Volume 114, Number 3, May-June 2009
Journal of Research of the National Institute of Standards and Technology

[J. Res. Natl. Inst. Stand. Technol. **114**, 195-199 (2009)]

# *Pass-Fail Testing: Statistical Requirements and Interpretations*

Volume 114 | Number 3 | May-June 2009

**David Gilliam, Stefan Leigh, Andrew Rukhin, and William Strawderman**

National Institute of Standards and Technology,
Gaithersburg, MD 20899

david.gilliam@nist.gov
stefan.leigh@nist.gov
andrew.rukhin@nist.gov
william.strawderman@nist.gov

Performance standards for detector systems often include requirements for probability of detection and probability of false alarm at a specified level of statistical confidence. This paper reviews the accepted definitions of confidence level and of critical value. It describes the testing requirements for establishing either of these probabilities at a desired confidence level. These requirements are computable in terms of functions that are readily avail-able in statistical software packages and general spreadsheet applications. The statistical interpretations of the critical values are discussed. A table is included for illustration, and a plot is presented showing the minimum required numbers of pass-fail tests. The results given here are applicable to one-sided testing of any system with performance characteristics conforming to a binomial distribution.

**Key words:** binomial distribution; confidence bounds; confidence coefficient; critical value; probability of detection; probability of false alarm.

**Accepted:** April 27, 2009

**Available online:** http://www.nist.gov/jres

## 1. Introduction

In evaluating the efficacy of equipment that is meant for detection of hidden contraband or dangerous substances, the instrument is often subjected to testing that measures its performance against requirements set forth in protocols set by national or international standards organizations. Performance requirements in these standards include those for probability of detection (PD) and probability of false alarm (PFA) at a specified level of statistical confidence.

The detection systems considered in this paper are all assumed to behave according to a binomial distribution. Only two outcomes are considered for independent trials with contraband present: the detection system either correctly reports detection or does not. Furthermore, the probability of detection must remain constant during the period of the testing. Otherwise, it may be meaningless to perform binomial model based tests to determine estimates of this quantity. Similarly, for tests with contraband absent, the detection system either correctly reports no detection, or it falsely reports the presence of contraband: and the probability of a false alarm is presumed to remain fixed throughout the period of testing.

For a detection system, PD or PFA can only be determined accurately by a sufficient number of trials. However, there is a number called the confidence level (CL) that gives some sense of adequacy of the results from a series of trials of a given size.

CL is defined in terms of the binomial probability mass function, also called the binomial discrete density function, $b(m; n,p)$,

$$\begin{aligned} b(m;n,p) &= Pr(\mathrm{BIN}(n,p)=m) \\ &= \frac{n!}{m!(n-m)!}\, p^m (1-p)^{n-m}\,, \end{aligned} \tag{1}$$

where $m = 0,1, \ldots, n$, denotes the number of successful detections or false alarms) in $n$ independent trials with $p$ = PD, or $p$ = PFA, $0 \le p \le 1$ (see Johnson, Kotz, and Kemp, 1992.) The number of successes in $n$ repeated independent trials conforms to this function if each trial can be scored as either success or failure and the probability for success is fixed.

In Sec. 2 we discuss the definitions of CL and related critical values in detection problems. Section 3 gives statistical interpretation of these values in terms of hypothesis testing and confidence bounds. The note is concluded with Sec. 4 containing some examples.

## 2. Definitions and Test Requirements

The quantity CL can be loosely interpreted as the likelihood that any such system conforming to a binomial distribution with $m$ successes in a series of $n$ independent trials will have a true PD value greater or equal to a chosen value, $\mathrm{PD}_c$.

More formally, the accepted definition of CL in setting testing requirements is stated in terms of the equation below. The usage of this term is consonant with that of ASTM standard C 1236-99 (2005).

For a number $m$ of successes found in a series of $n$ pass-fail trials, with a fixed value of PD, designated $\mathrm{PD}_c$, the confidence level $CL(m, n, PD_c)$ is defined by the equation

$$CL(m,n,PD_c) = \sum_{j=0}^{m-1} b(j;n,PD_c)\,. \tag{2}$$

In other words, if for $x = 0, 1, \ldots, n$, $0 \le p \le 1$,

$$\begin{aligned}\mathrm{BINCDF}(x,n,p) &= Pr(\mathrm{BIN}(n,p) \le x)\\ &= \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}\end{aligned} \tag{3}$$

denotes the binomial cumulative distribution function, then (2) can be expressed as

$$CL(m,n,PD_c) = \mathrm{BINCDF}(m-1,\, n,\, PD_c)\,. \tag{4}$$

Note that under this definition $CL(m, n, PD_c)$ cannot exceed $1 - PD_c^n$.

To find the critical value $m_c$, i.e., the minimum value of $m$ establishing the $\mathrm{PD}_c$ of interest with a preselected, fixed level of confidence, CL, one must invert the inequality,

$$\mathrm{BINCDF}(m_c - 1, n, PD_c) \ge CL\,. \tag{5}$$

It follows that $m_c$ is well defined only if BINCDF $(n-1, n, PD_c) \ge CL$, i.e., if

$$1 - PD_c^n \ge CL\,. \tag{6}$$

Since BINCDF($x$, $n$, $p$) is a step-function in $x$ (i.e., is not strictly increasing), it does not have a proper inverse function. If we set $m_c - 1$, $1 \le m_c \le n$ to be the least integer such that BINCDF($m_c - 1$, $n$, $PD_c$) exceeds CL, then

$$m_c = INVBINCDF(CL,n,PD_c) + 1\,, \tag{7}$$

where $INVBINCDF(CL, n, p)$ is the inverse cumulative binomial distribution function (i.e., is the smallest nonnegative integer such that the cumulative distribution function evaluated at this value equals or exceeds CL.) Versions of this function are available in many statistical software packages, including MATLAB (*binoinv*), R (*qbinom*), NAG, GAMS, IMSL, S-PLUS, and SAS and in general spreadsheet applications, such as EXCEL (function *CRITBINOM*($n$, $p$, $CL$).)[1]

The binomial cumulative distribution function can be expressed through the incomplete beta-function,

$$\begin{aligned}\mathrm{BINCDF}(m-1,\, n,\, p) &= 1 - I_p(m,\, n-m+1)\\ &= \frac{\int_p^1 x^{m-1}(1-x)^{n-m}\,dx}{\int_0^1 x^{m-1}(1-x)^{n-m}\,dx}\,,\end{aligned} \tag{8}$$

$m > 0$, $n - m + 1 > 0$, (Abramowitz and Stegun, 1972), so that for fixed $m$ and $n$, BINCDF($m - 1$, $n$, $p$) is a decreasing function of $p$, $0 \le p \le 1$. This formula allows one to define BINCDF($m - 1$, $n$, $p$) for any real (non-integer) values $m$ and $n$ such that $0 < m < n + 1$.

An analogous definition of CL applies to testing for PFA in systems where no contraband or dangerous substance is present. For any chosen value of PFA, designated $\mathrm{PFA}_c$, the confidence level $CL(m, n, PFA_c)$, equals the probability that the number of false alarms occurring in a series of $n$ independent binary trials exceeds $m$. Thus, this level is defined by the equation

$$\begin{aligned}CL = CL(m,n,PFA_c) &= \sum_{k=m+1}^{n} b(k;n,PFA_c)\\ &= 1 - \mathrm{BINCDF}(m,n,PFA_c).\end{aligned} \tag{9}$$

[1] Any mention of specific commercially available statistical software packages or general spreadsheet applications does not imply endorsement of preference for these products by the NIST.

Similarly to the PD case,

$$CL \le 1-(1-PFA_c)^n. \tag{9}$$

To find the maximum value $M_c$ of $M$, $M = 0, 1, \ldots, n-1$, establishing the $\mathrm{PFA}_c$ of interest with a preselected, fixed level of confidence CL, one must invert the inequality

$$1-\mathrm{BINCDF}(M_c, n, PFA_c) \ge CL. \tag{11}$$

To express $M_c$ through the function INVBINCDF $(c, n, p)$, i.e., to establish the largest value $m$ satisfying (11), the formula,

$$\begin{aligned}&\mathrm{INVBINCDF}(c, n, p) = n-1\\ &\quad -\max\{x:\mathrm{BINCDF}(x, n, 1-p) \le 1-c\},\end{aligned} \tag{12}$$

can be employed. To prove (12), notice that for $x = 0, \ldots, n-1$,

$$\begin{aligned}&\mathrm{BINCDF}(x, n, p) =\\ &\quad 1-\mathrm{BINCDF}(n-x-1, n, 1-p),\end{aligned} \tag{13}$$

so that

$$\begin{aligned}&n-1-\mathrm{INVBINCDF}(c, n, p)\\ &= n-1-\min\{x:\mathrm{BINCDF}(x, n, p) \ge c\}\\ &= n-1-\min\{x:\mathrm{BINCDF}(n-x-1, n, 1-p) \le 1-c\}\\ &= \max\{x:\mathrm{BINCDF}(x, n, 1-p) \le 1-c\}.\end{aligned} \tag{14}$$

Therefore,

$$M_c = n-1-\mathrm{INVBINCDF}(c, n, 1-PFA_c), \tag{15}$$

so that $M_c \le n-1$ and $M_c$ is not defined when

$$INVBINCDF(CL, n, 1-PFA_c) = n,$$

i.e., when $(1-PFA_c)^n > 1-CL$.

Thus (15) and (7) show that under the same value of CL, when PD = 1– PFA, a simple formula,

$$m_c + M_c = n, \tag{16}$$

relates $m_c$ and $M_c$.

## 3. Hypothesis Testing and Confidence Bounds on Binomial Probability

We give here two statistical interpretations of Eq. (7) and Eq. (15). The first of these is related to a (lower) cconfidence limit for binomial probability $p$. Such limits are supposed to provide a data-dependent interval containing the unknown $p$ with a given probability called *confidence coefficient* (see Hahn and Meeker, 1991).

Assume that for the given CL, a lower confidence bound for PD = $p$ of confidence coefficient CL is desired: that is for a binomial observation $X \sim BIN(n, p)$, one requires a function $\underline{p} = \underline{p}(X, n, \mathrm{CL})$ such that

$$Pr(\underline{p}(X, n, CL) \le p) \ge CL. \tag{17}$$

The well known solution of this problem for $X \ge 1$, is

$$\begin{aligned}&\underline{p}(X, n, CL) =\\ &\quad \max\{p:\mathrm{BINCDF}(X-1, n, p) \ge CL\}.\end{aligned} \tag{18}$$

(e.g, Casella and Berger, 2002.) When $X = 0$, $\underline{p}(0, n, CL) = 0$.

Thus with $m_c$ defined by (7), the inequalities $\underline{p} < p$ (strict inequality) and $X \le m_c$ (non-strict inequality) are equivalent. Therefore, the critical value $m_c$ has the interpretation of the largest value of the binomial $BIN(n, p)$ variable such that the lower confidence bound for $p$ does not exceed $PD_c$.

A related interpretation is provided by the statistical hypothesis testing problem, $H_0: p \ge PD_c$ under the alternative: $H_1: p < PD_c$. The most powerful test of level $1-\mathrm{CL}$ rejects $H_0$ when the observed value $X$ exceeds the critical value $m$, $X > m$ (which means the same as $\underline{p}(X, n, CL) \ge PD_c$).

The critical value for PFA has a similar statistical interpretation, namely, $M_c$ is the largest value of the binomial variable for which the upper confidence bound for the binomial probability does not exceed $PFA_c$. Indeed, an upper confidence bound of confidence coefficient CL has the form,

$$\overline{p}(X, n, CL) = 1-\underline{p}(n-X, n, CL). \tag{19}$$

Identity (13) shows that

$$\begin{aligned}&\overline{p}(X, n, CL) =\\ &\quad \min\{p:\mathrm{BINCDF}(X, n, p) \le 1-CL\}.\end{aligned} \tag{20}$$

Thus, $\overline{p}(M_c, n, CL) \le PFA_c$,

but $\overline{p}(M_c+1, n, CL) > PFA_c$.

In terms of the hypothesis testing with $H_0: p \le PFA_c$ and the alternative: $H_1: p > PFA_c$, the most powerful test of level 1-CL rejects $H_0$ when the observed value $X$ exceeds the critical value $M_c$, $X > M_c$.

## 4. Examples

Consider an example in which one finds twenty-nine correct results in a single set of thirty trials. If the system under test conforms to a binomial distribution, then based on the result of twenty-nine out of thirty correct responses in that one set of tests, one can make multiple correct inferences, such as: the PD > 0.95 with 44 %, confidence, the PD > 0.90 with 81 %, confidence, or the PD > 0.85 with 95 % confidence.

One can easily construct a table which simultaneously includes requirements for both PD and PFA.

Table 1 gives the critical value $M_c$ and $n - m_c$ for 68 % confidence to show the general characteristics of these quantities. These are the maximum permissible numbers of incorrect results that may be tolerated in establishing the specified PD or PFA values at this level of confidence. If the tabulated value is indicated as "*", then the number of trials in that set is insufficient to establish the corresponding PD or PFA at this confidence level. One may generate tables of this kind for any CL, PD, and PFA using Eq. (7) and Eq. (15) by using the previously mentioned functions like *binoinv* or *CRITBINOM* from statistical software packages or spreadsheet applications. The actual value of $M_c$ and $n - m_c$ given by these functions in the cases marked by "*" is – 1.

The symmetry of testing requirements when PFA = 1 – PD permits tabulating the results for PFA and PD in a single table, but it does not imply that PFA should or must always be chosen equal to 1 – PD. The PD and PFA values may be assigned independently in any testing protocol. In fact, to avoid disruption of the stream of commerce by large numbers of false alarms, it is often necessary to require inspection equipment to have PFA smaller than 1 – PD.

By solving (6) or (10), we obtain a formula for the minimum number of required trials $n_k$ needed to establish a given value of PD or PFA for the same *CL*,

$$n_k = \lceil a \rceil, \tag{21}$$

with

$$a = \frac{\log(1 - CL)}{\log PD} = \frac{\log(1 - CL)}{\log(1 - PFA)}. \tag{22}$$

Here $\lceil a \rceil$ denotes the smallest integer exceeding $a$. This formula is useful in designing test protocols that give the most satisfactory requirement with the least amount of testing. Figure 1 shows $a$ plotted as a function of PD and CL. This function increases much more rapidly for $PD$ approaching 1 than for $CL \to 1$.

**Table 1.** Maximum permissible numbers of incorrect results for verifying a lower bound on PD or an upper bound on PFA with 68 % confidence

| PD→ | 0.95 | 0.90 | 0.85 | 0.80 | 0.75 | 0.70 | 0.60 | 0.50 |
|---|---|---|---|---|---|---|---|---|
| PFA→ | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.40 | 0.50 |
| n = 2 | * | * | * | * | * | * | * | 0 |
| n = 3 | * | * | * | * | * | * | 0 | 0 |
| n = 4 | * | * | * | * | 0 | 0 | 0 | 1 |
| n = 5 | * | * | * | * | 0 | 0 | 0 | 1 |
| n = 6 | * | * | * | 0 | 0 | 0 | 1 | 1 |
| n = 7 | * | * | * | 0 | 0 | 0 | 1 | 2 |
| n = 8 | * | * | 0 | 0 | 0 | 1 | 2 | 2 |
| n = 9 | * | * | 0 | 0 | 1 | 1 | 2 | 3 |
| n = 10 | * | * | 0 | 0 | 1 | 1 | 2 | 3 |
| n = 11 | * | 0 | 0 | 0 | 1 | 2 | 3 | 4 |
| n = 12 | * | 0 | 0 | 1 | 1 | 2 | 3 | 4 |
| n = 13 | * | 0 | 0 | 1 | 1 | 2 | 3 | 5 |
| n = 14 | * | 0 | 0 | 1 | 2 | 2 | 4 | 5 |
| n = 15 | * | 0 | 1 | 1 | 2 | 3 | 4 | 6 |
| n = 16 | * | 0 | 1 | 1 | 2 | 3 | 4 | 6 |
| n = 17 | * | 0 | 1 | 2 | 2 | 3 | 5 | 7 |
| n = 18 | * | 0 | 1 | 2 | 3 | 3 | 5 | 7 |
| n = 19 | * | 0 | 1 | 2 | 3 | 4 | 6 | 7 |
| n = 20 | * | 0 | 1 | 2 | 3 | 4 | 6 | 8 |
| n = 21 | * | 0 | 1 | 2 | 3 | 4 | 6 | 8 |
| n = 22 | * | 0 | 1 | 2 | 3 | 5 | 7 | 9 |
| n = 23 | 0 | 1 | 2 | 3 | 4 | 5 | 7 | 9 |
| n = 24 | 0 | 1 | 2 | 3 | 4 | 5 | 7 | 10 |
| n = 25 | 0 | 1 | 2 | 3 | 4 | 5 | 8 | 10 |
| n = 30 | 0 | 1 | 2 | 4 | 5 | 7 | 10 | 13 |
| n = 40 | 0 | 2 | 4 | 6 | 8 | 10 | 14 | 18 |
| n = 50 | 1 | 3 | 5 | 8 | 10 | 12 | 17 | 22 |
| n = 60 | 1 | 4 | 7 | 9 | 12 | 15 | 21 | 27 |
| n = 70 | 2 | 5 | 8 | 11 | 15 | 18 | 25 | 32 |
| n = 80 | 2 | 6 | 9 | 13 | 17 | 21 | 29 | 37 |
| n = 90 | 2 | 7 | 11 | 15 | 20 | 24 | 33 | 42 |
| n = 100 | 3 | 7 | 12 | 17 | 22 | 27 | 37 | 47 |

Similarly $n_k$ in (21) would increase much more rapidly for $PFA \to 0$ than for $CL \to 1$.

When only the minimum number of trials $n_k$ is performed, the system must give 100 % correct results to establish the specified PD or PFA at, the desired confidence *CL*. In statistical terms, $n_k$ is the smallest number of trials with 100 % correct detections such that the *CL*-lower confidence bound for detection probability exceeds the given value PD. The same is true when there are no false alarms with the *CL*-upper confidence bound on the false alarm probability being less than PFA. A table such as Table 1 will show how many errors may be permitted if a larger number of trials are carried out, while still establishing the specified PD or PFA at the desired CL.

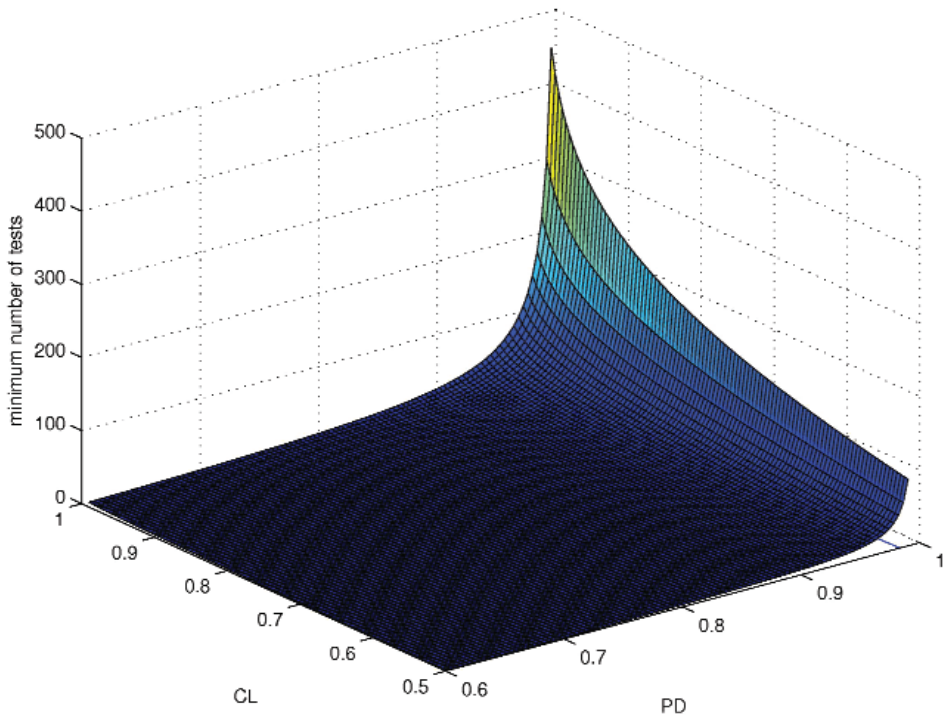


**Fig. 1.** The minimum required number of tests to establish a given value of PD (or 1-PFA) for a given CL.

## 5. Discussion and Conclusions

The formula for $n_k$ shows that requiring either PD or CL to be too near unity can result in impossibly large numbers of pass-fail tests. If such rigorous criteria are in fact required then one should search for some method of verification different from pass-fail testing.

The results presented here make it possible to design pass-fail testing protocols based on functions readily available in statistical software packages and general spreadsheet applications.

## 6. References

[1] M. Abramowitz and I. Stegun, Handbook of Mathematical Functions, Dover, New York (1972) p 263.
[2] ASTM International, Standard Guide for In-Plant Performance Evaluation of Automatic Vehicle SNM Monitors: C 1236-99, W. Conshohocken, PA (2005) pp 1-4.
[3] G. Casella and R. Berger, Statistical Inference, 2nd edition, Duxbury, Pacific Grove (2002) pp 425 427.
[4] G. J. Hahn and W. Q. Meeker, Statistical Intervals: A Guide for Practitioners, Wiley, New York (1991) p 25.
[5] N. Johnson, S. Kot, and A. Kemp, Univariate Discrete Distributions, New York: John Wiley (1992) pp 105-150.

***About the authors:*** *David Gilliam is a nuclear engineer/physicist in the Neutron Interactions and Dosimetry Group, Ionizing Radiation Division, Physics Laboratory. Stefan Leigh and Andrew Rukhin are mathematical statisticians in the Statistical Engineering Division, Information Technology Laboratory. Bill Strawderman a professor in the Department of Statistics at Rutgers University. He is also a Faculty Appointee at NIST. The National Institute of Standards and Technology is an agency of the U.S. Department of Commerce.*