A Statistical Evaluation Metric for Radioactive Isotope Detection in the Environment.

Robert A. August Jr., Ph.D

Abstract

Program managers and system designers trying to determine the minimum detectable amount (MDA) of a specific radioactive source that a proposed detector system can measure, require a statistically valid metric for system performance. A common assumption is that the simple metric $\boldsymbol{S}/\sqrt{\boldsymbol{B}}$ is valid when a detector system is used to determine the mean of a source distribution, $\boldsymbol{S}$, measured in the presence of a background with mean $\boldsymbol{B}$. However, this metric fails to associate $\boldsymbol{S}$ mathematically with the statistical error involved in the measurements necessary to determine this derived quantity, and can lead to gross overestimations of detector performance. The simple metric $\boldsymbol{D} = \boldsymbol{S}/\sqrt{\boldsymbol{S}+2\boldsymbol{B}}$ is shown to be one that can be used as a statistically valid measure of system performance when dealing with strong radiation sources. Ultimately even this metric is too generous for the many operational situations where system false alarms must be balanced against MDA. In those cases the statistically appropriate metric is shown to be $\boldsymbol{S} = \boldsymbol{F}\sqrt{2\boldsymbol{B}} + \boldsymbol{D}\sqrt{\boldsymbol{S}+2\boldsymbol{B}}$ ($\boldsymbol{F}$ being a false alarm metric) for the Poisson or Gaussian statistical distributions that are appropriate to many radiation detection situations, especially those concerning the smuggling of special nuclear material.

Assumptions and Limitations

This brief issue paper address the fundamental question of the statistical significance of a measurement of a radioactive source in a real environment. This paper is not concerned with detector or radiation specifics, just the question of counting statistics. In particular the question of a proper statistical metric to use in evaluating the detection capabilities of a given system design is addressed. This question is answered in detail in most introductory level text books on radiation detection, and I specifically refer here to Chapter 3 of the most recent edition of Glenn Knoll's "Radiation Detection and Measurement." [1] However, these are basic principles not specific to radiation detection and the same arguments can be found in most introductory level text books on statistics.

The focus of this paper is to address the statistics relevant to the general scenario where a system is being designed to combat radionuclear smuggling, and the designer wants to determine the limits of the system to detect a specific radionuclear threat. Most of these scenarios follow Poisson or Gaussian statistical distributions, and in fact most scenarios in a natural environment will have enough total counts to follow the simpler Gaussian distribution which will be assumed here. Those situations outside these simplifying assumptions will require a more complicated metric than considered here.

Statistics Appropriate to Radiation Detection

When measuring any radiation source in a field environment, the measurement will be made in the presence of an ambient background of those same radiations. Therefore, define the mean of the radiation distribution measured during a given time interval as $\boldsymbol{T}$, consisting of a mean $\boldsymbol{B}$ from the background distribution and a mean $\boldsymbol{S}$ from the source of interest.

$$(1) \quad \boxed{\boldsymbol{T} = \boldsymbol{S} + \boldsymbol{B}}$$

Note that this is true for any type of radiation measurements, and for both gross counting and spectral detection methods. The specifics of $\boldsymbol{T}$, $\boldsymbol{S}$ and $\boldsymbol{B}$ are very different depending on what is being measured and what method is used to make the measurement.

It is instructive for the purpose of bringing out general principals to ignore the sensitivity and selectivity of specific detection methods for the moment and consider only the mean of the source distribution, $\boldsymbol{S}$, and the mean of the background distribution, $\boldsymbol{B}$, independent of what detection method was used to arrive at them, or even what specific radiation is being measured. Independent of any specific technology it is always possible to define the basics as laid out in equation (1). In actual field detection it is not possible to measure $\boldsymbol{S}$ in isolation, but it is possible to measure $\boldsymbol{T}$ and $\boldsymbol{B}$. The mean of the source distribution (i.e. source strength) will actually be determined operationally by measuring

B while the source is not present and then subtracting this from the measurement of ***T*** (taking into account equal measurement times), so:

$$(2)\quad \boxed{\boldsymbol{S = T - B}}$$

The problem in most considerations of detector performance arises when the error on ***S*** is considered. As mentioned in the blue box below, signal to noise (background) is an important consideration for nuclear detection, but it leads to the misconception that this can be used directly as a measure of detector performance, commonly in the form: $\boldsymbol{S/\sqrt{B}}$. Such a metric may be useful in comparing one detector with another, but unfortunately this metric does not reflect a detector's statistical ability to detect a source being measured and can result in gross overestimations of detector performance. If one is trying to determine the minimum detectable amount (MDA) of a specific source that a detector

Selectivity and Sensitivity

Some gamma ray detectors measure only gross counts across a broad energy range from any specific source, and have a corresponding ***S*** and ***B***. More sophisticated spectral detectors are capable of measuring an energy spectrum from the same specific source. Given that nuclear isotopes have characteristic spectral signatures, such detectors allow the measurement to focus in on only those regions of the energy spectrum characteristic of the specific source being measured. ***S*** and ***B*** in that case will include only counts from this relevant subset of the energy spectrum, therefore signal to noise will be improved since the ratio ***S*** to ***B*** will be increased. Such spectral detectors have the ability to identify specific isotopes in the presence of competing sources of radiation. This ability to identify isotopes is referred to as selectivity. In general, the degree of selectivity of spectral detectors depends on the energy resolution of the particular detector; higher energy resolution detectors have higher selectivity. The other key measure of a detector is sensitivity, which is the efficiency with which a detector can transform radiation incident on the detector into measured counts. For example, if two different detectors with the same active surface area have one-hundred identical gammas incident upon them, and one measures fifty counts while the other measures only ten, then the former is five times more sensitive than the latter. Together selectivity and sensitivity are key metrics for radiation detectors.

system can measure in a given operational scenario [†](counting time, stand-off distance, etc.), then a statistically valid metric is required.

The real statistical error on equation 2 is determined after considering that $\boldsymbol{S}$ is a derived quantity and only $\boldsymbol{T}$ and $\boldsymbol{B}$ can be actually measured. Both $\boldsymbol{T}$ and $\boldsymbol{B}$ will have standard deviations given by the formulas $\boldsymbol{\sigma_T} = \sqrt{\boldsymbol{T}}$ and $\boldsymbol{\sigma_B} = \sqrt{\boldsymbol{B}}$ for Poisson and Gaussian distributions. The error on $\boldsymbol{S}$ is determined by combining these two standard deviations in quadrature:

$$(3) \quad \boxed{\boldsymbol{\sigma_S} = \sqrt{\boldsymbol{\sigma_T^2} + \boldsymbol{\sigma_B^2}}}$$

where $\boldsymbol{\sigma_T} = \sqrt{\boldsymbol{T}}$ and $\boldsymbol{\sigma_B} = \sqrt{\boldsymbol{B}}$, so equation (3) leads to:

$$(4) \quad \boxed{\boldsymbol{\sigma_S} = \sqrt{\boldsymbol{T} + \boldsymbol{B}}}$$

When making an actual measurement this is the error used. However, when designing a detection system, the designer would like to understand the systems ability to measure a specific real source with a mean $\boldsymbol{S}$, so requires a metric expressed in terms of $\boldsymbol{S}$. Therefore, combining equation (4) with equation (1) yields:

$$(5) \quad \boxed{\boldsymbol{\sigma_S} = \sqrt{\boldsymbol{S} + 2\boldsymbol{B}}}$$

This is equation (3.53) in Knoll[1] for the situation where $\boldsymbol{T}$ and $\boldsymbol{B}$ where counted for equal times. The mathematic handling of unequal counting times in also shown there and is a straightforward process. The standard deviation $\boldsymbol{\sigma_S}$ is the only appropriate error to use when developing a metric for comparison with $\boldsymbol{S}$, since it is the only true measure of the variance in this derived quantity.

As a measure of statistical significance, a factor $\boldsymbol{D}$ is defined such that: $\boldsymbol{D} \equiv \boldsymbol{S} / \boldsymbol{\sigma_S}$. Therefore from equation (4):

$$(6) \quad \boxed{\boldsymbol{D} = \frac{\boldsymbol{S}}{\sqrt{\boldsymbol{S} + 2\boldsymbol{B}}}}$$

[†] For simplicity, equal operational parameters are assumed for all measurements in this paper. So MDA is understood to be the counts from a source with the minimum detectable activity measured this way.

D is therefore an assessment of the detection system's ability to measure the source mean ***S*** in units of standard deviations (this is often expressed as detection probability (DP), especially when expressed in terms of percentage probability which requires reference to the appropriate Poisson or Gaussian statistical distributions). This allows us to use ***D*** to set a statistical threshold to determine the capability limits of any given system. If, for instance, ***D*** = 2 was chosen as the minimum number of standard deviations acceptable for a positive detection of the specific source in question, then a minimum detectable about of that source would be an amount with a mean $\boldsymbol{S} = \boldsymbol{D}\sqrt{\boldsymbol{S}+2\boldsymbol{B}} = 2\sqrt{\boldsymbol{S}+2\boldsymbol{B}}$.

False Alarms: When there is no source present, ***T*** = ***B***, therefore ***S*** = ***0***; however, we can only arrive at this conclusion experimentally by following the procedure already outlined: measuring ***B*** independently and then subtracting this from the measurement of ***T*** when presumably a source might be present. Since there is no source we will find from equation (1) that ***T*** = ***B***, and from equation (4) we find that the error on this measurement is $\boldsymbol{\sigma_S} = \sqrt{2\boldsymbol{B}}$. We can see this situation in Figure 1 where the derived distribution with this error is symmetrical around zero

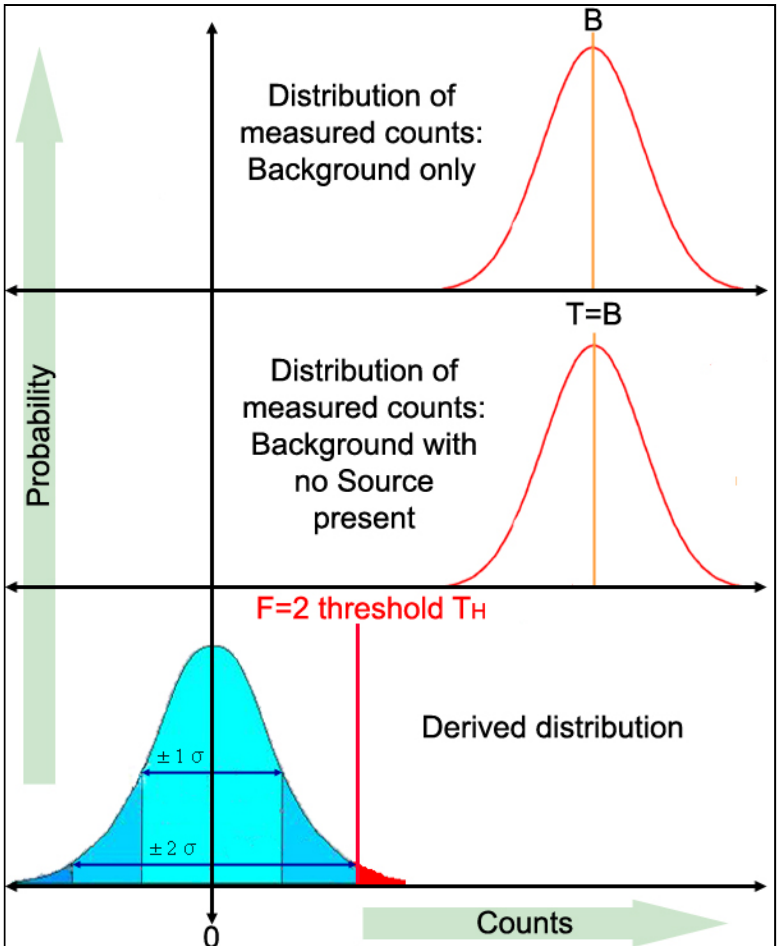


Figure 1: Gaussian count distributions for background and source measurements when no source is present. The counting times are assumed equal for both measurements. A situation is shown where $2\boldsymbol{\sigma_S}$ is chosen as the detection threshold. This means that 97.5% of the derived source distribution is below the ***F*** = 2 line, which corresponds to the probability of correctly rejecting the possibility of a source. The false alarm probability is 2.5%, corresponding to the red portion of the curve above the ***F*** = 2 line.

counts. This means there is a 50% probability that a derived measurement of $\boldsymbol{S}$ will yield a value greater than zero, which is an incorrect representation of the underlying reality and is therefore label a false alarm. When can use equation (6) to define:

$$\boldsymbol{T_h = F\sqrt{2B}} \qquad (7)$$

In this case, since there is no actual source, $\boldsymbol{S}$ in equation (6) has been replaced by a threshold number of counts $\boldsymbol{T_h}$, and $\boldsymbol{D}$ has been replaced by $\boldsymbol{F}$ which is the false alarm metric in sigma units. Therefore, when deriving $\boldsymbol{S}$ by measuring $\boldsymbol{B}$ and $\boldsymbol{T}$, only results greater than $\boldsymbol{T_h}$ will be considered a non-zero result. The portion of the distribution above this threshold corresponds to the probability of detection measurements resulting in an incorrect non-zero value for $\boldsymbol{S}$, generally known as the false alarm probability (false alarm probability (FAP) can be expressed in terms of percentage probability when referenced to the appropriate Poisson or Gaussian statistical distributions). This situation is shown in Figure 1 where $\boldsymbol{F = 2}$ has been chosen as the statistical limit for a measurement made where no source was actually present. This means that any specific measurement that yields counts above this line is a false alarm, while a measurement yielding counts below the line is correctly identified as no source present. In this case, with $\boldsymbol{F = 2}$ for a Gaussian distribution, the false alarm probability is 2.5%.

MDA and Missed Detections: To begin considering the concept of a minimum detectable amount (MDA) of a source, the concept of false alarms should be put aside for the moment. Figure 2 considers the situation when an amount of the source in question equal to a minimum detectable amount for the system used for the measurements is actually present when $\boldsymbol{D = 2}$ from equation (6) has been chosen as the statistical threshold. This figure shows the derived source distribution (which can't be measured directly) whose mean is the MDA, so that the $\boldsymbol{D}$ = 2 line is the threshold at which the MDA can be detected and is $\boldsymbol{S} = 2\sqrt{\boldsymbol{S + 2B}} = 2\boldsymbol{\sigma_S}$ exactly. So with Gaussian statistics there is a 2.5% probability that any real measurement of this source will yield a null result (the portion of the distribution below zero) which is a false negative (missed detection), and a 97.5% chance that such a measurement will yield non-zero counts (the portion of the curve above zero) and be considered a positive detection. The MDA defined in this way corresponds to an absolute lower limit to the MDA.

This means that for a system operated in a known background, equation (6) can be used to calculate the absolute minimum MDA for any chosen statistical threshold ***D***. It is instructive to consider what would happen in Figures 1 and 2 if a different choice of ***D*** or ***F*** were made, for example ***D*** = 1 or ***F*** = 1. In Figure 1 (when there is no source present) the red threshold line would move toward zero to match the 1σ line shown on the distribution. This would result in a larger red portion of the curve, meaning a higher false alarm probability (32% for a Gaussian). In Figure 2 (when there is actually a source present) the mean of the source distribution, ***S***, would move closer to zero so that the 1σ line shown on the distribution would match the zero axis. Thus the mean source counts, ***S***, would be less, meaning the system has a lower MDA for the source. However, it also means a higher probability of a null result (missed detection), 32% for a Gaussian. So choosing a lower ***D*** decreases the MDA but increases the probability of a missed detection. Similarly, choosing a higher ***D*** increases the MDA but decreases the probability of a missed detection. So in the end the choice of a minimum acceptable ***D*** or ***F*** is an

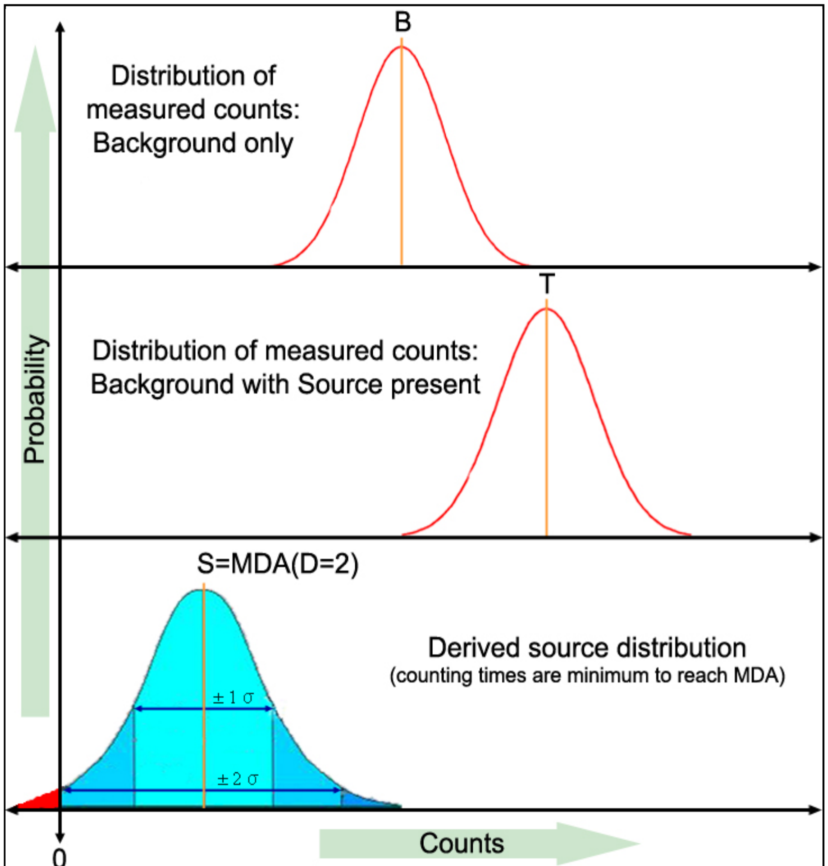


Figure 2: Gaussian count distributions for background and source measurements. The counting times are assumed equal for both measurements and are the minimum needed for a positive detection of the minimum detectable amount (MDA) of the source at ***D*** = 2. The source distribution shown is the derived (not directly measureable) source distribution. The result is that 97.5% of the distribution is above zero, corresponding to the probability of a positive detection, and 2.5% of the distribution is below zero, corresponding to the probability of a false negative.

operational and political one.

MDA and False Alarms: In virtually all realistic systems one actually wants to know the MDA of a system with respect to an acceptable rate of false alarms. The metric D as defined in equation (6) is concerned only with the probability that a measurement made in the presence of a source either detects that source (non-zero result) or doesn't. It defines an MDA based on that consideration alone, and has a 50% false alarm rate when no source is present. In most realistic operational situations there is also a concern that the detection system does not produce false alarms when no source is present. In such a case, the procedure is to define a threshold with no source present as in Figure 1, but then to define the MDA such that the derived source distribution with error produces a result with the acceptable probability of being above the threshold line (not just above zero). This requires a more complicated metric than equation (6), and results in higher MDA's, but yields a system with a known sensitivity to false alarms. In this case F and D are related by:

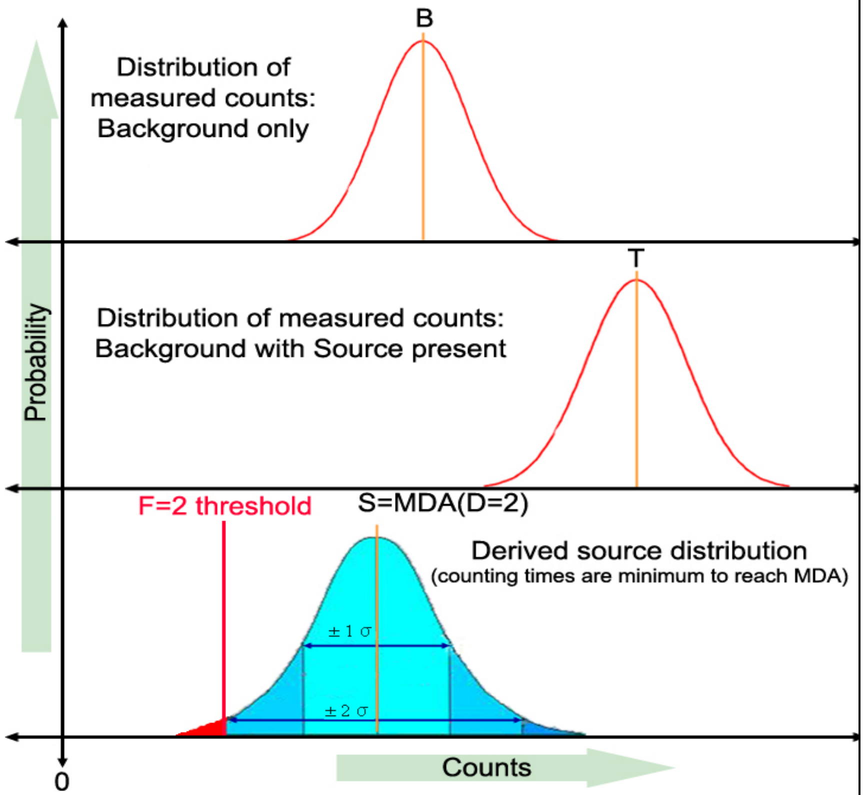


Figure 3: Gaussian count distributions for background and source measurements. The counting times are assumed equal for both measurements and are the minimum needed for a positive detection of the minimum detectable amount (MDA) of the source at ***D*** = 2 above a threshold set by ***F*** = 2. The source distribution shown is the derived (not directly measureable) source distribution. The result is that 97.5% of the distribution is above threshold, corresponding to the probability of a positive detection, and 2.5% of the distribution is below zero, corresponding to the probability of a false negative. The threshold is set as in Figure 1 so that the probability of false alarm with no source present is 2.5%

$$\boxed{\boldsymbol{S} = \boldsymbol{F}\sqrt{2\boldsymbol{B}} + \boldsymbol{D}\sqrt{\boldsymbol{S} + 2\boldsymbol{B}}} \tag{8}$$

This is shown in Figure 3 for the case where $\boldsymbol{F}$ = 2 and $\boldsymbol{D}$ = 2, with counts in the distribution such that its mean corresponds exactly to the MDA for these values. Note that this results in a higher MDA than in Figure 2, but that now the false alarm rate is 2.5% rather than 50%. In other words, this set-up can't detect as weak a source as figure 2, but it has traded that off for a significant improvement in false alarm rate.

Equation (8) is the general equation to use when designing a detection system. A designer must decide what are acceptable values of $\boldsymbol{F}$ for false alarm probability and $\boldsymbol{D}$ for detection probability for the source the system is designed to detect. Then the detectors are sized to ensure that an MDA of the target source produces enough counts in the allotted time for the mean of its distribution to reach a count threshold as defined by equation (8). Of course that system will react differently to amounts of the target source that are greater or less than an MDA, but the relationship between false alarms and detection probability will still be governed by equation (8). This will also be true when the system is exposed to radiation sources other than the target source.

ROC (Receiver Operating Characteristic) curves

A useful tool to determine how a detection system will react to varying amounts of different sources are ROC curves. The relationship between detection probability (DP) and false alarm probability (FAP) are shown in Figure 4 for the cases of random guessing, a perfect system, and a more realistic example. Because we are dealing with the statistics of radiation detection as governed by equation (8), a perfect system is impossible. However, a system designer will naturally strive for a system that pushes as close to the ideal of the top left corner of the graph as possible. Note that Figure 4 shows the curves not in terms of sigma units (though this is possible) which is the way that $\boldsymbol{F}$ and $\boldsymbol{D}$ appear in equation (8), but in terms of percent (1 = 100% in Figure 4). This requires reference to the appropriate Poisson or Gaussian statistical distributions relevant to the situation in question.

ROC curves can be very useful for program managers who are looking for systems that meet minimum detection probability and false alarm goals for a variety of sources operating under a variety of conditions. An comparison of the various sources and situations can be shown side by side on the same graph of ROC curves. Conversely, different systems operating against the same sources and situations can be compared. The difficulty comes when one tries to use ROC curves to compare a system's projected performance to that system's real-world operation. Experimentally measuring ROC curves on real-world systems is a statistical process that can involve very high numbers of repetitions of measurements taken under a variety of conditions. In such situations it is many times only possible to compare a systems performance against its ideal $\boldsymbol{F}$ and $\boldsymbol{D}$.

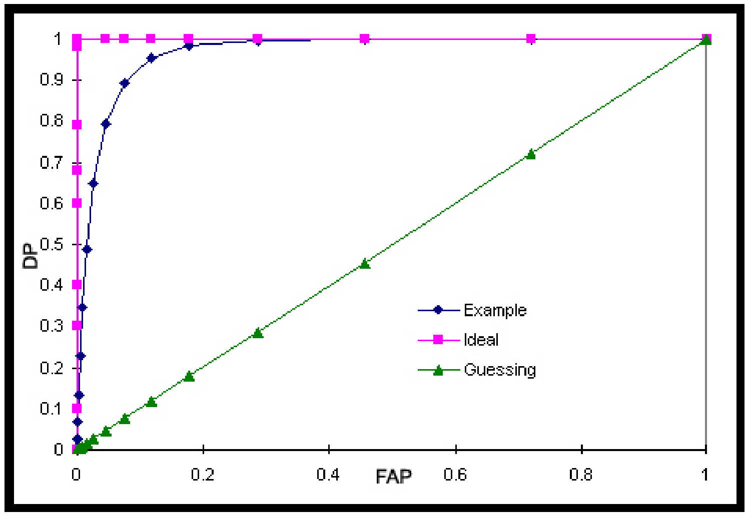
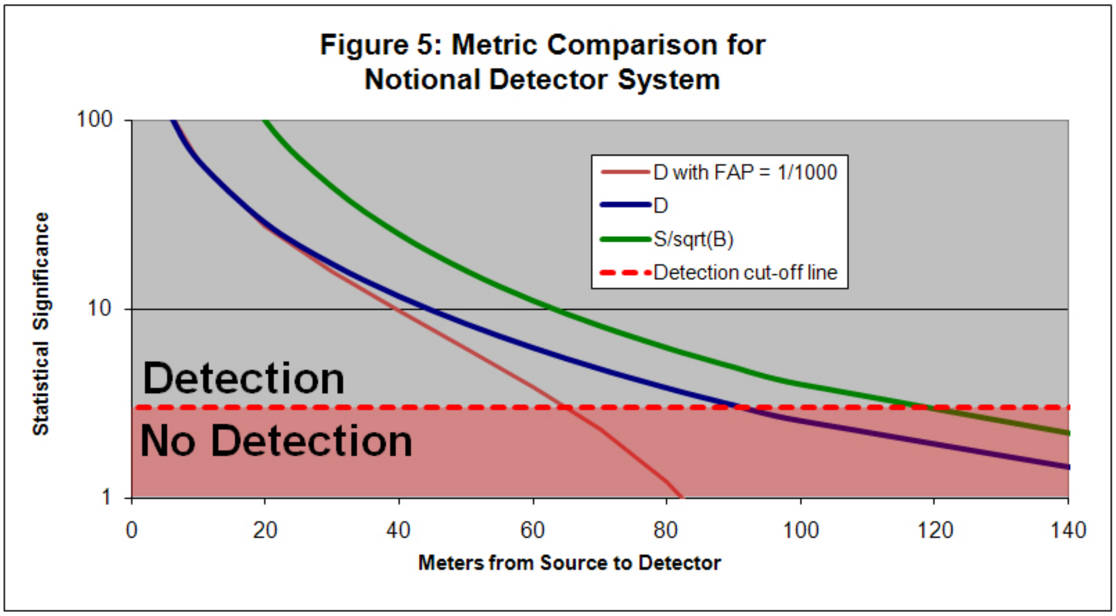


Figure 4: Example ROC curves showing the worst possible case of random guessing, the idealized case which has a point on the top left with 100% detection probability and 0% false alarm probability, and a more realistic case as might be encountered in a real system.

Comparing Different Statistical Metrics

The bottom line is that only metrics that mathematically associate $\boldsymbol{S}$ and $\boldsymbol{\sigma_S}$ are accurate measures of the statistical significance of $\boldsymbol{S}$ in radiation detection measurements in field environments. A comparison of $\boldsymbol{D}$ from equation (6) (statistically accurate measure ignoring false alarms) and $\boldsymbol{D}$ from equation (8) (false alarms explicitly taken into account) with the commonly quoted $\boldsymbol{S}/\sqrt{\boldsymbol{B}}$ shows the problem with inappropriate statistical metrics. For the case when $\boldsymbol{S}$ is small compared to $\boldsymbol{B}$ ($\boldsymbol{S} << \boldsymbol{B}$), then for equation (6) $\boldsymbol{D} \approx \boldsymbol{S}/\sqrt{2\boldsymbol{B}}$, which means that using the metric $\boldsymbol{S}/\sqrt{\boldsymbol{B}}$ instead of $\boldsymbol{D}$ from (6) results in a 41% overestimate of the statistical detection capability of the system being considered, a situation which is even worse when compared to ***D*** from (8). This is shown in Figure 5 for a notional detection system where equation (8) assumes ***F*** = 3 which is a false alarm probability of 1/1000 (a common operational choice) for Gaussian distributions. This system is drawn from a publication[2] where realistic parameters were



used for a system in a field environment. The figure shows this same system as evaluated by all three metrics. The small ***B*** assumption corresponds to where the curves cross the "detection cut-off" where a metric = 3 was assumed to be the limit of detection. Where the equation (8) curve crosses the corresponds to a 99.9% detection probability for a Gaussian distribution, making that point a desirable statistical mark for real word detection systems. This curve crosses the line at somewhat under 70 meters, while equation (6) curve crosses at about 90 meters, and the $S/\sqrt{B}$ crosses at over 120 meters. This final incorrect metric over-predicts the system capability for weak sources by nearly a factor of two. Considering that the weak source case is the relevant one for important issues like loose nukes, this is an unacceptable large over-prediction.

Conversely, when ***S*** is large compared to ***B*** (***S*** >> ***B***) you can see in Figure 5 that ***D*** as determined by equations (6) and (8) merge at the point corresponding to 20 meters stand-off between source and detector. This is because source is strong so that the mean ***S*** is large and its distribution has little overlap with the background distribution used to set thresholds for false alarms. So for strong sources $\boldsymbol{D} \approx \boldsymbol{S}/\sqrt{\boldsymbol{S}}$ for both equations (6) and (8); the metric $\boldsymbol{S}/\sqrt{\boldsymbol{B}}$ fails to give any meaningful insight into the statistical significance of the derived quantity $\boldsymbol{S}$ in this case.

Recommendations

System designs based on inappropriate metrics are guaranteed to fall short of expectations. It is very important that program managers and system designers use the appropriate metrics for evaluation of a system design. This paper demonstrates the danger of accepting even a commonly used metric like $\boldsymbol{S}/\sqrt{\boldsymbol{B}}$. As defined, $\boldsymbol{D}$ is a basic metric that is simple to apply and can give a statistically valid estimate of the MDA a system can measure of a given source. When defined as in equation (6) it is useful for systems being used with strong sources, but must be used in the equation (8) form when looking for sources that are weak relative to the background environment in which they are being measured. The principles sketched out here can be used to select an appropriate metric for any system once all the intended operational parameters of the system are understood.

---

[1] Knoll GF, *Radiation Detection and Measurement, 3rd ed.* John Wiley & Sons, New York (2000).
[2] August RA, "Technology Independent Metrics that Bound the SNM Detection Problem," Journal of Homeland Security, 7 January 2008.