JACQUELINE K. TELFORD

THE NUMBER OF TESTS NEEDED TO DETECT AN INCREASE IN THE PROPORTION OF DEFECTIVE DEVICES

One of the questions most frequently brought to a statistician is, How many tests will I need to run to detect an increase in the proportion of defective devices in a group? Minimizing the number of tests is important, since destructive testing of expensive devices is often involved. Different approaches can be used, depending on the assumptions made and how the testing is conducted. Five approaches are summarized in this article.

INTRODUCTION

Testing programs exist to determine if an increase has occurred in the proportion of defective devices in a group in relation to an initial value. The type of device being evaluated may be a complex system with many components, assemblies, and subsystems (e.g., rocket motors) or a single component of a system. At APL, missile systems are analyzed to detect a significant decrease in reliability, that is, a significant increase in the proportion of missiles that will fail. The number of missiles to be tested, the sample size, must be determined.

Binomial sampling to estimate a proportion of defective devices entails taking a random sample from a larger group and then classifying the devices as either defective or nondefective. The number of defective devices in the sample divided by the number of devices tested is an estimate of the proportion of defective devices in the larger group. (More devices may be defective because of aging chemicals, corrosion of parts, or cracks from the cumulative effect of environmental stresses.) The number of tests one must conduct to detect an increase in the defective proportion is especially important when destructive testing of expensive items is involved. The necessary sample size will depend on the initial proportion of defective devices, the amount of the increase in the defective proportion one wishes to detect, and the confidence one wants to have in the decision. The necessary sample size determines the number of devices to be tested and helps quantify the cost of the testing program.

The question to be answered by the testing and the different types of risks to be considered can lead to different statistical methods. The five methods discussed in this article, the first three of which have been used at APL to determine the number of missiles to be tested in weapon system evaluations, are as follows:

1. Fisher's Exact Test: tests whether two samples have the same proportion of defective devices.
2. One-Sample Neyman–Pearson Test: tests one sample against two specific proportions of defective devices (the initial defective proportion and an increase by a given amount).
3. Two-Sample Neyman–Pearson Test: tests two samples for an increase in the defective proportion by a given amount between the first and second samples.
4. Sequential Testing (also a Neyman–Pearson Test): tests a sample against two specific proportions of defective devices; a statistical test is made after observing whether each device is defective or nondefective to decide whether to accept one of the two values or to test another device.
5. Double Sampling Plan Test (also a Neyman–Pearson Test): tests a sample against two specific proportions of defective devices; the testing occurs in two groups, and a statistical test is made after the first group is tested to decide whether to accept one of the two values or to test the second group before deciding between the two values.

STATISTICAL FRAMEWORK

Sample sizes are usually determined in the context of statistical hypothesis testing. Statistical hypothesis testing has a particular framework and set of concepts and terminology into which a testing problem must be structured. For example, if one decides that the proportion of defective devices has increased when in fact it has not, this is a false alarm, and the false alarm rate is called α. On the other hand, if one decides that the defective proportion has not increased, when it has, this is another mistake, a failure to detect, and its probability is called β. Optimally, the probabilities of these two errors will be small; in textbooks α is usually set to 5% to indicate an event unlikely by chance alone. In a statistical context, confidence is the probability of correctly declaring that no increase has occurred and is equal to $1 - \alpha$ (or 95% if $\alpha = 5\%$). The probability of correctly declaring that an increase has occurred is called power and is equal to $1 - \beta$. Of course, one would like to have high confidence and high power in the testing, but to do so would require many tests. Determining acceptable levels for α and β is part of the problem formulation. In this article, a 25% false alarm rate and a 75% power will be used to limit the size of the examples. The choices of α and β should be determined by the consequences of the two types of errors. If one of the two is more serious than the other, then the probability of that error should be set to a small value and the other probability made considerably larger. On the other hand, if one is equally concerned about the two mistakes, then α and β can be equal.

The probabilities α and β are sometimes called producer's risk and consumer's risk, respectively. Imagine that a lot of identical components is received and in that lot a certain maximum proportion of defective items is acceptable. A sample of the lot is usually taken and tested to decide whether to accept or reject the whole lot. An incorrect decision to reject the lot is a false alarm and is a risk to the producer of the items (α). To decide incorrectly to keep the lot is a failure to detect the larger proportion of defects and is a risk to the consumer of the items (β). If the acceptance criterion is very stringent (very low consumer's risk, β), many good lots will be rejected, resulting in a high producer's risk (α). If the acceptance criterion is very loose (high consumer's risk), many bad lots will be accepted (low producer's risk). Thus, a trade-off exists between the two types of risk.

We assume in this article that the testing can be modeled by the binomial probability distribution. The assumptions for the binomial model are as follows:

1. The testing of each item will result in a classification of either defective or nondefective.
2. A constant proportion of defective items, denoted by p, exists in the population of items tested.

The probability of a certain number of defective devices (x) in a sequence of tests (n) can be computed as

$$\binom{n}{x} p^x (1-p)^{n-x} ,$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} .$$

The number of tests required is also a function of the initial proportion of defective devices and the amount of increase one wishes to detect with the preestablished amounts of α and β risks. To illustrate the procedures and computations used in testing, in this article the initial proportion of defective items is 0.15, and the amount of increase to be detected is 0.25. The assumption of no change in the defective proportion is called the null hypothesis, which is denoted by H_0. The assumption of an increase in the defective proportion is called the alternative hypothesis, which is denoted by H_A. The null and alternative hypotheses are written as

$$H_0\colon p = 0.15 \text{ versus } H_A\colon p = 0.40$$

or

$$H_0\text{: No Increase versus } H_A\text{: Increase of 0.25.}$$

Table 1 shows the decisions that can be reached and the associated probabilities of correctly or incorrectly making those decisions.

Table 1. Summary of decisions and probabilities.

	Actual situation	
Decision reached	H_0: No increase	H_A: Increase of 0.25
H_0: No increase	Confidence = $1-\alpha$	Failure to detect = β
H_A: Increase of 0.25	False alarm = α	Power = $1-\beta$

FISHER'S TEST

Ronald A. Fisher[1] published a test in 1934 variously known as the Fisher–Yates Test, Fisher's Exact Test, or Fisher–Irwin Test. Fisher's Test is a hypothesis test used to determine whether two binomial samples (the number of defective and nondefective devices in the two samples) can reasonably be expected to have the same proportion of defective devices. The acceptance or rejection of the hypothesis of no increase in the proportion of defective devices is based on computing the probability of the observed sets of defective and nondefective devices and also on more extreme sets with the same total number of defective devices:

$$\frac{\sum_{i=0}^{x_1} \binom{n_1}{i}\binom{n_2}{x_1+x_2-i}}{\binom{n_1+n_2}{x_1+x_2}} ,$$

where n_1 is the first sample size, n_2 is the second sample size, x_1 is the number of defective devices in the first sample, and x_2 is the number of defective devices in the second sample. If this probability is less than or equal to the prespecified false alarm rate, the null hypothesis is rejected, and the statement is usually made that an increase in the defective proportion has been detected at an $\alpha(100\%)$ risk level. Note that power (the probability of correctly deciding an increase has occurred) is not used in Fisher's Test.

The sample size is derived by ensuring that the false alarm rate (probability of declaring an increase when no increase has occurred) is no more than 25% in all cases for equal size samples with defective proportions differing by 0.25 or less. The smallest sample size for which the false alarm rate will be less than 25% is 12 (Table 2).

ONE-SAMPLE NEYMAN–PEARSON TEST

Jerzy Neyman and Egon Pearson[2] published a paper in 1933 that formulated the two types of errors (false alarm and failure to detect) discussed in the Statistical Framework section. The Neyman–Pearson Test is a hypothesis test designed to maximize the probability of

Table 2. Possible outcomes for Fisher's Test for a sample size of 12.

Initial sample $(x_1/n_1)^a$	Second sample $(x_2/n_2)^b$	False alarm rate (%)
0/12	3/12	11
1/12	4/12	16
2/12	5/12	19
3/12	6/12	20
4/12	7/12	21
5/12	8/12	21
6/12	9/12	20
7/12	10/12	19
8/12	11/12	16
9/12	12/12	11

[a]x_1 = number of defective devices in the first sample, n_1 = first sample size.
[b]x_2 = number of defective devices in the second sample, n_2 = second sample size.

correctly deciding that a change has occurred (the power of the test is maximized for a selected false alarm rate; see Table 1). The one-sample test compares one set of data with two hypotheses having certain proportions of defective devices. For example, the question could be stated as: Is the current defective proportion, p, 0.15 (H_0) or 0.40 (H_A)? The test is used to determine whether the observed results (and those more extreme) are more likely to have come from the H_0 or the H_A values of p and also to maximize the power of the test for a given false alarm rate. Bartlett[3] summarizes the major contributions of Neyman and Pearson to the foundations of statistical hypothesis testing by observing that "... there is no doubt that the general theory clarified considerably the current statistical procedures, and in particular counterbalanced Fisher's overemphasis of the null hypothesis, with its concomitant neglect of the consequences if alternative hypotheses were true."

Figure 1 is an example of the four probabilities in Table 1 computed using the binomial probability distribution and assuming one decides that the defective proportion is 0.40 (has increased by 0.25) if there are two or more defective devices in the sample. The green area in the H_0 histogram (Fig. 1A) is the probability of a false alarm (α), and the red area in the H_A histogram (Fig. 1B) is the failure to detect probability (β).

The sample size is determined by increasing it until the probabilities of both types of mistaken decisions (false alarm and failure to detect) are sufficiently small. These probabilities for sample sizes of 6, 9, and 12 are given in Table 3. Trade-offs between false alarm rate and power for different rejection criteria for the same sample size (12) are seen. For the same power, the false alarm rate decreases as the sample size increases. In addition, the power increases as the sample size increases for approximately the same false alarm rate. Because the binomial distribution is discrete, exactly 25% false alarm rates and 75% powers usually cannot be achieved for any given sample size.

A feature of the one-sample Neyman–Pearson Test is that the decision can be reached that an increase of 0.25 has occurred before the defective proportion is as high as that hypothesized by H_A. The decision to reject H_0 in favor of H_A can be reached with 2 defectives out of 6, 3 defectives out of 9, or 4 defectives out of 12. The defective proportion is consistently 0.33. That is, an increase in the proportion of defective devices to 0.40 can be detected before it is estimated as high as 0.40, since the test determines whether the data are more likely to have come from a binomial distribution with $p = 0.15$ or $p = 0.40$.

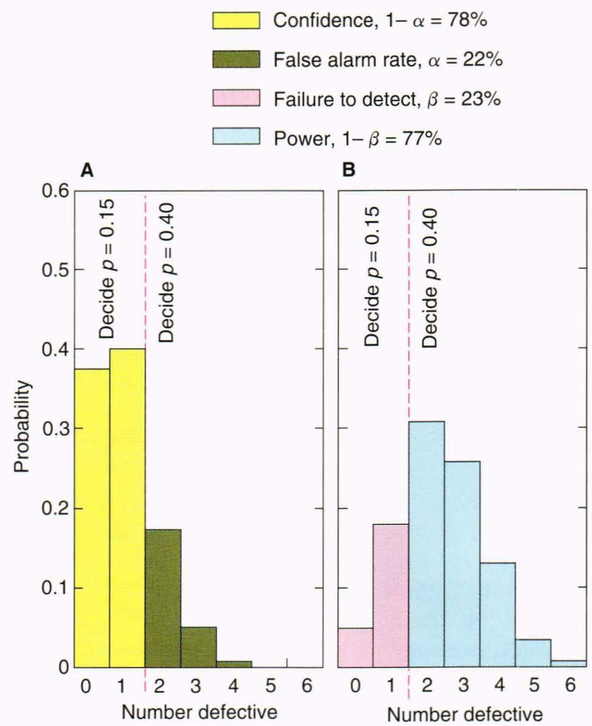

Figure 1. Histograms of the binomial probability for the number defective. **A**. Probability given $p = 0.15$ (H_0). **B**. Probability given $p = 0.40$ (H_A).

Table 3. One-sample Neyman–Pearson Test for sample sizes of 6, 9, and 12.

Sample size	Criterion: reject H_0 if the number defective is	False alarm rate (%)	Power (%)
6	2 or more	22	77
9	3 or more	14	77
12	4 or more	9	77
12	3 or more	26	92

TWO-SAMPLE NEYMAN–PEARSON TEST

The two-sample Neyman–Pearson test is used to compare two samples to decide whether the estimates from the two sets are more likely to be estimating the same defective proportion or if the second sample is estimating another defective proportion that is 0.25 greater than the first. The uncertainty in the two estimates is taken into account by this approach. An approximation using the normal distribution will be used, although the adequacy of the approximation may be questioned. The usual rule-of-thumb that both np and $n(1 - p)$ should be at least 5 may be satisfied, however.

The null and alternative hypotheses can be written as

$$H_0: p_2 - p_1 = 0 \text{ versus } H_A: p_2 - p_1 = 0.25 ,$$

where p_1 is the defective proportion in the first sample of size n_1, and p_2 is the defective proportion in the second sample of size n_2. The variance of the estimate of a binomial parameter p is $p(1-p)/n$. The variance of the difference in the defective proportions is $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$. Under the hypothesis of no change in the defective proportions, the variance of the difference can be simplified to $p(1-p)(1/n_1 + 1/n_2)$, where p is the common defective proportion.

Figure 2 shows the distributions of the estimates of the difference between two defective proportions given either H_0 or H_A. The centers of the H_0 and H_A curves are placed at 0 (no increase) and 0.25 (the increase we wish to detect), respectively. The area under each curve is 1 or 100%. The vertical line dividing the red and green areas is called the critical value and is the boundary at which we change from deciding that no increase in the defective proportion has occurred to deciding that an increase has occurred. The red and green areas are the probabilities of mistaken decisions (β and α, respectively). Moving the critical value to the right or left will decrease or increase α with the opposite effect on β. By placing the critical value at the point where the density curves intersect, β will be slightly larger than α, since the H_A curve is more spread out than the H_0 curve. Increasing the sample size increases the steepness of the curves. The sample size for testing is determined by increasing the sample size until the prespecified values for α and β are met. From Figure 2 we can see that

$$Z_\alpha \sqrt{\mathrm{Var}(p_1 - p_2)|H_0} + Z_\beta \sqrt{\mathrm{Var}(p_1 - p_2)|H_A} = 0.25 .$$

The coefficients Z_α and Z_β are values from the normal probability distribution and are determined by the specified false alarm rate and power.

The sample size needed to detect a 0.25 increase in the defective proportion from 0.15 to 0.40 can be derived assuming the same number of samples for the two sets of observations, a 25% false alarm rate, and a 75% power as follows:

$$0.6745\sqrt{(0.15)(0.85)(2/n)} +$$
$$0.6745\sqrt{(0.15)(0.85)(1/n)+(0.40)(0.60)(1/n)} = 0.25 .$$

Solving for n yields a necessary sample size of 9 for each of the two sets of observations.

SEQUENTIAL TESTING

Sequential sampling was developed by Abraham Wald[4] in 1944. Sequential testing is another form of the Neyman–Pearson test similar to the one-sample test in that the decision is made as to which of two fixed proportions of defective devices the observed test results more likely represent. Rather than testing all devices in the required sample size and then deciding which of the two proportions is more likely, the decision is made after each test result (defective or nondefective) whether to accept one of the two proportions or to test another device. Testing proceeds one device at a time, and the number of devices that will be tested is not known in advance. An expected number of devices to be tested can be computed, which is usually about half the fixed sample size required, since large increases in the defective proportion or a very low defective proportion will be detected early in the testing. To avoid the possibility of a very large sample size, a truncated sequential testing plan can be devised.

The statistical test after each observed test result (defective or nondefective) can be conveniently performed by plotting the number of tests and the number of defective devices on a graph where the shaded areas indicate acceptance of one or the other of the defective proportions, and the unshaded area indicates the need to continue testing devices. The false alarm rate and the failure-to-detect probability must be chosen in advance to implement the sequential testing procedure. Figure 3 is an example of the graph resulting from having a 75% probability of detecting a defective proportion of 0.40 and a 25% false alarm rate if the defective proportion is 0.15. The formulas used to develop Figure 3 are given by Crow et al.[5]

A truncated sequential test limits the number of tests. The truncation point for this example could be 11 devices, although a truncation point of 12 will be used here to be comparable with the fixed-sample-size test (one-sample Neyman–Pearson test). The possible outcomes and the associated decisions for the truncated sequential test with binomial data are shown in Figure 4. To compute the expected number of devices to be tested until a decision is reached that the defective proportion is 0.15 or 0.40, the probabilities of reaching the decision points must be calculated. From these probabilities, the expect-

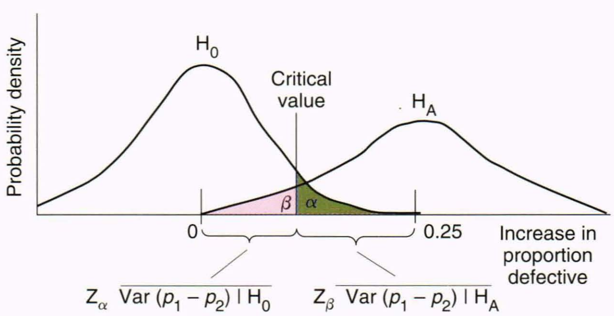

Figure 2. Distributions of increase in the defective proportion for the two-sample Neyman–Pearson Test.

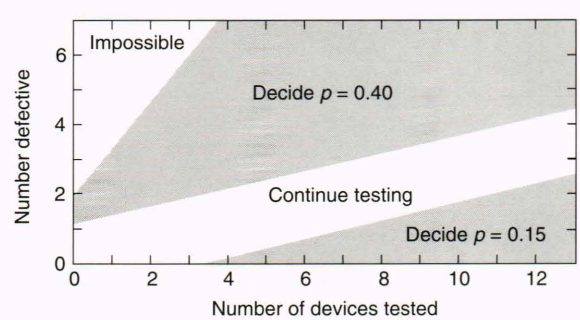

Figure 3. Sequential testing procedure.

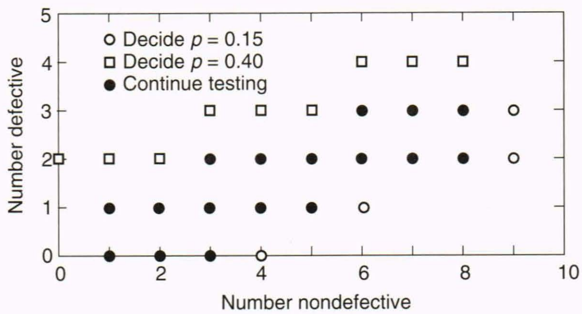

Figure 4. Truncated sequential testing procedure.

ed sample sizes are computed. The expected sample size when the proportion defective is 0.15 is 5.5, and when the proportion defective is 0.40, the expected sample size is 5. The false alarm rate and power can also be computed using the probabilities of reaching the boundary points. The false alarm rate is 16%, and the power is 75%. By comparing this false alarm rate and power with those in Table 3, we find that a test with a fixed sample size of 9 is the closest to the sequential tests with expected sample sizes of about 5. This comparison (9 versus 5) demonstrates the expected savings in the number of devices to be tested for a sequential test.

DOUBLE SAMPLING PLAN TEST

The double sampling plan test is also based on the Neyman–Pearson test and is an intermediate type of testing plan that incorporates features of a fixed-in-advance sample size approach and a sequential, one-at-a-time testing procedure. A double sampling plan entails testing all the devices in the first group (usually one-half or one-third of the total testable devices). If a certain minimum number of defective devices is found, the decision is reached without further testing that the lower of the two proportions is correct. If a certain maximum number of defective devices is equaled or exceeded, a decision is reached without further testing that the higher of the two proportions is correct. If an intermediate number of defective devices is found in the first group of tests, the rest of the devices are tested, and the decision between the two proportions is reached on the basis of the total defective devices. See Crow et al.[5] for a discussion of double sampling in quality control.

A double sampling plan test is illustrated in the boxed insert. Let the two samples be six devices each. A decision rule for deciding to accept one of the two proportions from the first sample only and for deciding between the proportions from both samples is evaluated by computing the false alarm rate and power when using the rule. Let X_1 be the number defective in the first sample and X_2 the number defective in the second sample.

The expected number of devices to be tested is 9.5 if the proportion defective is 0.15 and 9 if the proportion defective is 0.40. The false alarm rate for this double sampling plan is 25%, and the power is 87%.

This double sampling plan can be used, since the power and false alarm rate are within the 75% and 25% specifications. This double sampling plan is comparable

DECISION RULE FOR DOUBLE SAMPLING PLAN

Decide $p = 0.15$ if $X_1 = 0$ out of the first 6.

Decide $p = 0.40$ if $X_1 = 3$ or more out of the first 6.

If $X_1 = 1$ or 2, test the second sample of 6.

Then decide $p = 0.40$ if $X_1 + X_2$ is 3 or more.

Decide $p = 0.15$ if $X_1 + X_2$ is 2 or less.

Region Where $p = 0.40$ Decision Is Made

$X_1 = 3$, 4, 5, or 6

$X_1 = 1$ and $X_2 = 2$, 3, 4, 5, or 6

$X_1 = 2$ and $X_2 = 1$, 2, 3, 4, 5, or 6

to the fixed sample size of 12 in Table 3. The expected sample sizes of 9.5 and 9 may represent significant savings over the fixed sample size of 12.

Many other possible double sampling plans can be devised, such as using four tests in the first sample and then eight in the second sample. The false alarm rate and power must be calculated for each to see if the proposed sampling plan meets the specifications for false alarm rate and power. As stated by Burington and May,[6] "A systematic trial and error method is thus evolved for building various sampling plans of interest." Some standard test procedures with double sampling plans for quality control are cataloged in MIL-STD-105D, *Sampling Procedures and Tables for Inspection by Attributes.* These plans, however, usually apply to rather small false alarm rates and high powers. The plans generally assume that the first and second samples are of equal size or that the second sample is twice the size of the first sample. An extension of double sampling called multiple sampling[5] or grouped sequential sampling[6] might also be useful when testing is performed in groups of devices.

CONCLUSIONS

Fisher's Test can be used when one is only concerned about the false alarm rate and the sample sizes are small. Concern exists, however, that Fisher's Test is not powerful; consequently, several alternative tests have recently been published in the statistical literature. One of the Neyman–Pearson approaches should be considered when a certain increase in the proportion of defective items needs to be detected with a given false alarm rate and power. The one-sample Neyman–Pearson test requires a relatively small number of tests, but it assumes that the value for the initial defective proportion is known perfectly. The same is true for the formulations of the sequential and double sampling plan tests. The two-sample Neyman–Pearson test incorporates the uncertainty of both the initial defective proportion and the current defective proportion, since both are derived from testing. The sequential testing would probably give the earliest termination of testing if the failure rate is very high or low. Table 4 summarizes the sample sizes, false alarm rates, and powers for the five tests discussed.

Table 4. Comparison of five tests for detecting an increase in the proportion of defective devices.

Test	Maximum sample size	Expected sample size	False alarm rate (%)	Power (%)
Fisher's	12	12	24	61
One-sample	6	6	22	77
Two-sample	9	9	25	75
Sequential	12	5.5	16	75
Double sampling	12	9.5	25	87

Other approaches such as Bayesian or decision theoretic methods could be used for sample sizing and hypothesis testing. Some of the most commonly used methods, however, have been summarized in this article. Lloyd and Lipow[7] have produced a good reference book on reliability that discusses most of the topics summarized in this article and others, such as reliability growth modeling.

REFERENCES

[1] Fisher, R. A., *Statistical Methods for Research Workers*, Oliver and Boyd, Ltd., Edinburgh, pp. 96–97 (1958).

[2] Neyman, J., and Pearson, E. S., "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Phil. Trans. Roy. Soc. Ser. A* **231**, 289–337 (1933).

[3] Bartlett, M. S., "Egon Sharpe Pearson, 1895–1980: An Appreciation by M. S. Bartlett," *Biometrika* **68**, 1–12 (1981).

[4] Wald, A., *Sequential Analysis*, John Wiley and Sons, New York (1947).

[5] Crow, E. L., Davis, F. A., and Maxfield, M. W., *Statistics Manual*, Dover Publications, Inc., New York, pp. 212–213 and 220–221 (1960).

[6] Burington, R. S., and May, Jr., D. C., *Handbook of Probability and Statistics with Tables*, McGraw–Hill, New York, pp. 315–319 (1970).

[7] Lloyd, D. K., and Lipow, M., *Reliability: Management, Methods, and Mathematics*, Prentice–Hall, Inc., Englewood Cliffs, N.J. (1962).

THE AUTHOR

JACQUELINE K. TELFORD received a B.S. degree in mathematics from Miami University in 1973 and M.S. and Ph.D. degrees in statistics from North Carolina State University in 1975 and 1979, respectively. She was employed at the U.S. Nuclear Regulatory Commission in Bethesda, Maryland, from 1978 to 1983. Since joining APL in 1983, she has worked in the Systems Studies and Simulations Group of the Strategic Systems Department on reliability analysis and testing, test sizing, and planning for Trident programs.