NIST Technical Note 2118

False Alarm Testing for Radiation Detection Systems

Dennis D. Leber Leticia Pibida

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2118



NIST Technical Note 2118

False Alarm Testing for Radiation Detection Systems

Dennis D. Leber Statistical Engineering Division Information Technology Laboratory

Leticia Pibida Radiation Physics Division Physical Measurement Laboratory

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2118

October 2020



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 Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

National Institute of Standards and Technology Technical Note 2118 Natl. Inst. Stand. Technol. Tech. Note 2118, 25 pages (October 2020) CODEN: NTNOEF

> This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2118

Abstract

An operator of a radiation detection system that displays a high rate of false alarms may become desensitized to these alarms. This action is known as alarm fatigue and can have detrimental results as the operator may cease to respond to what may be valid alarms. To mitigate these adverse outcomes, an agency may seek to procure radiation detection systems with adequately low false alarm rates. In developing a test to confirm that the system satisfies the stated false alarm rate threshold requirement, the number of required observations or test duration is often an initial question. In this chapter, we demonstrate how an experimenter can develop a successful false alarm test with two provided pieces of information: the false alarm rate threshold requirement and a statement of acceptable risk or required confidence. Using the statistical hypothesis testing framework, we illustrate the meaning of risk and confidence from both the consumer's and producer's perspectives and provide guidance on selecting an informed false alarm rate threshold requirement and statement of acceptable risk. We consider the binomial and Poisson probability models that apply to testing of radiation detection systems that are employed with and without occupancy sensors, respectively. From these probability models 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 sample size and acceptance criterion tables to define fixed sample tests that satisfy a variety of false alarm thresholds and levels of acceptable risk for systems with and without occupancy sensors.

Keywords

False alarm; occupancy sensor; binomial random variable; Poisson random variable; hypothesis test; consumer and producer risks; false alarm rate threshold; power of a test; radiation detection systems.

Table of Contents

1.	Introduction	1						
2.	Choosing a False Alarm Rate Requirement	1						
3.	Stating the Test Requirement	3						
4.	Hypothesis Tests	3						
4	.1. The Null Hypothesis	4						
4	.2. Errors in Hypothesis Testing	4						
	4.2.1. Consumer and Producer Risks	4						
	4.2.2. Power of a Test	5						
	4.2.3. Acceptance Criterion	8						
5.	Systems with Occupancy Sensors	8						
5	.1. Binomial Probability Distribution	9						
5	.2. Power and Sample Size	9						
6.	Systems without Occupancy Sensors 1	3						
6	.1. Poisson Probability Distribution 1	4						
6	.2. Power and Sample Size 1	4						
7.	Summary 1	8						
Acl	Acknowledgments							
Ref	References 19							

List of Tables

Table 1: ANSI standards for homeland security application	3
Table 2: Hypothesis test truth table.	5
Table 3: Systems with occupancy sensors: required occupancies for stated false alarm	
threshold, acceptable risk (type I error) and maximum number of false alarms allowable to	
deem system as good.	12
Table 4: Systems without occupancy sensors: required test duration (e.g., hours) for stated	
false alarm rate requirement, acceptable risk (type I error) and maximum number of false	
alarms allowable to deem system as good.	17

List of Figures

1. Introduction

There are several types of radiation detections systems used for homeland security applications. These systems include personal radiation detectors and spectroscopic personal radiation detectors (PRDs and SPRDs), hand-held systems, radioisotope identification devices (RIIDs), backpack-type radiation detectors (BRDs), mobile systems, radiation portal monitors and spectroscopic radiation portal monitors (RPMs and SRPMs) and neutron detectors. Two primary metrics are used in quantifying the performance of these systems: the ability of the system to detect a source when a source is present (probability of detection), and the system's tendency to alarm in the absence of a source (probability of false alarm).

As the consequences of alarm fatigue have gained the attention of the healthcare community (Mitka, 2013), the impact of false alarms must also be considered in homeland security applications. Just as a high number of false alarms may lead a healthcare clinician to ignore a valid alarm, so too may be the case for a homeland security official who dons a radiation detection system with a large false alarm rate.

An experimenter may wish to confirm that a system under consideration provides a false alarm rate less than some predefined criterion, e.g., less than 1 false alarm in 1000 encounters. While the answer to this inquiry is a simple yes or no, there is a possibility of answering the question incorrectly because of the inherent uncertainty in the measurements used in the assessment (e.g., random behavior of nuclear decay process). This chapter provides guidance on developing an experimental sample size and acceptance criterion to determine whether a system satisfies a predefined false alarm performance criterion. Separately, for systems with and without occupancy sensors, we provide a sample size and acceptance criterion table to define a fixed sample test that will satisfy a variety of false alarm thresholds and levels of acceptable risk. Because false alarm testing entails the confirmation of a performance threshold where the threshold is taken as an upper bound, the concepts presented in this chapter parallel those presented by Leber, Pibida, and Enders (2019). For completeness, many of those ideas are repeated here.

2. Choosing a False Alarm Rate Requirement

A defensible and successful test always begins with a testable objective. A test to determine if a system satisfies a false alarm criterion will consist of two components:

- 1. a false alarm rate requirement, or more simply, a false alarm threshold; and
- 2. a statement of acceptable risk or required confidence.

Together, the defined false alarm threshold and statement of acceptable risk will lead directly to the required number of trials (or test duration) and acceptance criterion. If the number of trials required to support the false alarm 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 higher false alarm rate or assumes a higher level of risk than desired. This section presents a philosophical view on setting a false alarm threshold. A description of confidence and risk, and guidance on selecting an acceptable risk are presented in Section 4.2.

A defensible and successful test begins with a testable objective that includes a false alarm 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, false alarm threshold requirements may be directed by user needs, standard requirements, or acquisition requirements. When formulating false alarm threshold requirements based on user needs, one should consider the consequence of a false alarm. For example, a false alarm of a portal monitor during cargo inspection at a border crossing may result in major, unnecessary disruption of commerce flow and inspector effort due to the need to perform secondary screening or dismantle a cargo container full of goods. Even when the immediate consequence of a false alarm is minimal, the long-term toll of a high false alarm rate must be considered. Alarm fatigue, the desensitization to alarms, may lead to homeland security officials ignoring true alarms, the result of which may have detrimental impacts. This notion has been well studied in the medical community where patient deaths have been attributed to the behavior (Sendelbach & Funk, 2013). The radiation detection community may gain valuable insights by leveraging this knowledge base when setting false alarm threshold requirements.

The drive to set minimal false alarm threshold requirements must be balanced with the radiation detection capability of each type of technology; there is a tradeoff between the instrument detection capability and false alarm rate. Often, though not universally, a radiation detection system's ability to detect a source when a source is present is inversely related to its probability of a false alarm. That is, a system that provides a desirable false alarm rate (rarely alarms in the absence of a source) may also provide an undesirable true alarm rate (rarely alarm in the presence of a source). This tradeoff between false alarms and true alarms plays an everimportant role in defining performance requirements for a radiation detection system.

A thorough understanding of a user's practical operational requirements can help facilitate the definition of false alarm threshold requirements. Based on the effort required to adjudicate an alarm, one might consider the number of false alarms a user could handle during an eight hour work shift, for example. This requirement may be phrased in terms of a rate relative to a unit of time, e.g., no more than three false alarm per eight hours, or in terms of a frequency of occurrences relative to total inspections, e.g., no greater than one false alarm in every 1000 inspections.

Several American National Standard Institute (ANSI) standards have been developed for radiation detectors used for homeland security applications (Table 1). Most of these standards have requirements for the systems' probability of false alarms. A common requirement in these standards in terms of occurrences is to have a probability of false alarm no greater than one alarm in 1000 occupancies. In terms of a rate per unit of time, a common requirement is to have a false alarm rate no greater than one alarm per two hours.

ANSI Standard	Standard Title
ANSI N42.32	Performance Criteria for Alarming Personal Radiation Detectors for Homeland
	Security
ANSI N42.33	Portable Radiation Detection Instrumentation for Homeland Security
ANSI N42.34	Performance Criteria for Hand-held Instruments for the Detection and Identifica-
	tion for Radionuclides
ANSI N42.35	Evaluation and Performance of Radiation Detection Portal Monitors
ANSI N42.38	Performance Criteria for Spectroscopy-Based Portal Monitors Used for Homeland
	Security
ANSI N42.43	Performance Criteria for Mobile and Transportable Radiation Monitors Used for
	Homeland Security
ANSI N42.48	Performance Requirements for Spectroscopic Personal Radiation Detectors
	(SPRDs) for Homeland Security
ANSI N42.53	Performance Criteria for Backpack-Based Radiation-Detection Systems Used for
	Homeland Security

Table 1: ANSI standards for homeland security application

3. 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 needed to prove that a system satisfies a false alarm threshold. For this purpose, a testable requirement has two key parts: 1.) a false alarm threshold, and 2.) a statement of acceptable risk or required confidence. For example, *the radiation detection system shall provide a false alarm rate no greater than 0.001 (one alarm in 1000 occupancies) with 95 % confidence*, is a testable requirement. In this example, one alarm in 1000 occupancies is the false alarm 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 false alarm threshold and a statement of acceptable risk (or required confidence).

We are interested in drawing a conclusion about the true value of the system's false alarm rate, but all that we have available is an uncertain estimate 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.

4. Hypothesis Tests

One statistical method that may be used to support the task of confirming that a system meets a specified false alarm threshold is the hypothesis test (Montgomery & Runger, 2014). Hypothesis testing begins with a specific conjecture called the *null hypothesis*. Data are 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 system's false alarm rate satisfies a specified false alarm threshold.

4.1. The Null Hypothesis

The true state of a system falls into one of two categories when its true false alarm rate is compared to a specified false alarm threshold; that is, the true false alarm rate is less than (or equal to) the specified false alarm threshold, or it is not. We label a system as "good" if the true false alarm rate is less than the specified false alarm threshold, and "bad" otherwise. Therefore, two possible positions exist for the null hypothesis conjecture: 1) the system is good; or, 2) the system 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 system is good, we adopt as the null hypothesis that the system is bad. For example, if we seek to prove that a radiation detection system satisfies the stated false alarm rate threshold of 0.001, then we state the null hypothesis as *the radiation detection system's true false alarm rate is greater than 0.001*.

Because the hypothesis test relies on the idea of proof by contradiction, we adopt as the null hypothesis conjecture that the system 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 system's false alarm rate satisfies the false alarm threshold, i.e., the system is good. Failure to reject the null hypothesis is not evidence that the system is bad, but rather that insufficient evidence was found to support the conclusion that the system meets the specified false alarm threshold; that is, we fail to deem the system as good.

4.2. Errors in Hypothesis Testing

A system has a true but unknown false alarm rate. It follows that the system has a true but unknown state, either "good" or "bad", as would be determined by comparing its true false alarm rate to the stated false alarm threshold. The statistical hypothesis test provides a framework for an experimenter to deem a system as "good", based on an estimate of the system's false alarm rate. Because the estimated false alarm rate is uncertain (all measurements carry uncertainty), our conclusion about the true state of the system 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.

4.2.1. 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 system to be "good" when in fact, the system'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 $1 - \alpha$ 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 system is truly "good," but we fail to deem the system as "good." Statisticians refer to this as a *type II error* and denote the probability of its occurrence with the Greek letter β . These errors are illustrated in the truth table displayed in Table 2.

		System's True State					
		"Good"	"Bad"				
ssis Test	Deem	Correct Decision	Type I				
usion	"Good"		Error				
Hypothe	Fail to deem	Type II	Correct Decision				
Concl	"Good"	Error					

Table 2: Hypothesis test truth table.

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, if a radiation detection system is to be used to monitor cargo at a border crossing where false alarm events trigger a significant search effort and disruption to commerce flow, the consumer of this radiation detection system, e.g., the U.S. Customs and Border Protection (CBP), will seek to avoid purchasing and deploying a system with a false alarm rate larger than specified in the purchase agreement. 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 a 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*. When referring to consumer risk in this manuscript, we simply use *risk* and when discussing producer risk, we spell out the term.

4.2.2. 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 system as "good" as a function of the system's true but unknown false alarm rate.

An ideal test would deem a system as "good" with certainty (i.e., a probability of one) when the system's true false alarm rate is less than the false alarm threshold and never deem a system as "good" when the system's true false alarm rate is larger than the false alarm threshold. Fig. 1 provides a power curve for this ideal test when the false alarm threshold is $p^* = 0.1$.

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 Section 3, this statement of maximum acceptable consumer risk, paired with the false alarm threshold provides the necessary basis for constructing the hypothesis test.



Fig. 1: Power curve for an ideal test with a false alarm threshold $p^* = 0.1$.

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 a system will provide more errant alarms, resulting in additional alarm adjudication actions for the operator. 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" system 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.

Fig. 2 illustrates power curves for the ideal test $(n = \infty)$ and tests of sample size n = 50,100,250,500, each with a consumer risk (type I error probability) no greater than 0.05 and a false alarm threshold, $p^* = 0.1$. We first observe that for the limited sample tests when $n \neq \infty$, the power to the right of the false alarm threshold is similar. That is, for each of these tests, when the system under test has a true false alarm rate p > 0.1, i.e., a "bad" system, the probability of deeming the system 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) $\alpha = 0.05$ and a false alarm threshold $p^* = 0.1$.

Conversely, when the system under test is "good", i.e., true false alarm rate $p \le 0.1$, the probability of correctly deeming the system as good varies across the tests of different sample sizes. For example, consider a "good" system with true false alarm rate p = 0.05. From Fig. 2, we observe that the probability of deeming this system as good to be 0.28 when the test has n = 50. As the sample size of the test is increased, so too is the probability of deeming this system as good: 0.44 when n = 100, 0.88 when n = 250, 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 false alarm threshold and acceptable consumer risk defined, the experimenter selects the test that satisfies the tradeoffs between test burden (sample size) and desired producer risk.

4.2.3. Acceptance Criterion

Each individual trial of a test will produce a false alarm or not. If the total number of false alarms observed during the entire test is less than or equal to the predefined acceptance criterion, then the system is deemed as "good".

The acceptance criterion is the largest number of false alarms allowed to be observed to deem the system 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 system is "bad" (Section 4.1), a rejection of the null hypothesis results in an acceptance of the system. Thus, for simplicity, we refer to the rejection of the null hypothesis as the *acceptance criterion*. The following sections provide details on deriving an acceptance criterion for systems with and without occupancy sensors.

5. Systems with Occupancy Sensors

Some radiation detection systems used for homeland security applications may be equipped with occupancy sensors (e.g., RPMs, SRPMs) that inform the system when a vehicle, package or person is within the detection zone. The way the false alarm test is designed is dependent on whether the system is equipped with an occupancy sensor or not. For systems with occupancy sensors, the probability of a false alarm is estimated based on the binomial probability distribution. That is, each occupancy is viewed as an independent Bernoulli trial with some true false alarm rate, p, that we estimate by the ratio of the number of false alarms observed, x, to the total number of occupancies considered, n, i.e., $\hat{p} = \frac{x}{n}$. We then use the binomial

probability distribution to guide the quantification of our uncertainty in our estimated false alarm rate, \hat{p} .

In this section we develop a fixed sample test for radiation detection systems equipped with occupancy sensors to determine if the system's true but unknown false alarm probability, p, is less than some predefined, fixed false alarm threshold that we denote by p^* . 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. We are only concerned with a one-sided test, that is, investigating $p \le p^*$. We provide guidance for determining the sample size, *n*, and the acceptance criterion to prove that a system satisfies the false alarm threshold. We also illustrate how power curves, such as those displayed in Fig. 2, are generated.

5.1. Binomial Probability Distribution

Experiments with two, and only two possible outcomes, such as head and tail, defective and non-defective, or alarm and no alarm are known as Bernoulli trials (Casella & Berger, 2002). The probability of one of the two outcomes (e.g., "alarm") is denoted by p, while the probability of the complementary outcome ("no alarm") is given by 1 - p.

The total number of events observed, X (e.g., alarms), in a sequence of independent and identical Bernoulli trials is distributed as a binomial random variable. The binomial probability 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 binomial distribution, as described by Casella and Berger (2002), is defined in Eq. (1).

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

When a radiation detection system is equipped with an occupancy sensor, each occupancy with no source present is viewed as a Bernoulli trial. The number of false alarms observed in a sequence of such occupancies is modeled as a binomial random variable.

5.2. Power and Sample Size

Provided a false alarm 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 (Clopper & Pearson, 1934). The exact method directly utilizes the definition of the binomial distribution provided in Eq. (1). See Agresti and Coull (1998) 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* system's true but unknown false alarm rate
- p^* false alarm threshold
- α maximum acceptable risk (type I error probability)
- *n* sample size
- *c* acceptance criterion
- *X* total number of false alarms observed during the entire test

As stated in Section 3, the first step in designing a defensible and successful false alarm test is defining the false alarm threshold, p^* , and stating the maximum acceptable risk, α . Because we view the false alarm threshold in this chapter as an upper bound, any system with a true false alarm rate, p, that is less than or equal to p^* is considered "good"; otherwise, the system is considered "bad".

We deem a system as good if the total number of false alarms observed during the test, X, is less than or equal to the acceptance criterion, c. From the definition of the binomial distribution (Eq. (2)), we can calculate the probability of deeming a system with true false alarm rate p as good for any acceptance criterion, c, and sample size, n. That is, we calculate the probability that the number of false alarms, X, will be less than or equal to the acceptance criterion, c, for a binomial random variable with sample size n and true false alarm probability p.

$$P(\text{deem system good}) = P(X \le c \mid n, p) = \sum_{x=0}^{c} {n \choose x} p^{x} (1-p)^{n-x}$$
(2)

As an example, consider a test with false alarm threshold $p^* = 0.1$, maximum acceptable risk $\alpha = 0.05$, sample size n = 30, and acceptance criterion c = 1; we calculate the probability of deeming a system as good with true false alarm rate p = 0.2 by:

$$P(\text{deem system good}) = P(X \le 1 \mid n = 30, p = 0.2) = \sum_{x=0}^{1} {30 \choose x} 0.2^x (1 - 0.2)^{30 - x} = 0.0105$$

Since the true state of this example system is "bad" (true false alarm rate p = 0.2 is greater than the false alarm threshold $p^* = 0.1$), we desire a low probability of deeming the system as good. This example calculation can be carried out for many different true false alarm 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: Power curves for test with parameters n = 30, c = 1, and $p^* = 0.1$ (black points and curve) and n = 46, c = 1, and $p^* = 0.1$ (blue curve). The horizontal dashed 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 systems with true false alarm rates greater than the false alarm threshold $p^* = 0.1$ are defined as bad systems. We observe from Fig. 3 that the probability of deeming a truly bad system as good is as high as 0.184 (at $p = 0.1 + \varepsilon$, 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 = 46 resolves the issue in this example by providing a maximum probability of deeming a bad system 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 identify the minimum sample size test which occurs when no false alarms are allowed for acceptance of the system, i.e., c = 0. 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 3 provides the required sample size, n, and number of allowable false alarms, c, for a range of false alarm thresholds and acceptable risk levels.

False	Accentable	Number of Allowable False Alarms										
Alarm Threshold	Risk	0	1	2	3	4	5	6	7	8	9	10
0.001	0.01	4603	6636	8403	10042	11601	13105	14567	15996	17398	18779	20140
0.001	0.05	2995	4742	6294	7752	9151	10511	11840	13146	14432	15702	16959
0.001	0.10	2302	3889	5321	6679	7992	9273	10530	11769	12993	14204	15404
0.001	0.15	1897	3372	4722	6013	7266	8493	9702	10895	12076	13247	14410
0.001	0.20	1609	2994	4278	5514	6720	7905	9074	10231	11379	12517	13649
0.005	0.01	919	1325	1678	2006	2318	2618	2910	3196	3476	3752	4024
0.005	0.05	598	947	1258	1549	1829	2100	2366	2627	2884	3138	3389
0.005	0.10	460	777	1063	1335	1597	1853	2105	2352	2597	2839	3079
0.005	0.15	379	674	944	1202	1452	1698	1939	2178	2414	2648	2881
0.005	0.20	322	598	855	1102	1343	1580	1814	2045	2275	2502	2729
0.01	0.01	459	662	838	1001	1157	1307	1453	1596	1736	1874	2010
0.01	0.05	299	473	628	773	913	1049	1182	1312	1441	1568	1693
0.01	0.10	230	388	531	667	798	926	1051	1175	1297	1418	1538
0.01	0.15	189	337	471	600	726	848	969	1088	1206	1323	1439
0.01	0.20	161	299	427	551	671	790	906	1022	1137	1251	1364
0.05	0.01	90	130	165	198	229	259	288	316	344	371	398
0.05	0.05	59	93	124	153	181	208	234	260	286	311	336
0.05	0.10	45	77	105	132	158	184	209	234	258	282	306
0.05	0.15	37	67	94	119	144	169	193	216	240	263	286
0.05	0.20	32	59	85	110	134	157	180	204	226	249	272
0.1	0.01	44	64	81	97	113	127	142	156	170	183	197
0.1	0.05	29	46	61	76	89	103	116	129	142	154	167
0.1	0.10	22	38	52	65	78	91	104	116	128	140	152
0.1	0.15	19	33	46	59	72	84	96	107	119	131	142
0.1	0.20	16	29	42	54	66	78	90	101	113	124	135

Table 3: Systems with occupancy sensors: required occupancies for stated false alarmthreshold, acceptable risk (type I error) and maximum number of false alarms allowable to
deem system as good.

An experimenter uses Table 3 by identifying the row that corresponds to the stated false alarm 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 false alarms were to be allowed, i.e., c = 0. As one moves across the row in the main body of the table, the required sample size increases as the number of allowable false alarms increases. This increase in sample size reduces the producer risk (type II error).

Consider an experiment that seeks to prove that a system has a true false alarm rate no greater than one false alarm per 1000 occupancies, i.e., a false alarm threshold of $p^* = 0.001$ with a maximum acceptable risk of $\alpha = 0.05$. As noted in Table 1, this false alarm threshold is common among the ANSI standards for radiation detectors used for homeland security applications. From Table 3, we see that the experimenter could choose to perform a test with as few as n = 2995 trials, though the system would be deemed as good only if no false alarms were recorded over all 2995 trials. If the experimenter wished to increase the sample size, increase the number of allowable false alarms, and decrease the producer risk, he or she could do so by increasing the sample size to n = 4742 and allow for one false alarm, or n = 6294 with two false alarms, or n = 7752 with three false alarms, and so on. The power curves associated with this family of potential tests, which satisfy a stated false alarm threshold of $p^* = 0.001$ and a maximum acceptable risk of $\alpha = 0.05$ 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 false alarm threshold of $p^* = 0.001$ and maximum acceptable risk of $\alpha = 0.05$.

6. Systems without Occupancy Sensors

When a radiation detection system is equipped with an occupancy sensor, the encounter and observation are clearly defined. This is not the case for radiation detection systems without occupancy sensors where the system is constantly evaluating the surrounding environment. For systems without occupancy sensors, we define the system's false alarm rate based on a time period. In this case, we model the number of false alarms over a given time period using the Poisson probability distribution with intensity parameter λ . The expected false alarm rate is provided by λ which we estimate by the ratio of the number of false alarms, x, to the total number of time periods (e.g., hours) observed, n, i.e., $\hat{\lambda} = \frac{x}{n}$. We use properties of the Poisson probability distribution to guide the quantification of our uncertainty in our estimated false alarm rate, $\hat{\lambda}$, and the development of our hypothesis test.

In this section, we develop a fixed sample false alarm test for systems without occupancy sensors. We provide guidance for determining the sample size, n, and the acceptance criterion to prove that a system satisfies the false alarm threshold.

6.1. Poisson Probability Distribution

A random variable used to describe a number of occurrences of some phenomena over a fixed period of time or within a fixed region of space can often be modeled by the Poisson distribution (Casella & Berger, 2002). Examples include the number of radioactive particles that strike a detector during a fixed period of time and the number of bomb hits in a defined area.

The probability function for the Poisson distribution is provided in Equation (3).

$$P(x \mid \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \qquad x = 0, 1, 2, \dots; 0 \le \lambda$$
(3)

The single positive parameter λ is the expected number of occurrences per unit time, sometimes referred to as the mean occurrence rate or the intensity parameter. In addition to the expected value of the Poisson distribution, λ is also the variance of the distribution. The occurrence rate can be estimated by $\hat{\lambda} = \frac{x}{n}$, where x is the number of occurrences observed and n is the number of units of time over which the observation was made.

When a radiation detection system does not have an occupancy sensor, we model the number of false alarms observed over a period of time using a Poisson distribution. We estimate the system's false alarm rate with the Poisson distribution's occurrence rate, λ .

6.2. Power and Sample Size

We proceed in developing the parameters of our hypothesis test for a system without an occupancy sensor as we did in Section 5.2: we leverage the Clopper-Pearson "exact" method and directly utilize the definition of the Poisson distribution provided in Eq. (3). To distinguish the methods in this section pertaining to systems without occupancy sensors from the methods provided in Section 5, we provide the following, slightly altered, notation:

- λ system's true but unknown false alarm rate
- λ^* false alarm rate threshold requirement
- α maximum acceptable risk (type I error probability)
- *n* test duration in number of time units
- *c* acceptance criterion
- X total number of false alarms observed during the entire testing time

Again, to develop our defensible and successful false alarm test we begin by defining the false alarm threshold requirement, λ^* , and stating the maximum acceptable risk, α . Any system with a true false alarm rate, λ , that is less than or equal to λ^* is considered "good", otherwise, the system is considered "bad".

For the purpose of this discussion we consider the time unit to be one hour and we define our false alarm threshold in these terms. For example, the ANSI N42.32 standard requires that a system present no more than two false alarms during an eight-hour period, thus, we define $\lambda^* = \frac{2}{8} = 0.25$ alarms per hour. We tally the number of false alarms observed, X, over the course of the test period that consists of *n* hours and deem the system as good if the number of false alarms observed is less than or equal to the acceptance criterion, c. Using the definition of the Poisson distribution, we can calculate the probability of deeming a system with true false alarm rate λ as good for any acceptance criterion, c, and test duration, n. That is, we calculate the probability that the number of false alarms, X, will be less than or equal to the acceptance criterion, c, for a Poisson random variable with intensity parameter λ and test duration n (Eq. (4)).

$$P(\text{deem system good}) = P(X \le c \mid \lambda, n) = \sum_{i=0}^{x} \frac{e^{-\lambda n} \lambda n^{i}}{i!}$$
(4)

As an example, consider a test with a false alarm rate threshold of $\lambda^* = 0.1$ alarms per hour, maximum acceptable risk $\alpha = 0.05$, test duration n = 24 hours (one day), and acceptance criterion c = 1; we calculate the probability of deeming a system as good with a true false alarm rate of $\lambda = 0.25$ by:

$$P(\text{deem system good}) = P(X \le 1 | \lambda = 0.25, n = 24) = \sum_{i=0}^{1} \frac{e^{-6}6^i}{i!} = 0.0174$$

Because the true false alarm rate of this example system, $\lambda = 0.25$, is greater than the false alarm threshold, $\lambda^* = 0.1$, (the system is "bad") we desire to see that the test provides a low probability of deeming this system as good. We carry out the above calculation for many different true system false alarm rates ranging from 0 to 1 and provide these results as the power curve in Fig. 5.



Fig. 5: Power curves for test with duration n = 24 hours, acceptance criterion c = 1, and false alarm threshold $\lambda^* = 0.1$ alarms per hour (black points and curve) and n = 48 hours, acceptance criterion c = 1, and false alarm threshold $\lambda^* = 0.1$ alarms per hour (blue curve). The horizontal dashed line is maximum acceptable risk of $\alpha = 0.05$.

We immediately note that just beyond the false alarm threshold at $\lambda = 0.1 + \varepsilon$, where ε is some very small, negligible value, the probability of deeming the system as good (0.308) exceeds the maximum acceptable risk $\alpha = 0.05$. Thus, we must adjust either the test duration or the acceptance criterion – or both – to develop a test that satisfies the test requirements. Increasing the test duration to two days (n = 48 hours) rather than one (n = 24 hours) with acceptance criterion c = 1 resolves the issue in this example by providing a maximum probability of deeming a bad system as good of 0.048.

As we did in Section 5.2 for systems with occupancy sensors, here we use optimization routines in conjunction with Eq. (4) to develop Table 4 that provides test parameters n and c that satisfy the false alarm rate requirement and stated maximum acceptable risk for systems without occupancy sensors. Again, for a required false alarm rate threshold requirement and acceptable risk, a user may first identify the minimum duration test which occurs when no false alarms are allowed for acceptance of the system, i.e., c = 0, and consider increasing the duration and acceptance criterion to decrease the producer risk (type II error). 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.

False	Accentable	Number of Allowable False Alarms										
Alarm Rate Reqirement	Risk	0	1	2	3	4	5	6	7	8	9	10
0.01	0.01	461	664	841	1005	1161	1311	1458	1600	1741	1879	2015
0.01	0.05	300	475	630	776	916	1052	1185	1315	1444	1571	1697
0.01	0.10	231	389	533	669	800	928	1054	1178	1300	1421	1541
0.01	0.15	190	338	473	602	727	850	971	1090	1208	1325	1442
0.01	0.20	161	300	428	552	673	791	908	1024	1138	1252	1366
0.05	0.01	93	133	169	201	233	263	292	320	349	376	403
0.05	0.05	60	95	126	156	184	211	237	263	289	315	340
0.05	0.10	47	78	107	134	160	186	211	236	260	285	309
0.05	0.15	38	68	95	121	146	170	195	218	242	265	289
0.05	0.20	33	60	86	111	135	159	182	205	228	251	274
0.1	0.01	47	67	85	101	117	132	146	160	175	188	202
0.1	0.05	30	48	63	78	92	106	119	132	145	158	170
0.1	0.10	24	39	54	67	80	93	106	118	130	143	155
0.1	0.15	19	34	48	61	73	85	98	109	121	133	145
0.1	0.20	17	30	43	56	68	80	91	103	114	126	137
0.25	0.01	19	27	34	41	47	53	59	64	70	76	81
0.25	0.05	12	19	26	32	37	43	48	53	58	63	68
0.25	0.10	10	16	22	27	32	38	43	48	52	57	62
0.25	0.15	8	14	19	25	30	34	39	44	49	53	58
0.25	0.20	7	12	18	23	27	32	37	41	46	51	55
0.5	0.01	10	14	17	21	24	27	30	32	35	38	41
0.5	0.05	6	10	13	16	19	22	24	27	29	32	34
0.5	0.10	5	8	11	14	16	19	22	24	26	29	31
0.5	0.15	4	7	10	13	15	17	20	22	25	27	29
0.5	0.20	4	6	9	12	14	16	19	21	23	26	28
1	0.01	5	7	9	11	12	14	15	16	18	19	21
1	0.05	3	5	7	8	10	11	12	14	15	16	17
1	0.10	3	4	6	7	8	10	11	12	13	15	16
1	0.15	2	4	5	7	8	9	10	11	13	14	15
1	0.20	2	3	5	6	7	8	10	11	12	13	14

Table 4: Systems without occupancy sensors: required test duration (e.g., hours) for statedfalse alarm rate requirement, acceptable risk (type I error) and maximum number of falsealarms allowable to deem system as good.

An experimenter uses Table 4 by identifying the row that corresponds to the specified false alarm rate requirement and acceptable risk. Within that row, the first column in the main body of the table is the duration of the test required if the acceptance criterion were such that no false alarms were to be allowed, i.e., c = 0. The duration of the test is defined in the same time unit as the false alarm rate requirement. For example, if the false alarm rate requirement is no more than 1 alarm per ten hours ($\lambda^* = 0.1$), then the test duration is defined in hours; if the false alarm rate requirement is no more than 1 alarm per ten minutes ($\lambda^* = 0.1$), then the test duration is defined in minutes. As one moves across the row in the main body of the table, the required test duration increases as the number of allowable false alarms increases. This increase in test duration reduces the producer risk (type II error). Fig. 6 illustrates the power curves that correspond to the family of tests that prove that a system has a false alarm rate no greater than one alarm false alarm during a two hour period ($\lambda^* = 0.5$) with an acceptable risk of $\alpha = 0.05$ (test requirement of ANSI 42.35). This family of curves illustrates the reduction in producer risk when the test duration is increased.



Fig. 6: Power curves for family of tests that satisfy a stated false alarm rate requirement of $\lambda^* = 0.5$ and maximum acceptable risk of $\alpha = 0.05$.

7. Summary

The approach presented in this chapter to develop a false alarm test for radiation detection systems is similar to that used to confirm a performance threshold (Leber, Pibida, & Enders, 2019), but here our threshold is an upper bound not to be exceeded. With a false alarm threshold and acceptable level of risk specified, we've presented approaches to develop false alarm tests for both systems with and without occupancy sensors. The test for systems with occupancy sensors relies on estimates and uncertainties from the binomial probability distribution. The test for systems without an occupancy sensor relies on estimates and uncertainties from the binomial probability distribution.

The hypothesis tests provided here are mathematically equivalent to demonstrating that a one-sided $(1-\alpha)$ % upper confidence bound for the estimated false alarm rate is less than the false alarm threshold. Readers interested in pursuing this alternate approach can refer to Hahn and Meeker (1991) for guidance on one-sided upper confidence bound calculations for binomial and Poisson estimates.

Acknowledgments

This work was sponsored by the U. S. Department of Homeland Security. The authors would like to thank our colleagues, Denis Bergeron and Amanda Koepke, for their valuable input that improved this work.

References

- Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2), 119-126.
- Casella, G., & Berger, R. L. (2002). *Statistical Inference* (Second ed.). Pacific Grove, CA: Duxbury.
- Clopper, C. J., & Pearson, E. S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26, 404-413.
- Hahn, G. J., & Meeker, W. Q. (1991). *Statistical Intervals A Guide for Practitioners*. New York: John Wiley & Sons, Inc.
- Leber, D. D., Pibida, L., & Enders, A. L. (2019, June). Confirming a Performance Threshold with a Binary Experimental Response. *NIST Technical Note*. Retrieved from https://doi.org/10.6028/NIST.TN.2045
- Mitka, M. (2013). Joint Commission Warns of Alarm Fatigue: Multitude of Alarms From Monitoring Devices Problematic. *Journal of the American Medical Association*, 309(22), 2315-2316.
- Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics and Probability for Engineers* (Sixth ed.). Hoboken, NJ: John Wiley & Sons, Inc.
- Sendelbach, S., & Funk, M. (2013). Alarm Fatigue: A Patient Safety Concern. AACN Adv Crit Care, 24(4), 378-388.
- Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2), 117-186.