NIST Technical Note 2045

Confirming a Performance Threshold with a Binary Experimental Response

Dennis D. Leber
Leticia Pibida
Alexander L. Enders

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2045

National Institute of Standards and Technology
U.S. Department of Commerce

NIST Technical Note 2045

Confirming a Performance Threshold with a Binary Experimental Response

Dennis D. Leber
Statistical Engineering Division
Information Technology Laboratory

Leticia Pibida
Radiation Physics Division
Physical Measurement Laboratory

Alexander L. Enders
Oak Ridge National Laboratory
U.S. Department of Energy

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2045

July 2019

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

National Institute of Standards and Technology Technical Note 2045
Natl. Inst. Stand. Technol. Tech. Note 2045, 25 pages (July 2019)
CODEN: NTNOEF

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2045

Abstract

When designing a test to confirm that an artifact (e.g., a radiation detection system) meets a performance threshold where the artifact's performance is estimated based on a binary response, the number of required observations is often an initial question. To determine the required sample size, two pieces of information are necessary: the performance threshold; and a statement of acceptable risk or required confidence. In this chapter, we provide guidance on developing a defensible and successful test through the informed selection of a performance threshold and statement of acceptable risk. Using the statistical hypothesis testing framework, we illustrate the meaning of risk and confidence from both the consumer and producer's perspectives. We define the power of a test and demonstrate how an experimenter can use a power curve to balance the tradeoffs between test burden (costs) and producer risk (type II error) while satisfying the required confidence. We provide a sample size and acceptance criterion table to define a fixed sample test that will satisfy a variety of performance thresholds and levels of acceptable risk. We conclude with a general discussion of sequential sampling tests and provide important considerations and contrasts to their fixed sample counterparts.

Keywords

Binomial test response; consumer and producer risks; hypothesis test; performance threshold; power of a test; radiation detection systems.

Table of Contents

List of Tables

List of Figures

1. Introduction

Often, experimenters wish to confirm that a test artifact meets some predefined, fixed performance criterion or claim. For example: does the newly formulated pharmaceutical reduce the disease rate by 10 % or more; will the composite overwrapped pressure vessel fail prior to the completion of the 15-year space mission; or, can the radiation detection system detect the specified radiological source with at least 80 % probability? The answers to these questions are a simple yes or no, but because of the inherent uncertainty in the measurements used in the assessment, there is a risk of answering the question incorrectly. This chapter provides guidance on developing an experimental sample size and acceptance criterion to determine whether a test artifact satisfies a predefined and fixed performance criterion when the response variable observed is binary, such as a success or failure to detect. We provide a sample size and acceptance criterion table to define a fixed sample test that will satisfy a variety of performance thresholds and levels of acceptable risk. We conclude with a general discussion of sequential sampling tests and provide important considerations and contrasts to their fixed sample counterparts.

2. Choosing a Performance Threshold

A defensible and successful test always begins with a testable objective. As will be further defined throughout this chapter, a test to support the confirmation that a test artifact satisfies a performance threshold where the experimental response is binary will consist of two components:

1. a performance threshold; and
2. a statement of acceptable risk or required confidence.

Together, the defined performance threshold and statement of acceptable risk will lead directly to the required number of trials (samples) and acceptance criterion. If the number of trials required to support the performance threshold at the stated level of acceptable risk cannot be achieved due to budgetary or other constraints, then the value of performing a lesser test must be considered. Here, a lesser test is a test that maintains a lower performance threshold or assumes a higher level of risk than desired. This section presents a philosophical view on setting a performance threshold. A description of confidence and risk, and guidance on selecting an acceptable risk are presented in Sec. 5.2.

A defensible and successful test begins with a testable objective that includes a performance threshold and required level of confidence or acceptable risk. The number of trials necessary and the acceptance criterion follow directly from these test requirements.

For radiation detection applications, performance thresholds are often based on the consequences of not detecting a threat object with a single radiation detection system or a system of

systems. These performance thresholds may be directed by public policy, user needs, standard requirements (e.g., ANSI, IEC, TCS), or acquisition requirements (see Acronyms Section). Current technological limitations in passive radiation detection should be considered when defining the performance threshold. Care should be taken to consider the class of detector, source strength of interest and other test parameters. There are several classes of radiation detection systems that vary in detector size, capability, and intended use, e.g., PRDs, SRPMs, RIIDs, BRDs, mobile systems, RPMs, and SRPMs. It is unrealistic to expect a small, portable PRD to perform the same as a large, stationary RPM. Requirements in the ANSI and IEC standards, for example, provide reasonable estimates of the performance levels equipment can attain today. The test parameters may be based on how detection systems are used in an operational setting or defined by standard test methods. A risk of defining the performance threshold solely on user requirements is that unrealistic goals may be set for a given class of detection systems. There may be value in accepting a lower performing system so long as the system's limitations are well understood, and users can adjust their operating procedure to compensate for these limitations.

The consequence of failing to detect a threat object, detector technological limitations and intentions, test parameters, and user needs should be considered collectively in defining a performance threshold for radiation detection systems.

---

3. Binary Response Variable

An experiment with a binary response is often used in estimating a performance measure that takes the form of a rate, a ratio, or a probability. The probability that a radiation detection system correctly detects or identifies a given source is an example of such a performance measure. Experiments with two, and only two possible outcomes, such as head or tail, success or failure, or defective or non-defective are known as Bernoulli trials [1]. The probability of one of the two outcomes (e.g., "detect") is denoted by p, while the probability of the complementary outcome ("no detect") is given by $1 - p$.

The total number of events observed (e.g., detections) in a sequence of independent and identical Bernoulli trials is distributed as a binomial random variable. The binomial distribution is characterized by two parameters, n and p, where n represents the number of trials and p represents the probability of the outcome of interest. The binominal distribution, as described by Casella and Berger [1], is defined in Eq. (1).

$$P\left(X = x \mid n, p\right) = \binom{n}{x} p^{x} \left(1-p\right)^{n-x} \quad x = 0,1,2,\ldots,n; \quad 0 \le p \le 1 \qquad (1)$$

The focus of this chapter is to help develop a test whose objective is to determine if the test artifact's true but unknown performance measure, p, is above or below some predefined, fixed

performance threshold that we denote by p^*. We are only concerned with a one-sided test, that is, investigating $p \geq p^*$ (or, if appropriate, $p \leq p^*$). We provide guidance in this chapter for determining the sample size, *n*, and the acceptance criterion to prove that a test artifact meets or exceeds a performance threshold.

4. Stating the Test Requirement

Ensuring that a testable requirement has been stated is the initial, crucial step in identifying the sample size and acceptance criterion to prove that a test artifact satisfies a performance threshold. For this purpose, a testable requirement has two key parts: 1.) a performance threshold, and 2.) a statement of acceptable risk or required confidence. For example, ***the radiation detection system shall provide at least an 80 % probability of correct detection with 95 % confidence when exposed to source A under conditions X***, is a testable requirement. In this example, 80 % probability of correct detection is the minimum performance threshold and 95 % confidence is the statement of required confidence. Without these two key pieces of a test requirement, a test's necessary sample size and acceptance criterion cannot be determined.

A test requirement must contain both a performance threshold and a statement of acceptable risk (or required confidence).

We note that the performance threshold could be either a lower bound as in the example above, or an upper bound, as would be the case in assessing a detection system's false alarm performance. In this chapter, we focus on the performance threshold as a lower bound, assessing the probability of detection. In a subsequent chapter we focus on the performance threshold as an upper bound in false alarm testing.

We are interested in drawing a conclusion about the true value of the performance measure of the test artifact, but all that we have available is an uncertain estimate of the performance measure obtained from the test results. It is this uncertainty that leads us to the possibility of drawing the wrong conclusion. In the following sections, we present a rigorous approach to designing a test that allows for the probability of drawing an incorrect conclusion to be quantified and controlled.

5. Hypothesis Tests

The statistical method that may be used to support the task of confirming that a test artifact meets a specified performance threshold is the hypothesis test [2]. Hypothesis testing begins with a specific conjecture called the *null hypothesis*. Data is gathered that directly pertain to whether the null hypothesis is true. All possible outcomes of the data are considered in establishing an acceptance criterion. The collected data are examined and, in conjunction with the established acceptance criterion, the null hypothesis is either rejected or not. The following subsections provide details on implementing a hypothesis test to prove that a test artifact's performance measure of interest satisfies a performance threshold.

5.1. The Null Hypothesis

The true state of a test artifact can fall into one of two categories when its true performance measure is compared to a performance threshold; that is, the true performance measure meets or exceeds the performance threshold, or it does not. If we label a test artifact as "good" if its true performance measure satisfies the performance threshold, and "bad" otherwise, then two possible positions exist for the null hypothesis conjecture: 1.) the test artifact is good; or, 2.) the test artifact is bad. Because the hypothesis test relies on the idea of proof by contradiction, we state the null hypothesis conjecture as opposite of what we would like to prove. Thus, in our effort to prove that the test artifact is good, we adopt as the null hypothesis that the test artifact is bad. For example, if we seek to prove that a radiation detection system meets or exceeds a detection performance threshold of 80 %, then we state the null hypothesis as *the radiation detection system's true probability of correct detection against source A is less than 80 %*.

Because the hypothesis test relies on the idea of proof by contradiction, we adopt as the null hypothesis conjecture that the test artifact is bad and seek data to prove that it is good.

Based on the established acceptance criterion and the observed patterns in the collected data, we either reject the null hypothesis in favor of its alternative or fail to reject the null hypothesis. Rejecting the null hypothesis in this case leads us to the conclusion that the test artifact's performance measure satisfies the performance threshold, i.e., the test artifact is good. Failure to reject the null hypothesis is not evidence that the test artifact is bad, but rather that insufficient evidence was found to support the conclusion that the test artifact meets or exceeds a detection performance threshold, that is, we fail to deem the artifact as good.

5.2. Errors in Hypothesis Testing

A test artifact has a true but unknown value of its performance measure. It follows, that the test artifact has a true but unknown state, either "good" or "bad", as would be determined by comparing its true performance measure value to the stated performance threshold. The statistical hypothesis test provides a framework for an experimenter to deem an artifact as "good", based on an estimate of the artifact's true performance measure value. Because the estimated performance value is uncertain (all measurements carry uncertainty), our conclusion about the true state of the artifact may be incorrect. The following subsections describe the two ways in which we may draw an incorrect conclusion and how we can control the rate at which these errors occur through the definition of the test.

5.2.1. Consumer and Producer Risks

There are two ways that we may make a mistake. The first error, a *false positive*, happens when our hypothesis test leads us to deem the artifact to be "good" when in fact, the test artifact's true state is "bad." Statisticians refer to this mistake as a *type I error* and denote the probability

of its occurrence with the Greek letter α. We note here that the statistical term *confidence level* is defined as 1 – α and the statistical term *significance level* is defined as α.

The second error that could be made in carrying out a hypothesis test, a *false negative*, happens when the test artifact is truly "good," but we fail to deem the artifact as "good." Statisticians refer to this as a *type II error* and denote the probably of its occurrence with the Greek letter β. These errors are illustrated in the truth table displayed in Table 1.

Table 1: Hypothesis test truth table.

		Artifact's True State	
		"Good"	"Bad"
Hypothesis Test Conclusion	Deem "Good"	Correct Decision	Type I Error
	Fail to deem "Good"	Type II Error	Correct Decision

The severity of the consequences associated with each of the above described errors are often not equivalent and the sensitivity to each depends on perspective. For example, a potential consumer of a radiation detection system, such as the U.S. Customs and Border Protection (CBP), will go to great lengths to protect itself from purchasing and deploying a "bad" system because the consequence of such an action could have a detrimental impact on the American public if illicit radiological or nuclear material were able to compromise a protected area. Thus, the CBP will desire a test with a low probability of committing a type I error. On the other hand, it is in the best interest of the manufacturer of the radiation detection system under test to minimize the probability of a type II error as such an error may lead to his truly "good" system not being purchased. For these reasons, the risk associated with a type I error in this construct is termed *consumer risk*, and that associated with a type II error is termed *producer risk*.

5.2.2. Power of a Test

Fortunately, both the consumer risk and producer risk can be controlled through the design of the hypothesis test and the selection of the sample size. These risks can be evaluated prior to conducting a test and are illustrated through a test's *power curve* that displays the probability of deeming a test artifact as "good" as a function of the test artifact's true but unknown performance measure.

An ideal test would deem a test artifact as "good" with certainty (i.e., a probability of one) when the artifact's true performance measure value meets or exceeds the desired performance threshold and never deem a test artifact as "good" when the artifact's true performance measure value is below the performance threshold. Figure 1 provides a power curve for this ideal test when the performance threshold, $p^* = 0.80$.

Unfortunately, a test with no risk, such as the ideal test illustrated with the power profile displayed in Fig. 1, requires an infinite number of samples. Therefore, common practice is to state a maximum acceptable consumer risk (type I error probability) and construct a suitable

acceptance criterion and sample size. The resulting power curve is examined, and the sample size adjusted to satisfy the desired producer risk (type II error probability). As discussed in Sec. 4, this statement of maximum acceptable consumer risk, paired with the minimum performance requirement provides the necessary basis for constructing the hypothesis test.

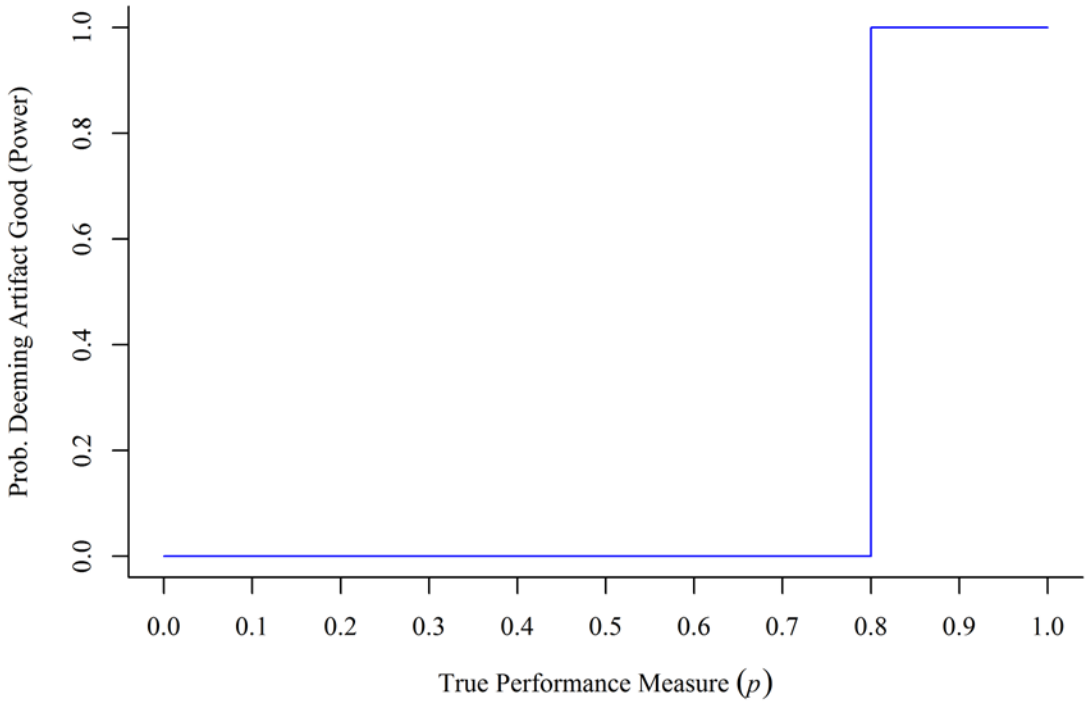

Fig. 1: Power curve for an ideal test with a performance threshold p = 0.8.*

An experimenter must carefully consider the consequence of committing a type I error before setting its maximum acceptable value. For experiments published in the medical and health science literature, where committing a type I error may have detrimental implications on human life, the maximum acceptable type I error is often selected to be very small, e.g., 0.01 or 0.001. For experimental results found in the physical science literature, when implications on human life are typically lower, type I error rates are often selected (by default) to be 0.05.

For homeland security applications, the type I error probability is interpreted as the probability of purchasing and deploying a "bad" detection system. Such an action would lead to a faulty detection system being fielded that, unbeknownst to the operator, does not perform as specified and may result in illicit material going undetected. It should be noted, however, that in certain operational situations the detection of illicit material may not be limited to the quality of a single detection system but rather to that of a system of detection systems. In these situations, the entire system design should be considered in determining acceptable risk. Thus, the selected type I error should be carefully considered, and selected based on the goals and policies set within the Department of Homeland Security (DHS).

A statement of acceptable risk, i.e., the type I error probability, defines the probability that a "bad" test artifact will be accepted. Type I error probability of ≤ 1 % is common practice in the medical and health science fields, where failure consequences are dire. Type I error probability of 5 % is common practice in the physical sciences. DHS goals and policies should drive their statement of acceptable risk.

Figure 2 illustrates power curves for the ideal test ($n = \infty$) and tests of sample size $n = 25, 50, 100, 500$, each with a consumer risk (type I error probability) no greater than 0.05 and a performance threshold, $p^* = 0.80$. We first observe that for the limited sample tests when $n \neq \infty$, the power to the left of the performance threshold is similar. That is, for each of these tests, when the artifact under test has a true performance measure $p < 0.80$, i.e., a "bad" artifact, the probability of deeming the artifact as good does not exceed 0.05.

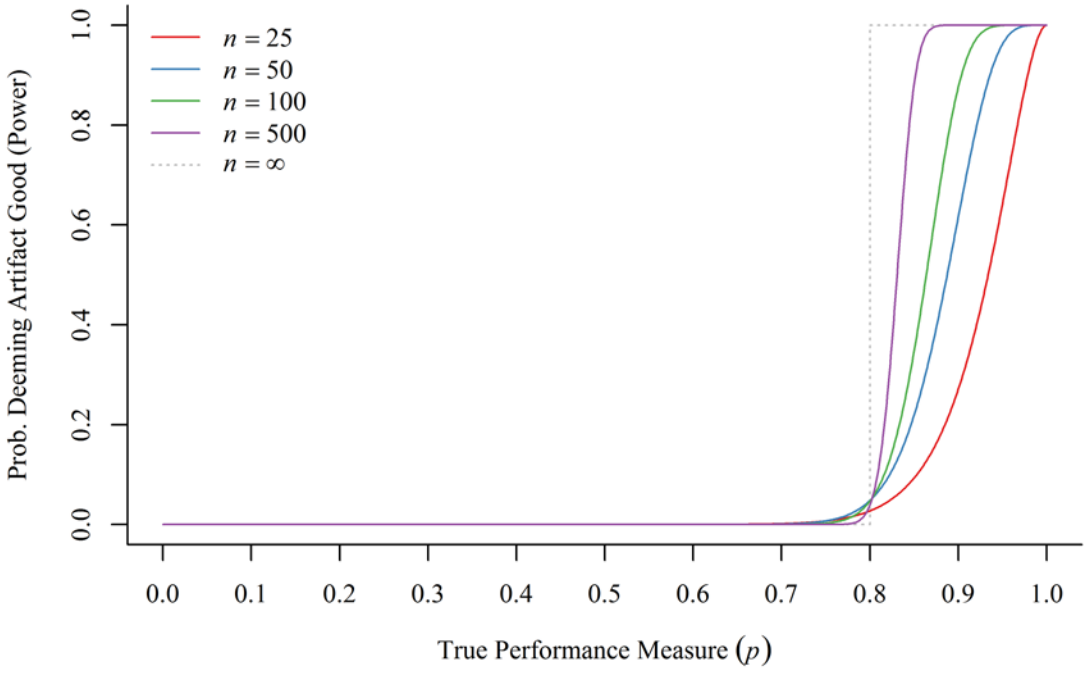

Fig. 2: Power curves for several tests of varying sample sizes, n, each with a maximum consumer risk (type I error probability) α = 0.05 and a performance threshold p = 0.8.*

On the other hand, when the artifact under test is "good", i.e., true performance measure $p \geq 0.80$, the probability of correctly deeming the artifact as good varies across the tests of

different sample sizes. For example, consider a "good" artifact with true performance measure $p = 0.9$. From Fig. 2, we observe that the probability of deeming this artifact as good to be 0.27 when the test has $n = 25$. As the sample size of the test is increased, so too is the probability of deeming this artifact as good: 0.62 when $n = 50$, 0.88 when $n = 100$, and 1.00 when $n = 500$. The complement of these probabilities are the producer risks (type II error probabilities) associated with each of the different tests. We see that as the sample size *n* increases, the producer risk decreases. Thus, the experimenter must consider and balance the tradeoffs between increasing sample size (test cost) and decreasing producer risk.

With the performance threshold and acceptable consumer risk defined, the experimenter selects the test that satisfies the tradeoffs between test burden (sample size) and desired producer risk.

5.2.3. Acceptance Criterion

Each individual trial of a test will produce a success or a failure; e.g., a detection or a failure to detect. If the total number of successes observed during the entire test is greater than or equal to the predefined acceptance criterion, then the test artifact is deemed as "good".

The acceptance criterion is the smallest number of successes needed to be observed to deem the test artifact as "good".

We note that most statistics references, when discussing the topic of hypothesis testing, refer to the *rejection region*: the set of realized observations that will result in a rejection of the null hypothesis. Because the formulation of our null hypothesis assumes that the test artifact is "bad" (Sec. 5.1), a rejection of the null hypothesis results in an acceptance of the test artifact. Thus, for simplicity, we refer to the rejection of the null hypothesis as the *acceptance criterion*. The following section provides details on deriving an acceptance criterion for a fixed sample test.

6. Fixed Sample Test

The total number of trials and the acceptance criterion for a fixed sample test are determined prior to making any test observations and must remain fixed and unchanged throughout testing for the performance requirements of the test to be attained. In this section, we develop sample sizes and acceptance criterion to support a fixed sample hypothesis test constructed to confirm that a test artifact satisfies a performance requirement. We also illustrate how power curves, such as those displayed in Fig. 2, are generated.

Provided a performance threshold and statement of acceptable risk (or required confidence), there are many statistical methods that can be leveraged to define the parameters of a hypothesis test when observing binary response data. Because of its coverage properties, we

chose to implement the approach based upon the Clopper-Pearson "exact" method [3]. The exact method directly utilizes the definition of the binomial distribution provided in Eq. (1). See Agresti and Coull [4] for a presentation of the exact method and several additional applicable methods and their properties.

We begin by defining the following notation, most of which has been previously defined in this chapter:

p	test artifact's true but unknown performance measure
p^*	performance threshold
α	maximum acceptable risk (type I error probability)
n	sample size
c	acceptance criterion
X	total number of successes observed during the entire test

As stated in the Sec. 4, the first step in designing a defensible and successful test is defining the performance threshold, p^*, and stating the maximum acceptable risk, α. Because we view the performance threshold in this chapter as a lower bound, any test artifact with a true performance measure, *p*, that is greater than or equal to p^* is considered "good", otherwise, the test artifact is considered "bad".

We deem a test artifact as good if the total number of successes observed during the test, *X*, is greater than or equal to the acceptance criterion, *c*. From the definition of the binomial distribution, we can calculate the probability of deeming a test artifact with true performance measure *p* as good for any acceptance criterion, *c*, and sample size, *n*, using Eq. (2). That is, we calculate the probability that the number of successes observed, *X*, will be greater than or equal to the acceptance criterion, *c*, for a binomial random variable with sample size *n* and success probability *p*.

$$P\left(\text{deem artifact good}\right) = P\left(X \geq c \mid n, p\right) = 1 - \sum_{x=0}^{c-1} \binom{n}{x} p^x \left(1-p\right)^{n-x} \quad (2)$$

As an example, consider a test with performance threshold $p^* = 0.8$, maximum acceptable risk $\alpha = 0.05$, sample size $n = 20$, and acceptance criterion $c = 18$; we calculate the probability of deeming a test artifact as good with true performance measure $p = 0.7$ by:

$$P\left(\text{deem artifact good}\right) = P\left(X \geq 18 \mid n = 20, p = 0.7\right) = 1 - \sum_{x=0}^{17} \binom{20}{x} 0.7^x \left(1-0.7\right)^{20-x} = 0.035$$

Since the true state of this example artifact is "bad" (true performance measure $p = 0.7$ is less than performance threshold $p^* = 0.8$), we desire a low probability of deeming the artifact as good. This example calculation can be carried out for many different true performance values ranging from 0 to 1 as illustrated in Fig. 3. The results of these calculations provide the basis for the power curve.

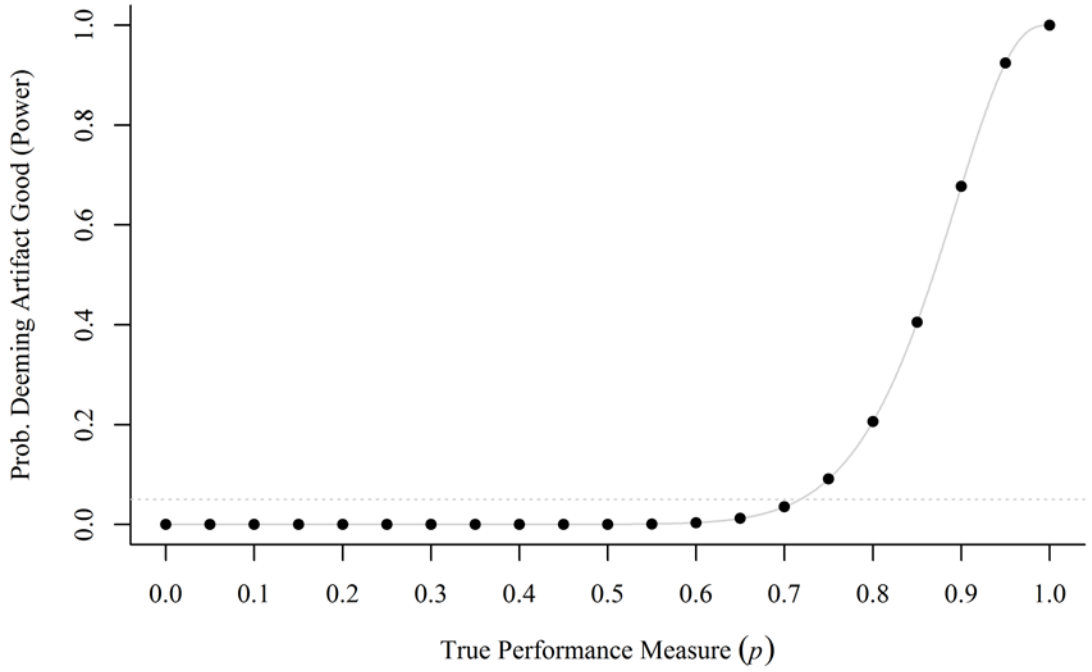

Fig. 3: Illustration of power curve construction for test with parameters n = 20, c = 18, and p = 0.8. The horizontal dotted line is maximum acceptable risk of $\alpha = 0.05$.*

Beyond illustrating the construction of the power curve, Fig. 3 highlights a problem with the underlying example. All artifacts with true performance measures less than the performance threshold $p^* = 0.8$ are defined as bad artifacts. We observe from Fig. 3 that the probability of deeming a truly bad artifact as good is as high as 0.206 (at $p = 0.8 - \varepsilon$, where ε is some very small, negligible value); this violates the stated maximum acceptable risk of $\alpha = 0.05$. To rectify this issue, either the sample size or the acceptance criterion – or both – must be altered. Increasing the sample size to $n = 22$ and the acceptance criterion to $c = 21$ resolves the issue in this example by providing a maximum probability of deeming a bad instrument as good of 0.048.

In practice, optimization routines can be used in conjunction with Eq. (2) to identify test parameters *n* and *c* that satisfy the stated maximum acceptable risk. An often-used strategy is to first investigate the minimum sample size test which occurs when no failures are allowed for acceptance of the artifact, i.e., $c = n$. From here the sample size is increased, with appropriate adjustments to the acceptance criterion to allow the type I error to be as large as possible without exceeding the stated maximum acceptable risk. The result of the increased sample size is a decrease in the producer risk (type II error) as was illustrated in Fig. 2. This exercise allows the experimenter to identify test parameters *n* and *c* that are of practical size, satisfy the stated maximum acceptable risk, and provide a producer risk that is satisfactory. Table 2 provides the required sample size, *n*, and number of allowable failures, $n - c$, for a range of performance thresholds and acceptable risk levels.

Table 2: Required sample size for stated performance threshold, acceptable risk (type I error) and maximum number of failures allowable to deem test artifact as good.

Performance Threshold	Acceptable Risk	Number of Allowable Failures										
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0.99	**0.01**	459	662	838	1001	1157	1307	1453	1596	1736	1874	2010
0.99	**0.05**	299	473	628	773	913	1049	1182	1312	1441	1568	1693
0.99	**0.10**	230	388	531	667	798	926	1051	1175	1297	1418	1538
0.99	**0.15**	189	337	471	600	726	848	969	1088	1206	1323	1439
0.99	**0.20**	161	299	427	551	671	790	906	1022	1137	1251	1364
0.95	**0.01**	90	130	165	198	229	259	288	316	344	371	398
0.95	**0.05**	59	93	124	153	181	208	234	260	286	311	336
0.95	**0.10**	45	77	105	132	158	184	209	234	258	282	306
0.95	**0.15**	37	67	94	119	144	169	193	216	240	263	286
0.95	**0.20**	32	59	85	110	134	157	180	204	226	249	272
0.90	**0.01**	44	64	81	97	113	127	142	156	170	183	197
0.90	**0.05**	29	46	61	76	89	103	116	129	142	154	167
0.90	**0.10**	22	38	52	65	78	91	104	116	128	140	152
0.90	**0.15**	19	33	46	59	72	84	96	107	119	131	142
0.90	**0.20**	16	29	42	54	66	78	90	101	113	124	135
0.85	**0.01**	29	42	53	64	74	84	93	103	112	121	130
0.85	**0.05**	19	30	40	50	59	68	76	85	93	102	110
0.85	**0.10**	15	25	34	43	52	60	68	77	85	93	100
0.85	**0.15**	12	22	31	39	47	55	63	71	79	87	94
0.85	**0.20**	10	19	28	36	44	52	59	67	75	82	90
0.80	**0.01**	21	31	39	47	55	62	69	76	83	89	96
0.80	**0.05**	14	22	30	37	44	50	57	63	69	76	82
0.80	**0.10**	11	18	25	32	38	45	51	57	63	69	75
0.80	**0.15**	9	16	23	29	35	41	47	53	59	65	70
0.80	**0.20**	8	14	21	27	33	39	44	50	56	61	67
0.75	**0.01**	17	24	31	37	43	49	54	60	65	70	76
0.75	**0.05**	11	18	23	29	34	40	45	50	55	60	65
0.75	**0.10**	9	15	20	25	30	35	40	45	50	55	59
0.75	**0.15**	7	13	18	23	28	33	37	42	47	51	56
0.75	**0.20**	6	11	16	21	26	31	35	40	44	49	53
0.70	**0.01**	13	20	25	30	35	40	44	49	53	58	62
0.70	**0.05**	9	14	19	24	28	33	37	41	45	49	53
0.70	**0.10**	7	12	16	21	25	29	33	37	41	45	49
0.70	**0.15**	6	10	15	19	23	27	31	35	39	42	46
0.70	**0.20**	5	9	14	18	21	25	29	33	37	40	44
0.60	**0.01**	10	14	18	22	25	29	32	36	39	42	45
0.60	**0.05**	6	10	14	17	21	24	27	30	33	36	39
0.60	**0.10**	5	9	12	15	18	21	24	27	30	33	36
0.60	**0.15**	4	8	11	14	17	20	23	26	28	31	34
0.60	**0.20**	4	7	10	13	16	19	22	24	27	30	33
0.50	**0.01**	7	11	14	17	19	22	25	27	30	33	35
0.50	**0.05**	5	8	11	13	16	18	21	23	26	28	30
0.50	**0.10**	4	7	9	12	14	17	19	21	24	26	28
0.50	**0.15**	3	6	8	11	13	16	18	20	22	25	27
0.50	**0.20**	3	5	8	10	12	15	17	19	21	24	26

An experimenter uses Table 2 by identifying the row that corresponds to the stated performance threshold and acceptable risk. Within that row, the first column in the main body of the table is the number of samples required if the acceptance criterion were such that no failures were to be allowed, i.e., $c = n$. As one moves across the row in the main body of the table, the required sample size increases as the number of allowable failures also increases. This increase in sample size reduces the producer risk (type II error). We remind the reader that under this

fixed sample test approach, the sample size and acceptance criterion must be selected prior to the beginning of test execution and cannot be revised during testing for the performance requirements of the test to be attained.

Consider an experiment with a stated performance threshold of $p^* = 0.85$ and a maximum acceptable risk of $\alpha = 0.01$. From Table 2, we see that the experimenter could choose to perform a test with as few as $n = 29$ trials, though the test artifact would be deemed as good only if all 29 trials resulted in successes. If the experimenter wished to increase the sample size, increase the number of allowable failures, and decrease the producer risk, he or she could do so by increasing the sample size to $n = 42$ and allow for one failure, or $n = 53$ with two failures, or $n = 64$ with three failures, and so on. The power curves associated with this family of potential tests, which satisfy a stated performance threshold of $p^* = 0.85$ and a maximum acceptable risk of $\alpha = 0.01$ are displayed in Fig. 4. From such a figure, the experimenter can view the benefit gained in producer risk by increasing the sample size.

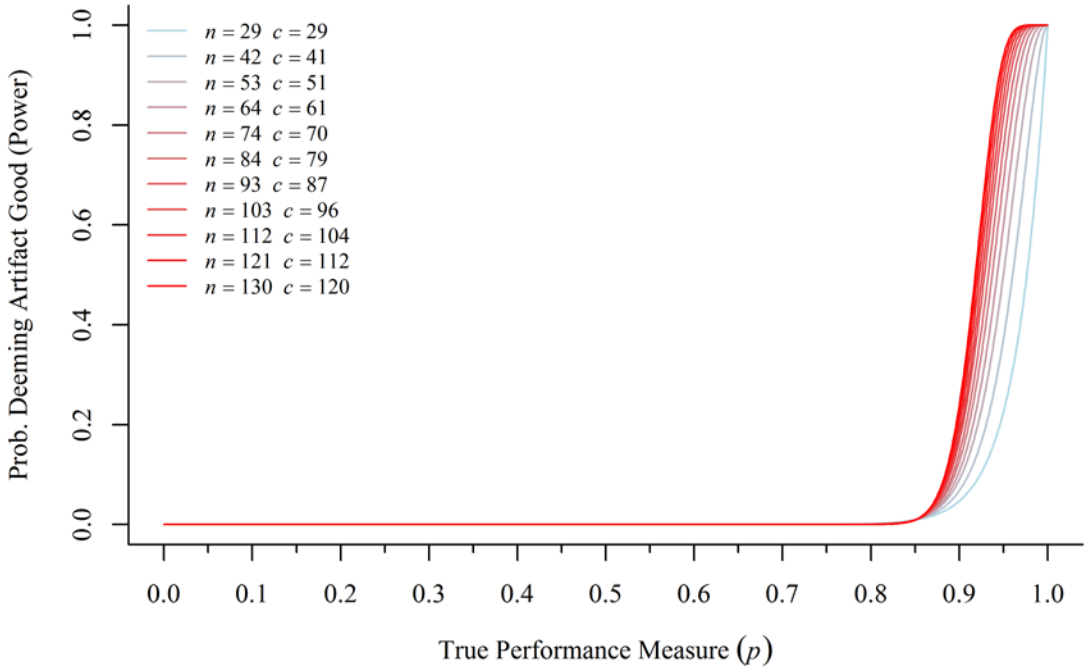

Fig. 4: Power curves for family of tests that satisfy a stated performance threshold of $p^* = 0.85$ and maximum acceptable risk of $\alpha = 0.01$.

7. Sequential Sampling Test

In Sec. 6 we presented an approach to develop a test for which the total number of trials and the acceptance criterion are determined prior to making any test observations and these remain fixed and unchanged throughout testing, i.e., a fixed sample test. In sequential sample testing, test observations are collected in batches with the outcomes of each batch immediately analyzed and a decision made based on the information gathered thus far in the experiment. The

possible decisions after each batch of data is collected are to either 1.) stop testing and deem the test artifact as good, 2.) stop testing and do not deem the test artifact as good, or 3.) continue testing and collect a subsequent batch of observations. The advantage of a sequential sampling test, or simply, *sequential test*, is that in certain situations the total number of observations required to confirm a performance threshold may be substantially fewer than required in a fixed sample test.

Beginning with the foundational work of Abraham Wald [5], there exists extensive literature on the topic of sequential testing. In this section, we introduce the concept of sequential testing through an example and offer several words of caution in implementation. However, due to the complexity and subtleties involved, we strongly encourage seeking guidance from a qualified party when designing and implementing a sequential test.

7.1. Structure

Fig. 5 provides a schematic of a three-phased sequential test that seeks to confirm that a test artifact satisfies a performance threshold where the response variable observed is binary. Each phase of the test consists of a batch of twelve binary observations, e.g., success or failures, for a maximum of 36 total observations. Beginning with phase 1 at the top of Fig. 5, $n_1 = 12$ binary observations are made, and the number of observed successes, X_1, are noted. If all twelve observations resulted in a success, i.e., $X_1 = 12$, then testing ceases and we deem the test artifact as good, or acceptable. This is denoted in Fig. 5 as the arrow labeled $X_1 = 12$ terminating at the circle labeled "A". If nine, ten, or eleven successes are observed in phase 1, i.e., $X_1 = 9$, $X_1 = 10$, or $X_1 = 11$, then testing continues with phase 2 where another batch of twelve binary observations are collected. If eight or fewer successes are observed in phase 1, i.e., $X_1 \leq 8$, then testing ceases and we conclude that we cannot deem the test artifact as good (for simplicity, these terminal paths are not displayed on the schematic). This same interpretation of the schematic follows for phases two and three where in phase three no further testing ensues, and the test artifact is either deemed good or not.

Before examining the probabilistic properties of the sequential test represented by Fig. 5, we make some general comments about its structure. First, there are three phases and in each of the first two phases testing may either cease or it may continue. The third phase is terminal, regardless of the outcome. There is nothing particular about the number of phases and in general, they may be chosen as a practical matter. Second, we note that in each phase the batch size is equal to twelve. Again, there is nothing particular about the size of the batches, and in fact they could be a small as a single observation and the batch sizes do not need to be equivalent across phases. As we will examine next, the constraints on batch size, number of phases, and terminal nodes are a result of the desired probabilistic properties. More specifically, like the fixed sample test, the probabilistic properties and resulting structure of the sequential test follow directly from a stated testable objective that must include a performance threshold and required level of confidence or acceptable risk.

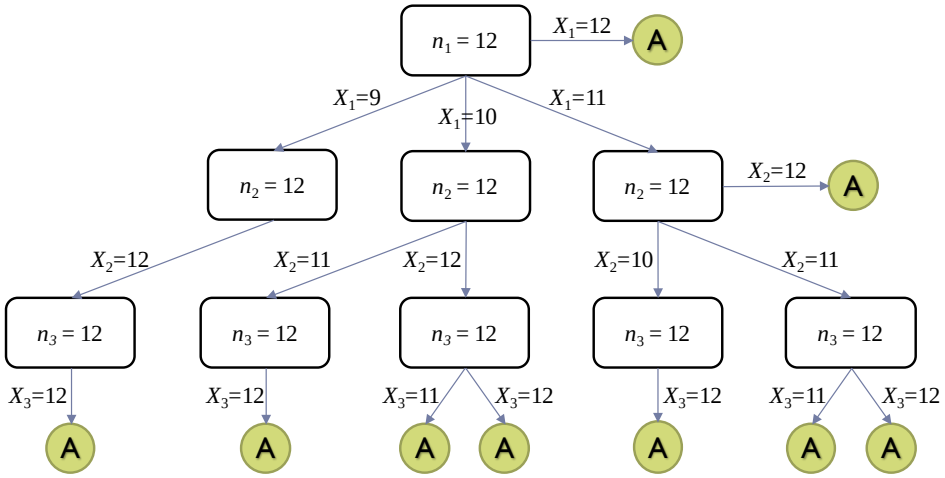

Fig. 5: Schematic of a three-phased sequential test. Each phase i = 1, 2, 3 consists of n_i = 12 observations with binary responses. The random variable X_i is the number of successes observed within the phase. The paths leading to an "A" are terminal, ending in deeming the artifact as good or acceptable. All other possible values for X_i that are not shown are terminal paths failing to deem the artifact as good.

7.2. Probabilistic Properties

We observe in Fig. 5 that a test artifact is deemed as good only if three or fewer failures are observed throughout the entire sequential test which may include up to 36 total observations. In comparing this to a similar sized fixed sample test, we note from Table 2 that a fixed sample test with a total sample size of 36 that allows for three failures supports a test objective with a performance threshold of $p^* = 0.85$ and a maximum acceptable risk of $\alpha = 0.20$. Can the sequential test in Fig. 5 support this same test objective?

Recall from the Sec. 6 that the maximum risk of a defined test is found by calculating the maximum probability of deeming a test artifact as good when it is truly bad, and that this maximum occurs at $p^* - \varepsilon$, where ε is some very small, negligible value. So, for the fixed sample test with performance threshold $p^* = 0.85$, sample size $n = 36$, and acceptance criterion $c = 33$ (three allowable failures), the maximum probability of deeming a test artifact as good when it is truly bad is given by:

$$P(\text{deem artifact good}) = P(X \geq 33 \mid n = 36, p = 0.85) = 1 - \sum_{x=0}^{32} \binom{36}{x} 0.85^x (1 - 0.85)^{36-x} = 0.191$$

which satisfies the stated maximum acceptable risk of $\alpha = 0.20$.

In a similar, but slightly more complicated manner, we can calculate the maximum probability of deeming a test artifact as good when it is truly bad for the sequential test of Fig. 5. We note that in the sequential test of Fig. 5 there are nine different ways that we will deem a test artifact as good, e.g., when 12 successes are observed in the first batch of 12 trials, or when 11 successes are observed in the first batch of 12 trials and 12 successes are observed in the second batch of 12 trials, and so on. Each of the nine paths to deeming a test artifact as good must be

captured in our probability calculation. So, with a performance threshold $p^* = 0.85$, the sequential test of Fig. 5 provides a maximum probability of deeming a test artifact as good when it is truly bad of:

$$
\begin{aligned}
P(\text{deem artifact good}) = {} & P(X_1 = 12 \mid n_1 = 12, p = 0.85) + \\
& P(X_1 = 11 \mid n_1 = 12, p = 0.85) * P(X_2 = 12 \mid n_2 = 12, p = 0.85) + \\
& P(X_1 = 11 \mid n_1 = 12, p = 0.85) * P(X_2 = 11 \mid n_2 = 12, p = 0.85) * P(X_3 = 12 \mid n_3 = 12, p = 0.85) + \\
& P(X_1 = 11 \mid n_1 = 12, p = 0.85) * P(X_2 = 11 \mid n_2 = 12, p = 0.85) * P(X_3 = 11 \mid n_3 = 12, p = 0.85) + \\
& P(X_1 = 11 \mid n_1 = 12, p = 0.85) * P(X_2 = 10 \mid n_2 = 12, p = 0.85) * P(X_3 = 12 \mid n_3 = 12, p = 0.85) + \\
& P(X_1 = 10 \mid n_1 = 12, p = 0.85) * P(X_2 = 12 \mid n_2 = 12, p = 0.85) * P(X_3 = 12 \mid n_3 = 12, p = 0.85) + \\
& P(X_1 = 10 \mid n_1 = 12, p = 0.85) * P(X_2 = 12 \mid n_2 = 12, p = 0.85) * P(X_3 = 11 \mid n_3 = 12, p = 0.85) + \\
& P(X_1 = 10 \mid n_1 = 12, p = 0.85) * P(X_2 = 11 \mid n_2 = 12, p = 0.85) * P(X_3 = 12 \mid n_3 = 12, p = 0.85) + \\
& P(X_1 = 9 \mid n_1 = 12, p = 0.85) * P(X_2 = 12 \mid n_2 = 12, p = 0.85) * P(X_3 = 12 \mid n_3 = 12, p = 0.85) \\
& = 0.272
\end{aligned}
$$

And hence, the sequential test of Fig. 5 cannot support the maximum acceptable risk of $\alpha = 0.20$ test objective.

A sequential sample test has a higher risk than a fixed sample test that allows for the same number of failures over the same total observations.

As we did in Sec. 6, we can calculate the probability of deeming a test artifact as good over the range of the true performance measure, *p*, to develop the power curve for this sequential test. In Fig. 6 we provide the power curves for the sequential test and the fixed sample test with a total sample size of 36 that allows for three failures. Also provided in Fig. 6 is a power curve for a fixed sample test with a total sample size of 32 that allows for three failures. This latter fixed sample test provides a power curve that is most similar to that of the sequential test over all possible fixed sample tests.

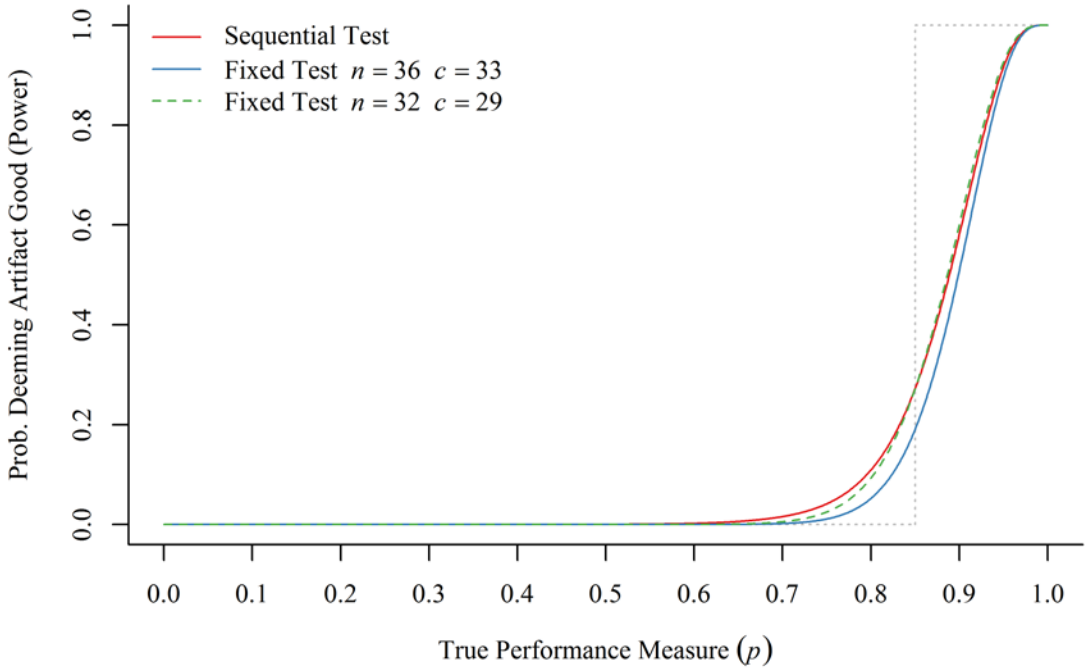

Fig. 6: Power curves for the sequential test and two fixed sample tests with sample sizes of 36 and 32 that allow for three failures.

From the previous calculations, we observed that a fixed sample test with a total sample size of 36 that allows for three failures does support the test objective with a performance threshold of $p^* = 0.85$ and a maximum acceptable risk of $\alpha = 0.20$. However, the sequential test of Fig. 5 that allows for up to three failures in a total of 36 observations cannot support the same test objective as the maximum risk associated with the sequential test is 0.272. This example illustrates the caution necessary when executing a fixed sample test in a sequential manner. If observations of a fixed sample test are revealed sequentially and it is noted at any time during testing that the number of allowable failures has been exceeded, then testing should cease as to conserve resources. Otherwise, the total number of trials and the acceptance criterion of a fixed sample test must remain fixed and unchanged throughout testing. Like the fixed sample test, designing a sequential test must begin with a test requirement that contains both a performance threshold and a statement of acceptable risk. The parameters of the sequential test which include the number of phases, the batch sizes, and the acceptance criteria at each phase can then be formulated to satisfy the stated test objective.

Like the fixed sample test, a sequential test must begin with a testable objective that contains both a performance threshold and a statement of acceptable risk.

7.3. Sample Size

Though there is a significant increase in the complexity involved with properly designing a sequential test to confirm a performance threshold with a binary experimental response, the tradeoff is the reduction in the expected number of observations required. Intuitively, from the sequential test illustrated by Fig. 5, we see that a very poor performing test artifact would almost surely end testing after only 12 observations as it is unlikely that the results provided in the first phase would satisfy the requirement to continue to phase two. More specifically, for any given test artifact with a true performance measure, *p*, we can calculate its expected number of test observations in each phase by the multiplying the probability that it enters the phase (using Eq. (1)) by the number of trials in the phase (in this case, 12). For example, the probability that a test artifact with a true performance measure $p = 0.65$ enters phase two of the sequential test of Fig. 5, is given by:

$$\begin{aligned} P(\text{enter phase 2}) = {} & P(X_1 = 9 \mid n_1 = 12, p = 0.65) + \\ & P(X_1 = 10 \mid n_1 = 12, p = 0.65) + \\ & P(X_1 = 11 \mid n_1 = 12, p = 0.65) \\ = {} & 0.341 \end{aligned}$$

and thus, its expected number of observations in phase two is $0.341 \times 12 = 4.09$. The overall expected number of test observations is the sum of the expected number of observations in each phase. Continuing with the example of the test artifact with a true performance measure $p = 0.65$, the overall expected number of observations is given by:

$$\begin{aligned} E(obs \mid p = 0.65) &= P(\text{enter phase 1}) \times n_{phase1} + P(\text{enter phase 2}) \times n_{phase2} + P(\text{enter phase 3}) \times n_{phase3} \\ &= 1 \times 12 + 0.341 \times 12 + 0.011 \times 12 \\ &= 16.22 \end{aligned}$$

The expected number of required test observations over the range of the true performance measure values, *p*, for the sequential test illustrated by Fig. 5 is provided in Fig. 7. We see that when the test artifact's true performance measure is less than 0.5, then we expect to require only 12 observations. The expected number of observations then increases as the true performance measure increases, until it reaches its maximum value of 25.6 observations at the true performance measure $p = 0.876$. The expected number of required observations then rapidly shrinks back to 12 as the true performance measure approaches one. Note that the maximum expected number of observations, 25.6, is less than the 32-observation fixed test that provided a power profile most similar to this sequential test.

Though a sequential test is more complicated to properly design, considerable savings may be provided through a reduction in the expected number of observations required.

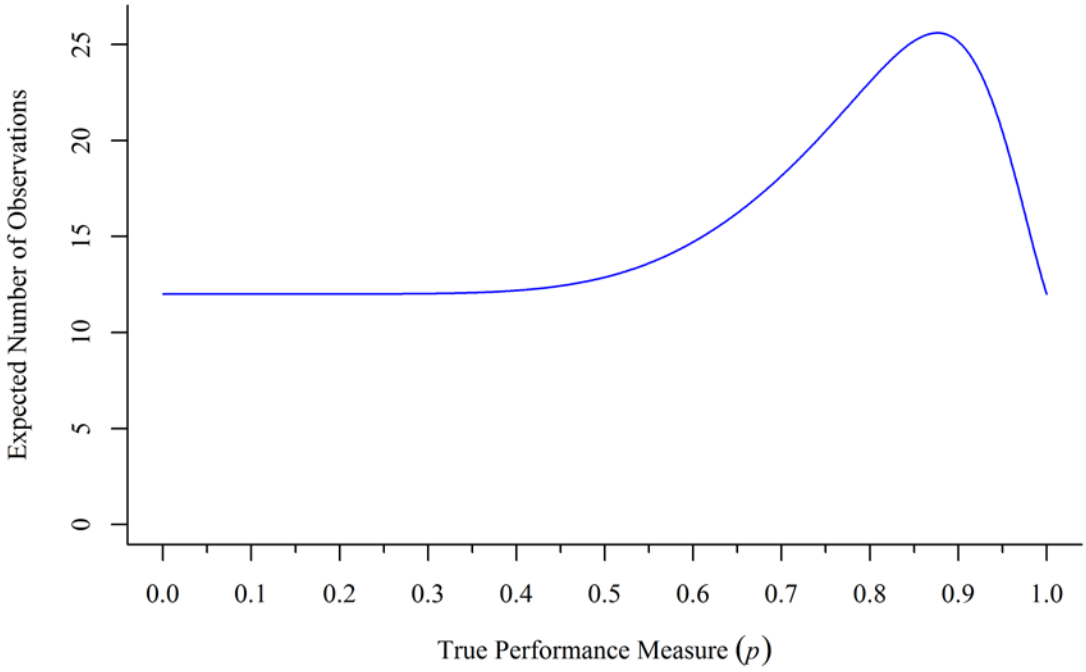

Fig. 7: Expected number of observations for the sequential test of Fig. 5.

Acknowledgments

This work was sponsored by the U. S. Department of Homeland Security. The authors would like to thank our colleagues, Ryan Fitzgerald, William Guthrie, and James Yen, for their valuable input that improved this work.

Acronyms

ANSI	American National Standard Institute
BRD	Backpack-type radiation detector
CBP	Customs and Border Protection
DHS	Department of Homeland Security
IEC	International Electrotechnical Commission
PRD	Personal radiation detector
RIID	Radioisotope identification device
RPM	Radiation portal monitor
SPRD	Spectrometric personal radiation detector
SRPM	Spectrometric radiation portal monitor
TCS	Technical Capability Standard

References

[1] G. Casella and R. L. Berger, Statistical Inference, Second ed., Pacific Grove, CA: Duxbury, 2002.

[2] D. C. Montgomery and G. C. Runger, Applied Statistics and Probability for Engineers, Sixth ed., Hoboken, NJ: John Wiley & Sons, Inc., 2014.

[3] C. J. Clopper and E. S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika,* vol. 26, pp. 404-413, 1934.

[4] A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician,* vol. 52, no. 2, pp. 119-126, 1998.

[5] A. Wald, "Sequential Tests of Statistical Hypotheses," *The Annals of Mathematical Statistics,* vol. 16, no. 2, pp. 117-186, 1945.