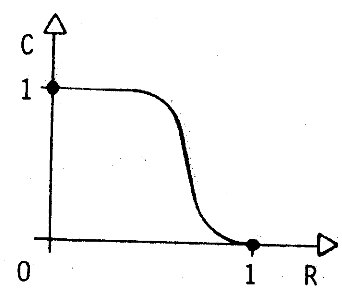


Figure 1. Two points which must always be on any function C(R).

So we know two points on this curve (viz., (0,1) and (1,0)) regardless of the level of technology. The shape of the curve <u>between</u> these two points is what the level of technology determines.

The remainder of this note is concerned with three topics:

a. General features which the C(R) curve always has.

b. What effects technology (of various kinds) has on the shape of the C(R) curve.

c. What the system user would like to do to the shape of the curve.

a. General features which the C(R) curve always has.

The C(R) function is always monotone decreasing in the RC plane. Why is this so? Well, let P_s denote the probability that the system will survive (or the probability that the electrical stress at a point inside the aircraft will not exceed the threshold of failure at that point, or "whatever"). Now, the function C(R) is just the confidence warranted by the data that P_s is at least R (for reliability). And for two values of R, say R_{low} and R_{high}, the confidence that P_s is at least R_{low} can be no less than the confidence that P_s is at least R_{high} (since the confidence in the latter is included in the confidence for the former, since R_{high} is at least R_{low}). Symbolically,

$$C(P_s \geq R_{low}) \geq C(P_s \geq R_{high}) \quad .$$

Or, more tersely,

$$C(R_{low}) \geq C(R_{high}) \quad .$$

Hence confidence C(R) must always be a monotone decreasing function of reliability R (where "reliability R" here means a lower bound on P_s).

Second, the C(R) plot is always "S-shaped". That is, qualitatively it is always as represented in this figure:

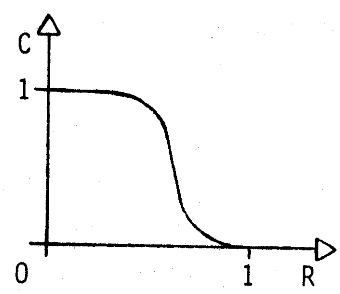


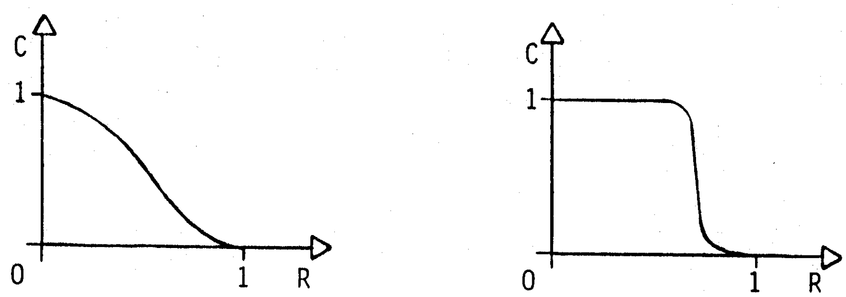
Figure 2. Generic graph of C(R).

The reasons for this are technical, and arise from the statistics of the binomial (or hypergeometric) distribution underlying this discussion (which in turn comes from P_s, which can be thought of as a binomial parameter). So we won't go into them in this note. The point that matters is that this S-curve has a point

(R value) of steepest descent. And that point is the <u>best guess</u> of P_s. (If we were being technical we would say that the value of R at which C(R) has its maximum absolute first derivative is the maximum likelihood estimate of P_s, denoted by $\hat{P}_s$. But we aren't, so we won't.)

b. Effect of technology on C(R).

In general a "sharp" assessment, i.e., one supported by a lot of assessment technology, meaning both a lot of good prediction theory and a lot of good experimental measurement, will have a clearly defined "step" in the "S shape". Qualitatively the effect is this:



Not much assessment technology, e.g., not much data.

A lot of assessment technology, e.g., a lot of data.

Figure 3. Effect of improving assessment technology.

The result is that $\hat{P}_s$ is much more clearly defined in the right figure than in the left. In the limit as, for example, the amount of data becomes infinite the curve becomes a <u>perfect step</u>, right at P_s:

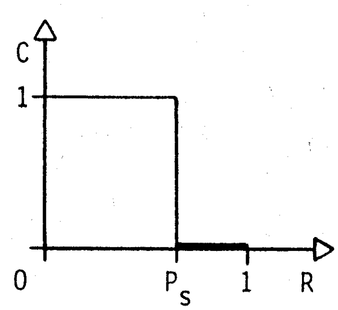
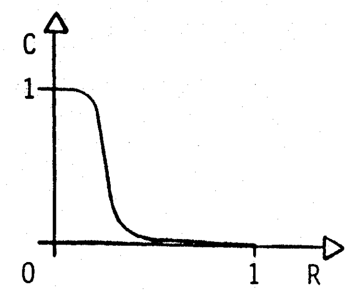
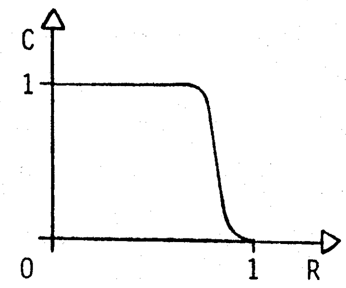


Figure 4. Result of perfect assessment (technology).

Superior hardening technology, in contrast, has a different effect on the C(R) curve, viz., it moves the step (however well or poorly defined) toward the right. Pictorially the effect of improving hardening technology is this:



Poor hardening technology.



Better hardening technology.

Figure 5. Effect of improving hardening technology.

c. What we would like to do to the C(R) curve.

If the system which the C(R) curve describes our knowledge of is useful to us, then we would like to have reasonably high confidence that most copies of that system will work. For consider the difference between having acceptable confidence in R = 80% vs having acceptable confidence in an R of only 50% . In the former case we must be prepared for the possibility that up to 20% of the copies of the system will not survive. So if we need 100 copies of the system "on target" we will have to procure 125 copies in order to be reasonably confident of delivering the 100 . But in the latter case, where R equals only 50% , we must be prepared for the possibility that up to 50% of the copies will fail, and so will have to procure 200 copies to be reasonably confident of putting at least 100 on target. Thus in the former case we will have to buy 25% more than we actually need, but in the latter case we will have to buy 100% more than we actually need.

Suppose the two kinds of technology together at present support a C(R) curve like this:

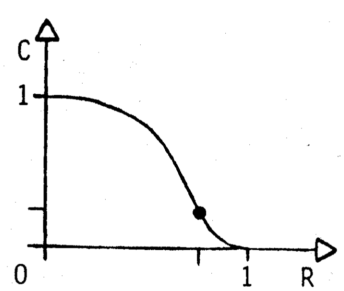


Figure 6. Today's function C(R) for the system.

It's a little bit "washed out" (poor assessment technology) and it's point of steepest descent isn't very far to the right (poor hardening technology). The dot on the curve is at an acceptable value of R, say .8 , but the confidence level there is rather low, say 15% . We might be prepared to do the 25% over-buy required by the R = .8 , but we have very little confidence (15%) that this level of over-buy will put the required number of items (or more) on target. So what do we do, <u>within existing technology</u> (all kinds), to get the level of confidence up to an acceptable level? As we move the dot to higher levels of confidence on the curve, it also moves to the left ... to values of R requiring more over-buy.

What can improving technology do for us in this situation?

Well, improving hardening technology alone will move the steep point of the S toward the right ... at some hardening expense. This will increase the height of the curve at R = .8 (and at all other values of R as well).

And improving assessment technology alone will improve the "definition" of the S . This might have the effect (perhaps at less expense) of moving the solid line in the following figure up to the dashed line:

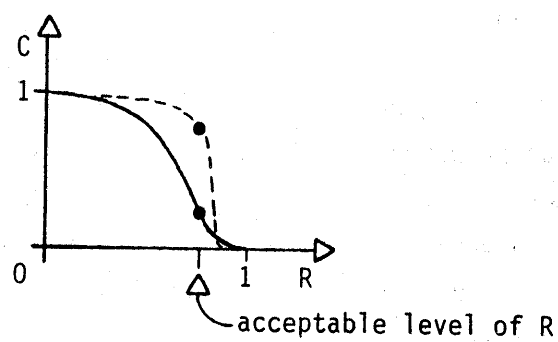


Figure 7. Possible result of improving assessment technology.

<u>Some closing remarks.</u>

As the foregoing shows, one can spend one's money on over-buy (i.e., on redundancy), on hardening technology, and on assessment technology. And investments in these three areas trade off against one another. So what one really has is a resource allocation problem, which can be discussed in terms of of the function C(R) and its graph.

It should be noted, though, that the last figure, showing what assessment technology can do for us, doesn't <u>have</u> to come out so rosy after additional investment in assessment technology. The S has to become better defined (more of a crisp step function) ... but this doesn't mean the resulting value at R = .8 will necessarily be higher than before the assessment technology advance. It could turn out lower, as in this figure:

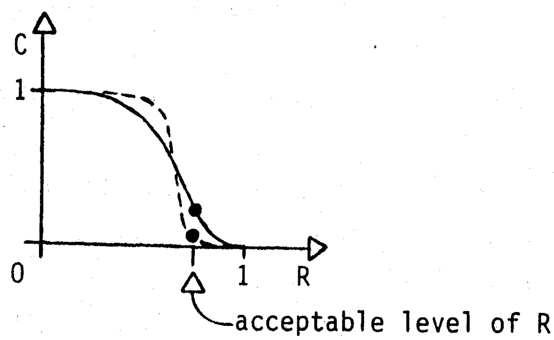


Figure 8. Another possible result of improving assessment technology.

(In both cases the location of the step place, i.e., the point of steepest descent, moved a little. But this wasn't because any hardening was gained or lost; it was only because $\hat{P}_s$ is, after all, a random variable, so that a new assessment with or without improved assessment technology can be expected to yield a slightly different value of $\hat{P}_s$.)

Finally, an observation about what might constitute a "reasonable" level of confidence. A thoroughgoing technical treatment would involve the costs of Type I and Type II error vs the costs of reducing those errors (in probability, or in expectation), and the statisticians have evolved many other ways of gauging the quality of estimates of quantities such as P_s . But in this note we make no pretense of offering a thoroughgoing treatment. Instead we offer a common sense rule of thumb. In general it doesn't make much sense to require very high levels of confidence in low values of reliability ("I'm totally sure it will do something, but I'm not sure what"), nor vice versa ("What it's going to do is work, but I'm not very sure of that"). After all, these can be had, approximately, from the "maximum ignorance" (i.e., before any data; uniform prior) curve:

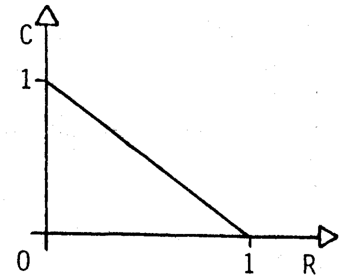


Figure 9. "Maximum ignorance" confidence distribution.

Instead, a good rule of thumb is to have the confidence level approximately comparable to the level of reliability at which the confidence is being evaluated. Pictorially this rule of thumb amounts to seeking the intersection of two curves, one being the C(R) dictated by the data (i.e., by the present level of the technologies) and the other being $C(R) = R$ (i.e., "the 45° line"):

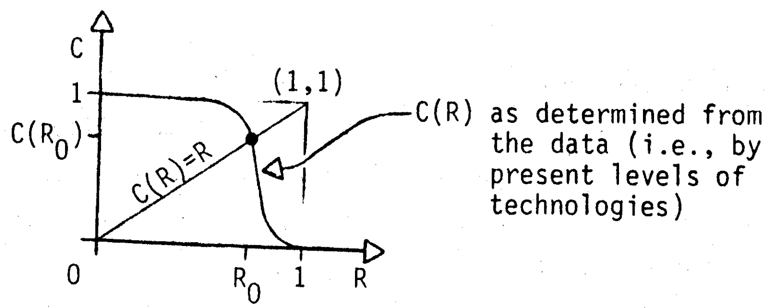


Figure 10. A rule of thumb for picking a useful single point on the C(R) curve.

Is the indicated point at a level of R high enough so that we can afford to buy a fraction $(1-R_0)/R_0$ more than we need? If not, then we need to buy some more technology.

C(R) for some limiting cases.

It might be useful, for the sake of perspective, to consider the C(R) curve in some limiting cases. To do this, let L denote the number of copies of the system which were tested (assumed realistically) and let M denote the number of those L which pass the test.

If all tested copies pass the test (i.e., $M = L$), then we get:

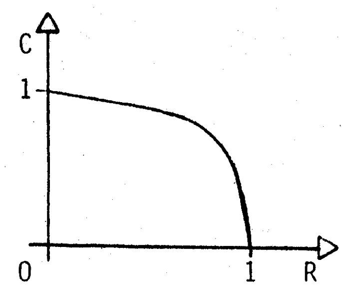


Figure 11. Everything passes.

Observe that the value of R for steepest descent is indeed at $\hat{P}_s = \frac{M}{L} = 1$. Conversely, a purely concave (as seen from below) C(R) curve can arise only from data in which everything passes.

If all tested copies fail the test (i.e., $M = 0$), then we get:

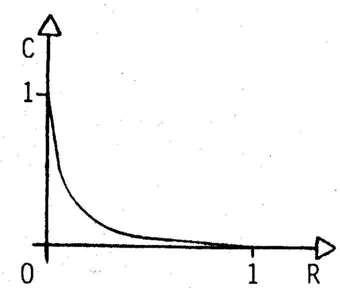


Figure 12. Everything fails.

Observe that the value of R for steepest descent is here at $\hat{P}_s = \frac{M}{L} = \frac{0}{L} = 0$ (assuming, of course, that $0 < L$). And again, conversely, such a purely convex (as seen from below) C(R) curve can arise only from data in which every tested item fails the test.

In the limit as the sample size goes to infinity if virtually all items pass we get:

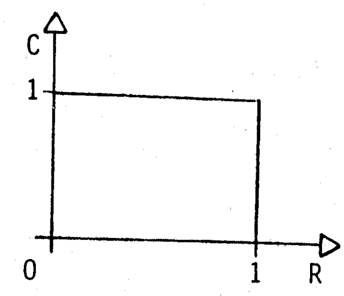


Figure 13. Everything passes in a very large sample.

On the other hand as the sample size goes to infinity if virtually all tested items fail we get:

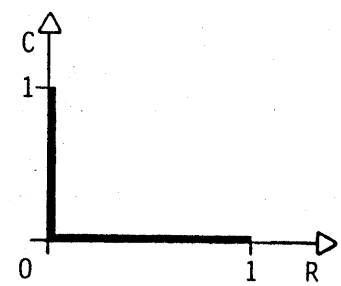


Figure 14. Everything fails in a very large sample.

Finally, as the sample size goes to infinity if we have a "mixed bag" of passing and failing then we get:

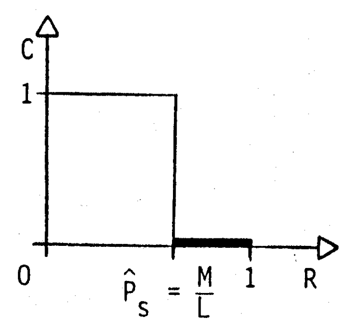


Figure 15. General C(R) for a very large sample.

The effects of error and different interests on C(R).

A system "passes" a test if the threshold of failure of that system is greater than the stress to which the system is subjected. But suppose that, instead of being able to test the system directly by applying the stress to it and seeing whether it "breaks" or "smokes", we are in a position in which we must calculate or measure the system's threshold, the level of stress to which the system will be subjected, or both. In that case we don't actually observe the system in the stressful environment, and observe directly the response of the system to that environment. Instead we do an algebraic comparison of the estimated values of the threshold and stress, to see which is greater. We refer to the difference threshold minus stress (e.g., in dB) as the margin of the system in that environment. If the margin is positive then the threshold is greater than the stress and the system would have survived, or passed, the test if it had actually been administered. If the margin is negative then the threshold is less than the stress and the system would fail the test.

However, since we are using estimates of threshold and stress we have only an estimate of the margin which is their difference. Therefore, by dealing with a calculated value of the margin rather than actually performing the test we incur the risk that the error in our estimate of the margin will occasionally result in our misclassifying a copy of the system as a "fail" (slightly negative estimated margin) when if it were actually tested it would in fact pass (actual margin in fact positive), or vice versa.

In view of this misclassification risk, why would we ever substitute such an algebraic comparison of estimates for an actual field test of the system? Basically because of cost, usually. For example, it may be very expensive to apply the actual stress (say a nuclear war) to the system. So we may elect instead to simulate the stress at a less than realistic ("sub-threat") level, and then for our estimate of the real stress use an extrapolation (e.g., linear) of our measurement of the coupled value.

But the risk is still there, so we might consider some of its consequences.

It is at this point that the different interests of various users of reliability data can begin to make themselves felt. As a crude simple example suppose that if we estimate the margin m to be $\hat{m}$ then the error in our estimates is such that we can be sure only that the true margin is somewhere between $\hat{m}-e$ and $\hat{m}+e$ (where e is a positive constant characteristic of our assessment technology, e for error). (Why is this just a "crude simple example"? Because we are adopting a rectangular distribution for the error, and because we are characterizing both the stress and the threshold as univariate. We are also assuming that we know perfectly the value of the parameter e of the rectangular distribution.) In such a situation one party might take the position that only values of $\hat{m}$ below $0-e = -e$ should be classified as "failures", since even though negative any estimate above -e might be in error by an amount great enough so that the actual margin of the system was in fact positive (i.e., the system would in fact pass if actually tested, despite a calculated value of $\hat{m}$ which was slightly negative). A person with this viewpoint might feel that if $\hat{m} = -.9e$ then it has simply not been shown conclusively that that copy of the system would fail if it were actually tested, and so that copy of the system should not be thrown in with the "failed" category. To put it more succinctly, this person might feel that a copy of the system should be treated as passable unless and until proven otherwise. This typically is the view of the "producer", who would have the burden of proof rest on any who would accuse his product of any inadequacy.

Conversely (or some might say perversely), the "consumer" might feel the burden of proof should rest with the person who claims this copy of the system is adequate, i.e., that a copy of the system is suspect until proven sound. This person might therefore advocate classifying as "failed" any copy of the system for which $\hat{m}$ came out less than +e , since even if positive a margin estimate less than +e could be in error be enough so that the actual system margin was negative (i.e., the system would fail if actually tested under realistic conditions, despite a calculated value of $\hat{m}$ which was somewhat positive).

Another way to look at this is from the point of view of hypothesis testing. Suppose the null hypothesis H_0 is that "this copy of the system will work even if realistically stressed." Then the producer would like to reduce "producer's risk", i.e., the probability of Type I error, which is the probability of rejection of this hypothesis H_0 when it is as a matter of fact actually true. To do this he classifies any "doubtful" copies as "passes". That is, if the calculated value of $\hat{m}$ for the margin m of a copy of the system is within e of zero then that's good enough; the producer throws that copy into the "pass" bin.

The consumer, on the other hand, would like to reduce "consumer's risk", i.e., the probability of Type II error, which is the probability of accepting this hypothesis H_0 when it is in fact false. To do this he classifies any "doubtful" copies of the system as "fails". That is, if the calculated value $\hat{m}$ for the margin m of a copy of the system is within e of zero then he throws it into the "fail" bin.

What are the consequences of these different attitudes toward how error in the data should be handled? Well, suppose the estimated values $\hat{m}$ of the margins m of various copies of the system are given by the x's in this figure:

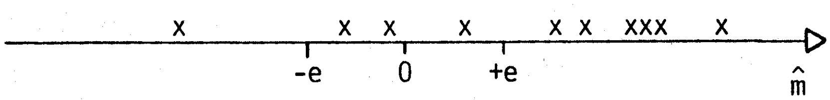


Figure 16. Estimated values of margin, with error.

Then the second and third estimates from the left are close enough to zero so that the fact that they are negative might be attributable entirely to error in the margin estimation process. So the producer would be inclined to classify these two copies of the system as "passes". As a result, in the notation of Appendix A, above, the producer would count a sample size of $L = 10$ and number of successes $M = 9$. In contrast the consumer would point out that the fourth estimate from the left is close enough to zero so that the fact that it is positive might be spurious, due only to error in the margin estimation process. So he would be inclined to classify that one as a "fail", just to be safe, and also all those to the left of it, a fortiori. So the consumer would count a sample size of $L = 10$, just as the producer did, but he would want to say the number of successes was $M = 6$.

In a nutshell, the producer wants to define M as the number of <u>possible</u> successes whenever there is possible error around, and the consumer wants to define M as the number of <u>guaranteed</u> successes.

Therefore we always have

$$M_{\text{consumer}} \leq M_{\text{producer}} \qquad ,$$

so to speak.

So what?

Well, a careful consideration of the details provided in the references of this note will show that for any R the function C(R) is monotone increasing in M. That is, if we compare the C(R) curve which the producer will get by classifying all "doubtfuls" (values of $\hat{m}$ within e of zero) as successes with that which the consumer will get by classifying all doubtfuls as failures, we will find that the producer's C(R) curve will always be as high as or higher than the consumer's C(R) curve. Pictorially:

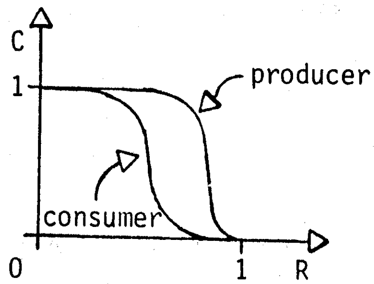


Figure 17. Comparing C(R) for producer vs consumer.

Notice that the points of maximum descent of these two curves are at $\hat{P}_s$ equal to $\frac{M_{producer}}{L}$ and $\frac{M_{consumer}}{L}$, respectively.

Finally, investment in assessment technology, in addition to seeking better prediction models and larger sample sizes, can also encompass an effort to reduce the value of e in the foregoing discussion (e.g., by better measurement and data processing techniques). By thereby reducing the width of the "gray area" between -e and +e in Figure 16, above, this can be expected to reduce the number of estimates $\hat{m}$ (x's in Figure 16) which will fall in that area of uncertainty, and therefore reduce the number of copies of the system concerning which there will be classification disagreement. Since this will bring $M_{producer}$ and $M_{consumer}$ closer together (in distribution), the result will be to reduce the distance between the producer's C(R) and the consumer's C(R). Thus another net effect of investment in assessment technology is to bring together the two curves in Figure 17, above.