A Comparison of Three Methods for Calculating Confidence Intervals around D-Prime*

Abby Kaplan
UC Santa Cruz

June 2009

1 Introduction

Signal detection theory (Macmillan and Creelman 1991) is concerned with modeling how individuals identify signals in a noisy stream of data. One popular metric provided by SDT is d', which is a measure of how sensitive an individual is to whether a noisy input contains a given signal or not. Typically, an experiment involves presenting individuals with a number of trials containing only noise (n_N) and a number of trials that also contain the signal in question (n_S). The two measurements used to calculate d' are the probability that the individual will correctly identify a signal trial as containing the signal (p_h, the probability of a 'hit') and the probability that the individual will incorrectly identify a noise trial as containing the signal (p_f, the probability of a 'false alarm'). d' is defined as

$$d' = cdf_{0,1}^{-1}(p_h) - cdf_{0,1}^{-1}(p_f) \ , \tag{1}$$

where $cdf_{0,1}^{-1}(p)$ is the inverse-normal transform of p.

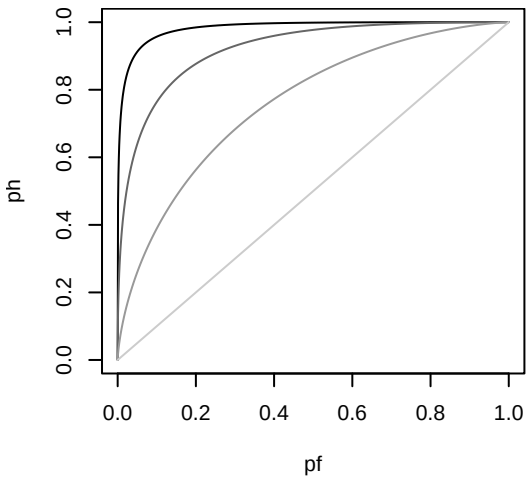
When p_h is equal to p_f, d' is 0 – that is, the individual is unable to distinguish signal trials from noise trials. When p_h is greater than p_f, d' is positive; the greater d' is, the more sensitive the individual is. Negative d's, which indicate that the subject can distinguish between the two types of trials but is responding to them incorrectly, are usually not considered. Figure 1 illustrates some of the values of d' associated with various points in the $p_f \times p_h$ space. Each curve joins all points in the space that yield a given d'. Curves are given for $d' = 3$ (black), 2 (dark gray), 1 (medium gray), and 0 (light gray).

Frequently, the goal of the experimenter is to compare observed $\hat{d}'$s[1] across different conditions, in order to determine the conditions in which individuals are more sensitive to

*Many thanks to Ryan Bennett, Keith Johnson, Grant McGuire, Jeremy O'Brien, Matt Tucker, and Paul Willis.

[1]Throughout, I use P to denote the true value of a given parameter and $\hat{P}$ to denote the value that is observed on some particular occasion.

Figure 1: d' as a function of p_h (y-axis) and p_f (x-axis). Each curve joins all points in the $p_f \times p_h$ space that yield a given d'. Curves are given for $d' = 3$ (black), 2 (dark gray), 1 (medium gray), and 0 (light gray).



the signal in question. One common procedure is to observe a large number of $\hat{d}'$s within each condition (for example, one $\hat{d}'$ for each experimental subject in an across-subjects design) and then compare the two (or more) groups of $\hat{d}'$s using a t-test or some other statistical measure. However, it is sometimes the case that there is no obvious or meaningful way to 'group' the observations from each condition to obtain sets of $\hat{d}'$s in this way – for example, if the experimenter wishes to compare the sensitivity of a single subject in two different conditions in a within-subjects design. In order to compare *individual* values of $\hat{d}'$, we must know the sampling distribution of d' for various values of p_h, p_f, n_S, and n_N.

This paper examines three methods for calculating the sampling distribution of various d's. The first involves an analytic approximation of the variance of d' presented by Gourevitch and Galanter (1967). The second method, described by Miller (1996), involves computing the full sampling distributions of a number of candidate d's in order to find the upper and lower bounds of the desired confidence interval around a given $\hat{d}'$. The third method is similar to Miller's, but uses the technique of maximum likelihood estimation in order to make the computations involved scalable to large datasets.

The appendix to this paper describes several functions written in R (R Development Core Team 2007) related to calculating confidence intervals according to each of these three methods, available from the author.

2 Method 1: Gourevitch and Galanter's (1967) *G*

Gourevitch and Galanter's (1967) G statistic tests whether two different observed $\hat{d}'$s are distinct. Among other assumptions, G relies on the approximation of the variance of $\hat{d}'$ given

in equation (2) (Gourevitch and Galanter's equation (5), with some symbols renamed):

$$Var(\hat{d'}) \approx \frac{\hat{p}_h(1-\hat{p}_h)}{n_S(ord\ \hat{z}_h)^2} + \frac{\hat{p}_f(1-\hat{p}_f)}{n_N(ord\ \hat{z}_f)^2} \tag{2}$$

As above, $\hat{p}_h$ and $\hat{p}_f$ are the observed probabilities of hits and false alarms, and n_S and n_N the number of signal and noise trials, respectively. The values $ord\ \hat{z}_h$ and $ord\ \hat{z}_f$ are the height of the normal density function at the inverse-normal transforms of $\hat{p}_h$ and $\hat{p}_f$.

If we assume that the sampling distribution of d' is normal, we can use the observed variance of $\hat{d'}$ to construct confidence intervals in the ordinary way: the 95% confidence interval for $\hat{d'}$, for example, is the interval spanning 1.96 standard deviations on either side of $\hat{d'}$. According to Miller (1996, 70), this is the standard method for computing confidence intervals around $\hat{d'}$.

Figure 2 illustrates the variances predicted by equation (2) for $n_S = n_N =$ 10, 20, 50, and 100 trials. As above, each graph plots hit rates on the y-axis and false alarm rates on the x-axis. The shading on each graph shows the predicted variance of $\hat{d'}$ for each possible combination of $\hat{p}_h$ and $\hat{p}_f$; darker shading represents a smaller variance. Note that the white region around the edge of each graph contains those points where either $\hat{p}_h$ or $\hat{p}_f$ is 0 or 1 and (2) is therefore undefined; naturally, the region shrinks as the number of trials increases. The dark contour lines represent discrete values of d' from 0 to 3 in increments of 0.5.

From these examples, we can observe three properties of Gourevitch and Galanter's approximation:

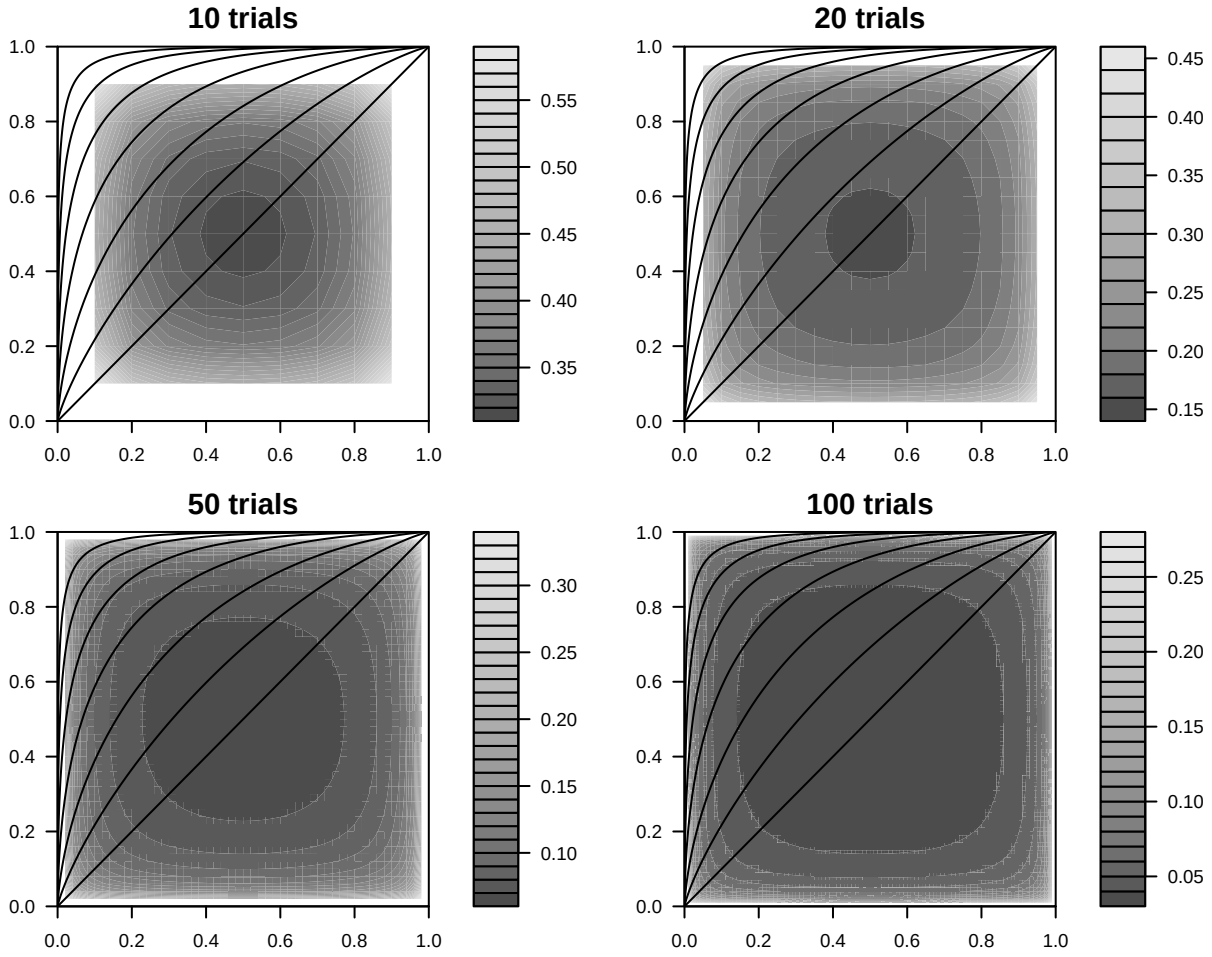
1. The variance decreases as the number of trials increases. We see this both from the scales associated with each graph (the range of the variances decreases as the number of trials increases) and from the graphs themselves (the smaller variances occupy larger portions of the $\hat{p}_f \times \hat{p}_h$ space as the number of trials increases). This is exactly as expected: we can put more confidence in a $\hat{d'}$ obtained from many trials than in one obtained from few trials.

2. The variance decreases when $\hat{p}_h$ and $\hat{p}_f$ are closer to 0.5. Again, this is expected: a small change in $\hat{p}_h$ (or $\hat{p}_f$) results in a small change to its inverse-normal transform when $\hat{p}_h$ is close to 0.5, but in a large change when $\hat{p}_h$ is close to 0 or 1.

3. The variance decreases when $\hat{p}_h$ is closer to $1-\hat{p}_f$ – that is, when there is less response bias.[2] (We can observe this in the graphs by noting that for each d' contour, the variance is smallest where the contour intersects the minor diagonal.) This is a corrollary of the previous observation: of all possible combinations of p_h and p_f that yield a given d', it is the one where $p_h = 1 - p_f$ that minimizes the distance between p and 0.5 for *both* ps.

[2]When $\hat{p}_h = \hat{p}_f$, the individual is said to be 'unbiased' because the probability of a correct response is the same for both signal and noise trials.

Figure 2: Gourevitch and Galanter's (1967) approximation of the variance of $\hat{d}'$ for various numbers of trials. The number of trials given for each graph is $n_S = n_N$. The y-axis of each graph plots possible observed hit rates and the x-axis possible false alarm rates. Shading gives the approximate variance for each combination of $\hat{p}_h$ and $\hat{p}_f$ (scales given separately for each graph). Dark contour lines plot discrete values of d' from 0 to 3 in increments of 0.5.



As an analytic approximation of the distribution of $\hat{d}'$, this technique has the important advantages of being simple to implement and computationally tractable. However, (2) assumes that the number of trials is large enough that the observed hit and false alarm rates converge on a normal distribution – an assumption that may not always be met in practice.

3 Method 2: Direct Computation with Assumption of No Bias

3.1 Probability Mass Function of $\hat{d}'$

As discussed above, $\hat{d}'$ is a function of two observed proportions, $\hat{p}_h$ and $\hat{p}_f$. If we assume that $\hat{p}_h$ and $\hat{p}_f$ follow binomial distributions, then the distribution of $\hat{d}'$ can be calculated accordingly, as detailed in the rest of this section. This distribution can then be used in various ways to calculate confidence intervals for $\hat{d}'$.

The probability mass function (pmf) of a discrete random variable is a function that returns, for each possible value of the variable, the probability that the random variable has that value. (For example, the pmf of a random variable representing tosses of a fair coin returns .5 for 'heads' and .5 for 'tails'.) The cumulative distribution function (cdf) of a random variable is a function such that the value of the function at a point a is equal to the probability that the random variable has a value less than or equal to a.

H, the number of observed hits, and F, the number of observed false alarms, can be modelled as binomial random variables with parameters n_S or n_N and p_h or p_f, respectively. The pmf of a binomial(n,p) random variable X is

$$pmf_X(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{(n-x)} \ . \tag{3}$$

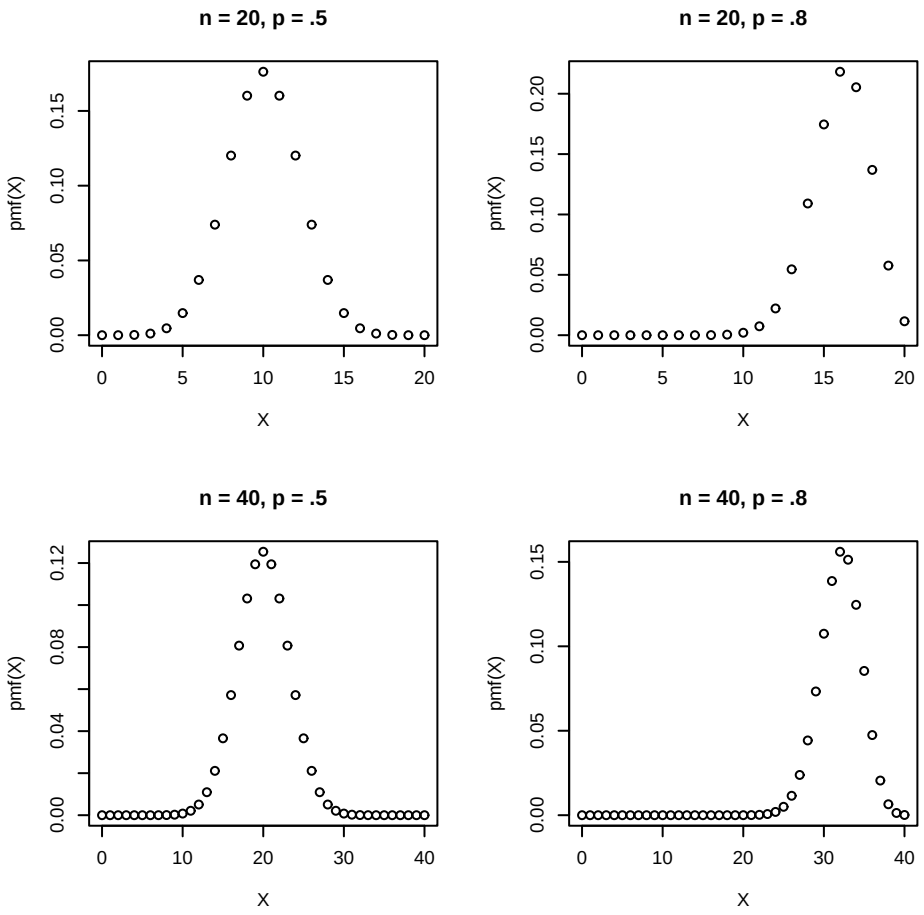
Figure 3 shows the pmfs of binomial random variables with various values of n and p.

As discussed above, $\hat{d}'$ is calculated, not from H and F directly, but rather from the z-transformed hit rate $\frac{H}{n_S}$ and the z-transformed false alarm rate $\frac{F}{n_N}$. Thus, we define two new random variables $X = cdf_{0,1}^{-1}(\frac{H}{n_S})$ and $Y = cdf_{0,1}^{-1}(\frac{F}{n_N})$. It is possible to determine the pmfs of X and Y, as follows (Casella and Berger 2002, 48): define $g(h) = cdf_{0,1}^{-1}(\frac{h}{n_S})$ and sets $\mathcal{H}$ and $\mathcal{X}$ such that $\mathcal{H} = \{h : pmf_H(h) > 0\}$ (that is, the set of all values of H that have a non-zero probability of occurring) and $\mathcal{X} = \{x : x = g(h), h \in \mathcal{H}\}$ (that is, the set of all values of X that are derived from possible values of H). Then for $x \in \mathcal{X}$,

$$\begin{aligned} pmf_X(x) &= \sum_{h \in g^{-1}(x)} pmf_H(h) && (4) \\ &= \sum_{h \in n_S (cdf_{0,1}^{-1})^{-1}(x)} pmf_H(h) && (5) \\ &= \sum_{h \in n_S cdf_{0,1}(x)} pmf_H(h) && (6) \\ &= pmf_H(n_S cdf_{0,1}(x)) && (7) \end{aligned}$$

and similarly for pmf_Y. In other words, the probability of observing a given value of X is the sum of the probabilities of observing any value of H that yields that value of X. Note that

Figure 3: Pmfs of binomial random variables with various settings of n and p



(6) simplifies to (7) because $cdf_{0,1}$ is a monotonically increasing function. Figure 4 shows the pmfs of the z-transformations of the same binomial random variables shown in figure 3.

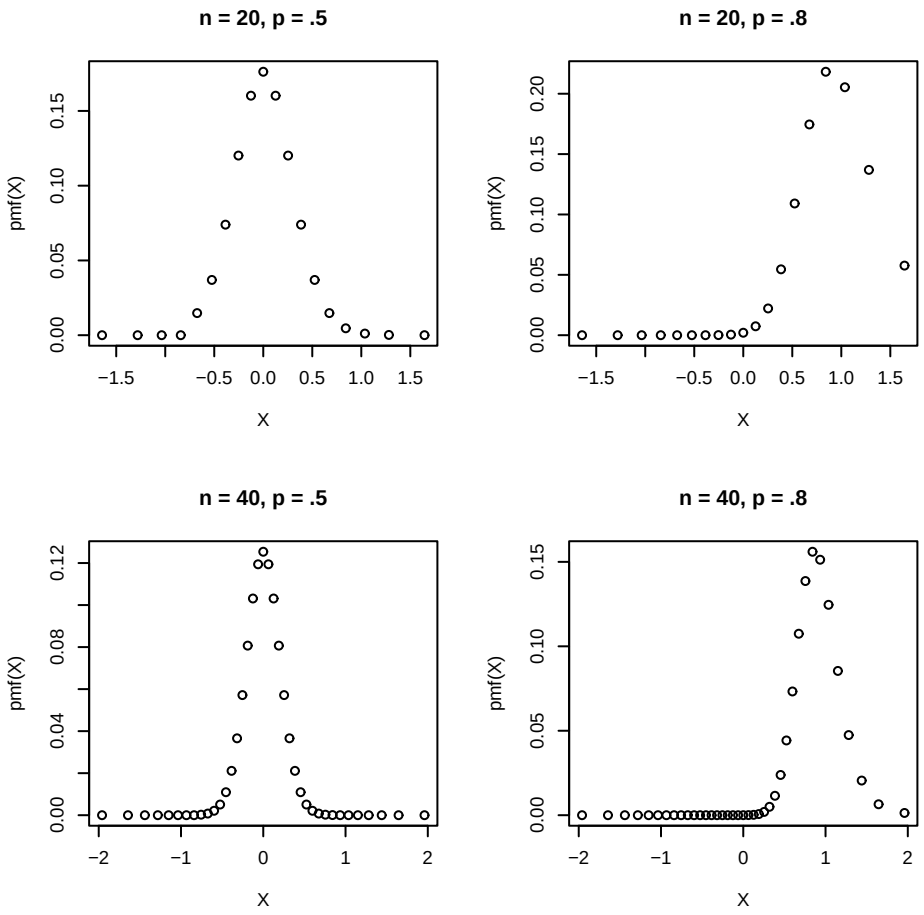
If H and F are independent[3], then X and Y are independent, and the joint pmf of X and Y is

$$pmf_{XY}(x, y) = pmf_X(x)pmf_Y(y) \ . \tag{8}$$

Since d' is the difference between the z-transformed hit rate and the z-transformed false alarm rate, we define a new random variable $D = X - Y$. Given the sets $\mathcal{A} = \{(x, y) :$

[3]Note that this assumption is not the same as saying that the probability of observing a given hit rate is independent of the probability of observing a given false alarm rate. Rather, this assumption claims that given the true probabilities p_h and p_f of observing a given hit or false alarm rate, the difference between the observed $\hat{p}_h$ and the true p_h is independent of the diffference between the observed $\hat{p}_f$ and the true p_f.

Figure 4: Pmfs of z-transformed binomial random variables with various settings of n and p



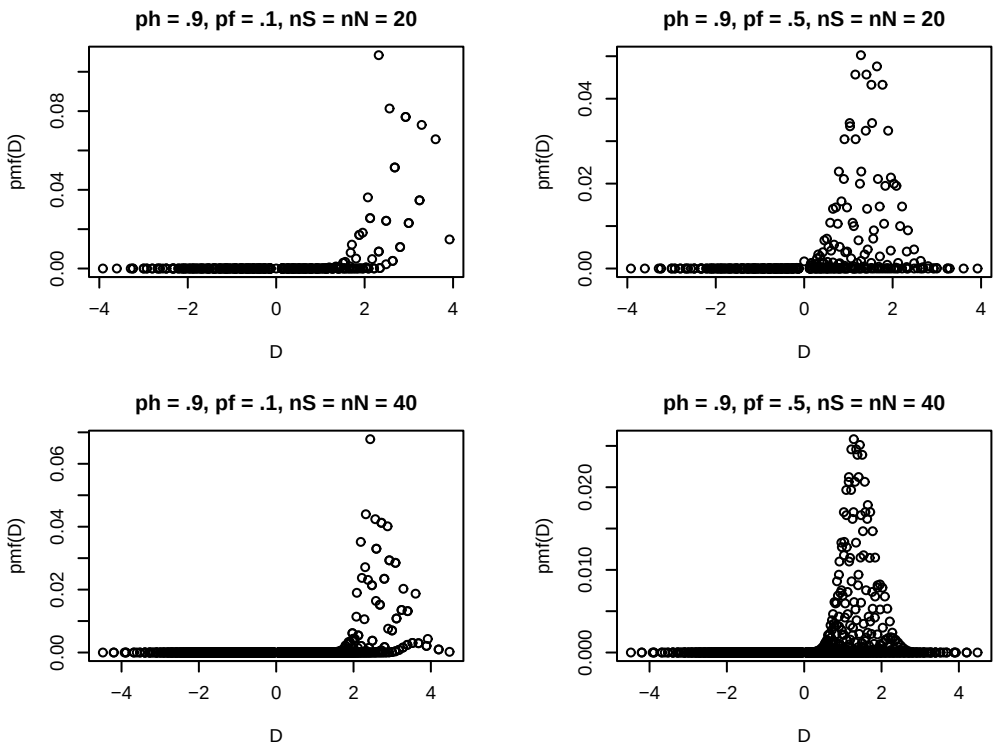
$pmf_{XY}(x, y) > 0\}$, $\mathcal{B} = \{d : d = x - y, (x, y) \in \mathcal{A}\}$, and $A_d = \{(x, y) \in \mathcal{A} : x - y = d\}$, we can define the pmf of D as follows (Casella and Berger 2002, 157):

$$pmf_D(d) = \sum_{(x,y)\in A_d} pmf_{XY}(x, y) \tag{9}$$

In other words, the probability of observing a given value of D is equal to the sum of the probabilities of every combination of (z-transformed) hit rate and false alarm rate that would result in that particular value of D. Figure 5 shows the pmfs of D for various values of p_h, n_S, p_f, and n_N.

The most salient characteristic of pmf_D is that it does not produce a sequence of points that fall along a simple curve; rather, the probabilities associated with the $\hat{d}'$s in the graphs in figure 5 appear to fall inside a region defined by such a curve. (This is most apparent when d' is small and n_S and n_N are large.) Although perhaps initially surprising, this is

Figure 5: Pmfs of d' with various settings of p_h, n_S, p_f, and n_N

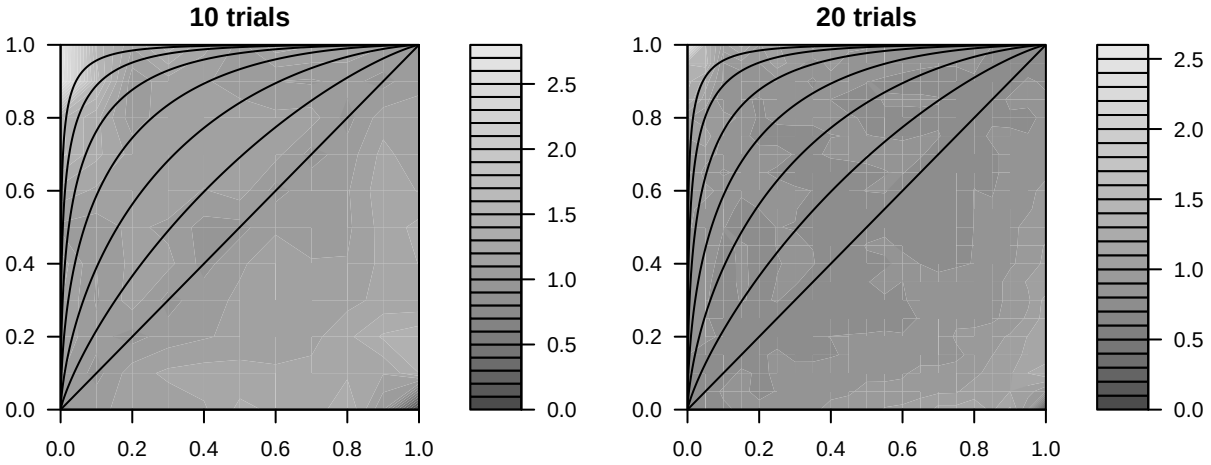


exactly the right result. Many of the points that fall on the edge of what appears to be the curve are those that fall within the normal range of expected variation. For example, when $p_h = .9$, $p_f = .1$, and $n_S = n_N = 20$ (the upper left graph in figure 5), the probability of observing a $\hat{d}'$ of 2.563 (equal to the true d') is .0813, and the probability of observing a $\hat{d}'$ of 2.926 (if $\hat{p}_h$ is .95 or $\hat{p}_f$ is .05) is also relatively high, at .0770. (In fact, the $\hat{d}'$ that is most likely to be observed is 2.318 – *not* the true d'! This is because there are several possible observations of relatively high probability, such as $p_h = .9$ and $p_f = .15$, that all yield a $\hat{d}'$ of 2.318.) There are also values of $\hat{d}'$ that are very close to the true value of 2.563 but involve such exotic observations that their probabilities are vanishingly small. For example, if $n_S = n_N = 40$, then it is possible to obtain a $\hat{d}'$ of 2.558 if $\hat{p}_h = .975$ and $\hat{p}_f = .275$ (or $\hat{p}_h = .725$ and $\hat{p}_f = .025$). However, the probability of either of these observations is extremely small (.0000715). It is these unlikely observations that happen to produce $\hat{d}'$s similar to the most likely ones that 'fill out' the interior of the curve.

3.2 Miller's (1996) Direct Computation of Confidence Intervals

Miller (1996, 70) outlines a procedure for computing a confidence interval around a given $\hat{d}'$ directly from the sampling distribution of d', as follows:

Figure 6: Computed upper bounds of 95% confidence intervals for $\hat{d}'$s for various numbers of trials, according to Miller's (1996) method. The number of trials given for each graph is $n_S = n_N$. The y-axis of each graph plots possible observed hit rates and the x-axis possible false alarm rates. Shading gives the size of the upper half of the 95% confidence interval for each combination of $\hat{p}_h$ and $\hat{p}_f$ (scales given separately for each graph). Dark contour lines plot discrete values of d' from 0 to 3 in increments of 0.5.



1. Select a value of d' as a candidate for the upper (or lower) bound of the confidence interval around $\hat{d}'$.

2. Compute the sampling distribution of d', choosing p_h and p_f such that $p_h = 1 - p_f$ and keeping n_S and n_N the same as in the experiment.

3. If the observed $\hat{d}'$ is located at the appropriate percentile of the sampling distribution of d' (for example, the 2.5th percentile for the upper bound of a 95% confidence interval), stop. Otherwise, choose another candidate d' and go back to step 2.

Miller does not specify how new candidates for d' are selected, nor does he specify when the search terminates. In my implementation of his procedure, selection of new candidate d's procedes by binary search, and the search terminates when the appropriate d' has been identified to a precision of 5 decimal places. This procedure replicates the values given in Miller's table 3 to within three decimal places.

Figure 6 illustrates the confidence intervals calculated according to Miller's method for $n_S = n_N = 10$ and 20 trials. Analyzing larger numbers of trials was not computationally feasible: as the number of trials increases, so does the number of computations required to determine the full sampling distribution of a given d', and so does the number of possible values of $\hat{p}_h$ and $\hat{p}_f$. It is the latter problem that proved to be fatal; calculating confidence intervals for single $\hat{d}'$s with 50 or 100 observations is quite feasible with this method. The layout of the graphs is analogous to that of the graphs in figure 2.

From these examples, we can learn four things about the confidence intervals produced by this method:

1. The size of the confidence interval shrinks as the number of trials increases, as expected.

2. The size of the upper half of the confidence interval approaches 0 as $\hat{d}'$ approaches negative infinity. This is a consequence of the non-normality of the distribution of $\hat{d}'$. As the two left-hand graphs in figure 5 illustrate, as d' moves further away from 0, more of the probability mass of its sampling distribution becomes concentrated *above* d'. In other words, $\hat{d}'$ is a biased estimator of d', tending to yield more extreme values than the true d'. At extreme values of d', the half of the confidence interval around $\hat{d}'$ that is closer to 0 (the upper half when d' is small and the lower half, not illustrated in figure 6, when d' is large) will approach 0.

3. As was the case with Gourevitch and Galanter's approximation, the size of the confidence interval decreases as $\hat{p}_h$ approaches $1 - \hat{p}_f$. However, in this case, the size of the confidence interval is *not* monotonically decreasing as the distance between $(\hat{p}_f, \hat{p}_h)$ and $(.5, .5)$ decreases.

4. The graphs are not symmetric across the minor diagonal. It is not clear to me why this should be the case.

Because Miller's method is based on fully computed distributions of relevant $\hat{d}'$s rather than an approximation, we expect the confidence intervals produced by this method to be more accurate than those derived from Gourevitch and Galanter's approximation. However, recall that sampling distributions are computed only for candidate d's with no bias – that is, d's where $p_h = 1 - p_f$. This assumption of no bias is likely to yield inaccuracies when the observed $\hat{p}_h$ and $\hat{p}_f$ do not have this property. In addition, as discussed above, it is precisely when $p_h = 1 - p_f$ that the sampling variance of d' is smallest; thus, it is likely that Miller's method tends to underestimate the true variance of $\hat{d}'$.

4 Method 3: Maximum Likelihood Estimation

Miller's (1996) method for calculating the confidence interval around $\hat{d}'$ is computationally costly because it requires computation of the entire pmf of *every* d' selected as a candidate for the upper (or lower) bound of the confidence interval. Fortunately, it is possible to obtain a confidence interval by computing the pmf of the single $\hat{d}'$ that is actually observed[4], using the technique of maximum likelihood estimation.

Maximum likelihood estimation makes use of the likelihood function, defined as follows (Casella and Berger's (2002) definition 6.3.1):

[4]Actually, as will be seen below, only the joint pmf of $\hat{p}_h$ and $\hat{p}_f$ is needed, further reducing the computational cost associated with this method.

Definition 1 *Let $f(x|\theta)$ denote the joint pdf or pmf of the random sample $\mathbf{X} = (X_1, \ldots, X_n)$. Then, given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by*

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

is called the likelihood function.

Informally, where a pmf returns the probability of an observation for a given parameter setting, a likelihood function returns the probability of a parameter setting for a given observation. For the purposes of a single observed $\hat{d'}$, the random sample of definition 1 consists of a single ordered pair (h, f) of the observed number of hits and false alarms. θ, on the other hand, is the ordered pair (p_h, p_f), the true hit and false alarm rates. (Note that n_S and n_N are fixed by the number of trials actually carried out.) Thus, the likelihood function is

$$\begin{aligned} L((p_h, p_f)|(h, f)) &= pmf_{HF}((h, f)|(p_h, p_f)) && (10) \\ &= pmf_H(h|p_h)pmf_F(f|p_f) \ . && (11) \end{aligned}$$

Figure 7 shows the likelihood functions for observations from binomial distributions with various sample sizes. Note that the most likely value of p is the one that is actually observed, and that the spread of the likelihood function becomes wider as the sample size decreases. Note also that while pmf_H and pmf_F are discrete, L is continuous: it is defined over proportions, which can take any real value between 0 and 1. The parameter setting p_h that defines the shape of pmf_H can be any proportion, even though the output of pmf_H is always an integer.

We can now define the maximum likelihood estimator (MLE) for observations of H and F, following definition 7.2.4 of Casella and Berger (2002, 316):

Definition 2 *For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A maximum likelihood estimator (MLE) of the parameter θ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{x})$.*

In other words, the MLE of $(\hat{p}_h, \hat{p}_f)$ – our best guess at the true probabilities that a subject will produce hits and false alarms – is the value of $(\hat{p}_h, \hat{p}_f)$ that maximizes $L((p_h, p_f)|(h, f))$.

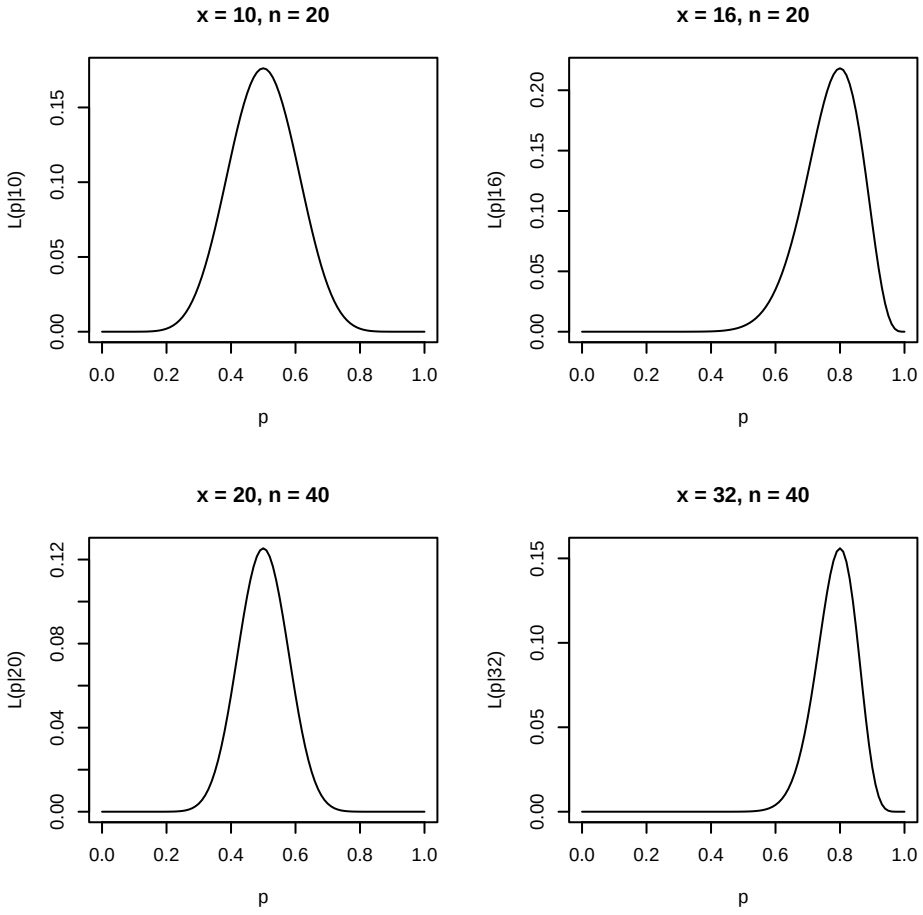
Of course, we are not interested in (p_h, p_f) directly, but rather in d', which is a *function* of p_h and p_f. Recall that d' is defined as

$$d'(p_h, p_f) = cdf_{0,1}^{-1}(p_h) - cdf_{0,1}^{-1}(p_f) \ . \quad (12)$$

Fortunately, MLEs have a useful property known as *invariance* (Casella and Berger 2002, 320):

Theorem 1 *If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.*

Figure 7: Likelihood functions for observations from binomial distributions with various settings for n



In other words, if we know the MLE $\hat{\theta}_{HF}$ of (p_h, p_f), then the MLE of d' for the same data is $d'(\hat{\theta}_{HF})$.

The MLE is the foundation of the likelihood ratio test (LRT). The LRT statistic is defined as in Definition 3 (Casella and Berger 2002, 375):

Definition 3 *The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^C$ is*

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})} .$$

A likelihood ratio test (LRT) is any test that has a rejection region of the form $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$,

where c is any number satisfying $0 \leq c \leq 1$.

For example, suppose we want to test the hypothesis that d' is less than 2, given the observation (h, f). To use the LRT statistic, we would find the maximum of $L(\theta|(h, f))$ when θ is restricted to the pairs (p_h, p_f) that yield a d' less than 2, and divide that number by the maximum of $L(\theta|(h, f))$ with no restrictions on θ. The result is the confidence level α with which we can reject the hypothesis that d' is less than 2.

Finally, we can invert the LRT statistic to obtain confidence intervals, as follows (Casella and Berger 2002, 421-422):

Theorem 2 *For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each $\mathbf{x} \in \mathcal{X}$, define a set $C(\mathbf{x})$ in the parameter space by*

$$C(\mathbf{x}) = \{\theta_0 : \mathbf{x} \in A(\theta_0)\}.$$

Then the random set $C(\mathbf{X})$ is a $1 - \alpha$ confidence set. Conversely, let $C(\mathbf{X})$ be a $1 - \alpha$ confidence set. For any $\theta_0 \in \Theta$, define

$$A(\theta_0) = \{\mathbf{x} : \theta_0 \in C(\mathbf{x})\}.$$

Then $A(\theta_0)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.

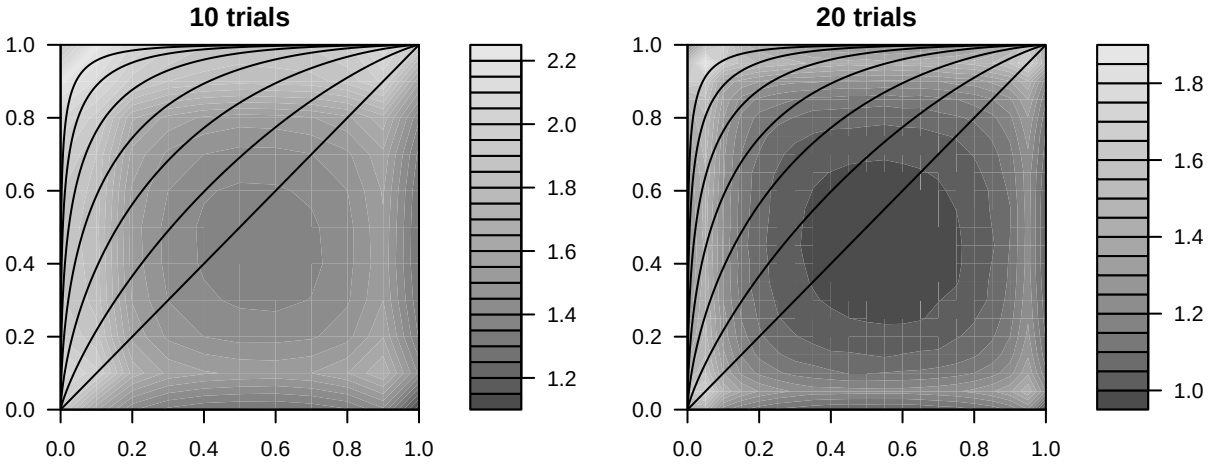
In other words, the 95% confidence set around an observed d' is the set of all d's that result from values of (p_h, p_f) that are not rejected by the LRT statistic at $\alpha = .05$.

Since p_h and p_f may take on any real value between 0 and 1 and d' is a continuous function of (p_h, p_f), the confidence set for a given $\hat{d'}$ is actually a confidence interval. However, since there is no analytic solution to maximizing the likelihood function of $\hat{d'}$ in the manner described above, we must procede by directly computing the likelihood function for various values of (p_h, p_f) to some arbitrary level of precision. In my implementation, I found that computing these values to the third decimal place was computationally feasible. The upper and lower bounds yielded by this procedure are generally accurate to the second decimal place when $\hat{p}_h$ and $\hat{p}_f$ are close to 0.5, and much more accurate when $\hat{p}_h$ and $\hat{p}_f$ are more extreme. Note that since the likelihood function is a function of the parameter values (p_h, p_f) with $\hat{p}_h$, $\hat{p}_f$, n_S, and n_N fixed, the length of time required to compute the confidence interval depends on the precision to which (p_h, p_f) is calculated, but *not* on the number of trials.

Figure 8 shows the size of the upper halves of the confidence intervals for $\hat{d'}$s for various combinations of $\hat{p}_h$ and $\hat{p}_f$ for $n_S = n_N = 10$ and 20 trials. As was the case for Miller's method, calculations for more trials was not computationally feasible because the number of possible combinations of $\hat{p}_h$ and $\hat{p}_f$ increases with the number of trials. The layout of the graphs is the same as that for figures 2 and 6.

These graphs reveal four things about the confidence intervals computed by maximum likelihood estimation:

Figure 8: Computed upper bounds of 95% confidence intervals for $\hat{d}'$s for various numbers of trials, according to maximum likelihood estimation. The number of trials given for each graph is $n_S = n_N$. The y-axis of each graph plots possible observed hit rates and the x-axis possible false alarm rates. Shading gives the size of the upper half of the 95% confidence interval for each combination of $\hat{p}_h$ and $\hat{p}_f$ (scales given separately for each graph). Dark contour lines plot discrete values of d' from 0 to 3 in increments of 0.5.



1. The overall pattern by which the size of the confidence interval depends on the location of $\hat{d}'$ in the $\hat{p}_f \times \hat{p}_h$ space is strikingly similar to the pattern produced by Gourevitch and Galanter's approximation.

2. As was the case for both of the previous methods, the size of the upper half of the confidence interval decreases when $\hat{p}_h$ or $\hat{p}_f$ approaches 0.5 or when the number of trials increases.

3. As was the case for Miller's method, the size of the upper half of the confidence interval is smaller for a $\hat{d}'$ of less than 0 than for its positive counterpart. This can be seen by the fact that the concentric shaded regions in the graphs in figure 8 are centered slightly below and to the left of $(0.5, 0.5)$.

4. The size of the upper half of the confidence interval appears to reach a maximum when $\hat{p}_h$ or $\hat{p}_f$ reaches near-perfect performance (that is, exactly one miss or one false alarm), and decreases again with perfect performance.

5 Comparison

5.1 Simplifying Assumptions

Of the three methods described here for computing confidence intervals around $\hat{d}'$, it is Gourevitch and Galanter's (1967) approximation that makes the most simplifying assump-

tions – most obviously, that the sampling distribution of d' is normal. As illustrated in figure 5, if the sampling distributions of p_h and p_f are binomial (a belief to which even Gourevitch and Galanter subscribe, pg. 30), this is simply not the case. The assumption of normality is best satisfied when $\hat{d}'$ is small and the number of trials is large; thus, it is safest to use Gourevitch and Galanter's approximation when those conditions are met.

Miller's (1996) method for direct computation of confidence intervals around $\hat{d}'$ is a great improvement on this point. However, to make his method computationally feasible, he is forced to test only alternative hypotheses where there is no response bias (where $p_h = 1 - p_f$). As discussed above, it is in precisely this circumstance where the sampling distribution of d' is expected to be lowest; thus, there is a danger that Miller's method is not conservative enough, yielding confidence intervals that are too small.

The maximum likelihood estimation technique shares with both of the other methods described here the assumption that the sampling distributions of p_h and p_f are underlyingly normal. As implemented here, maximum likelihood estimation improves on Miller's method by considering cases in which there *is* a response bias. However, this method is only as good as the maximum likelihood estimation technique in general. Although MLEs generally have very good behavior for large sample sizes, this is not necessarily the case for smaller sample sizes.

Therefore, it is impossible to know *a priori* which of the three methods described here yields confidence intervals that most accurately reflect the sampling distribution of d'. The most accurate calculation would involve a more computationally intense version of Miller's method that considered candidate d's with any possible combination of p_h and p_f; however, this undertaking was not feasible for the present paper.

5.2 Computational Tractability

As an analytic solution to the problem of finding confidence intervals, Gourevitch and Galanter's approximation is by far the most efficient of these three methods. When many such confidence intervals must be computed, or when the dataset is very large, researchers may find this method to be the most feasible – provided that the number of trials is large enough and that $\hat{d}'$ is small enough.

Miller's method is the least tractable of the three. This method requires computing the full pmf of every d' selected as a candidate for the upper or lower bound of the confidence interval. Because the number of computations required for a given pmf increases polynomially with the number of trials, this method may not scale well to large datasets. However, in my implementation, I found the time required to compute confidence intervals to be quite reasonable for values of n_S and n_N up to at least a few hundred.

The time required to compute confidence intervals using maximum likelihood estimation depends not on the size of the dataset but on the desired level of precision. The computations that are actually carried out are essentially the same as those for Miller's method (since the likelihood function is the same as the pmf, but varies the parameters rather than the observations); importantly, however, only *one* distribution needs to be computed. Thus, while not nearly as fast as Gourevitch and Galanter's approximation, this method does

represent an improvement over Miller's method for large datasets. In producing the graphs in the next section, I found that Miller's method was substantially faster than maximum likelihood estimation for $n_S = n_N = 40$, but substantially slower for $n_S = n_N = 80$.

5.3 Size of Confidence Intervals

The following graphs compare the upper bounds of the confidence intervals produced by the three methods described in this paper. Upper bounds were computed for every combination of $p_h, p_f \in \{.1, .2, \ldots, .8, .9\}$ and $n_S = n_N \in \{10, 20, 40, 80\}$. Within each block, each graph represents a single setting for $n_S = n_N$ and plots all of the resulting upper bounds (one for each combination of p_h and p_h) for one method against those of another method. The symbols used in the scatterplots represent the approximate size of the $\hat{d}'$ associated with each upper bound: 0 for $\hat{d}'$s of 0, 1 for $\hat{d}'$s with an absolute value of up to 1, and so on. Black symbols are used for positive $\hat{d}'$s (and 0) and gray symbols for negative values. The upper bound of a confidence interval for a negative d' is expected to correspond to the lower bound of the confidence interval for the corresponding positive d'. Also included in each graph is the line $x = y$.

As noted above, for all three methods, the size of the confidence interval decreases as the number of trials increases, as expected. In addition, the size of the confidence interval tends to increase as $\hat{d}'$ increases, although not necessarily monotonically.

The maximum likelihood estimation method exceptionlessly produces the largest confidence intervals, and Miller's method generally produces the smallest.

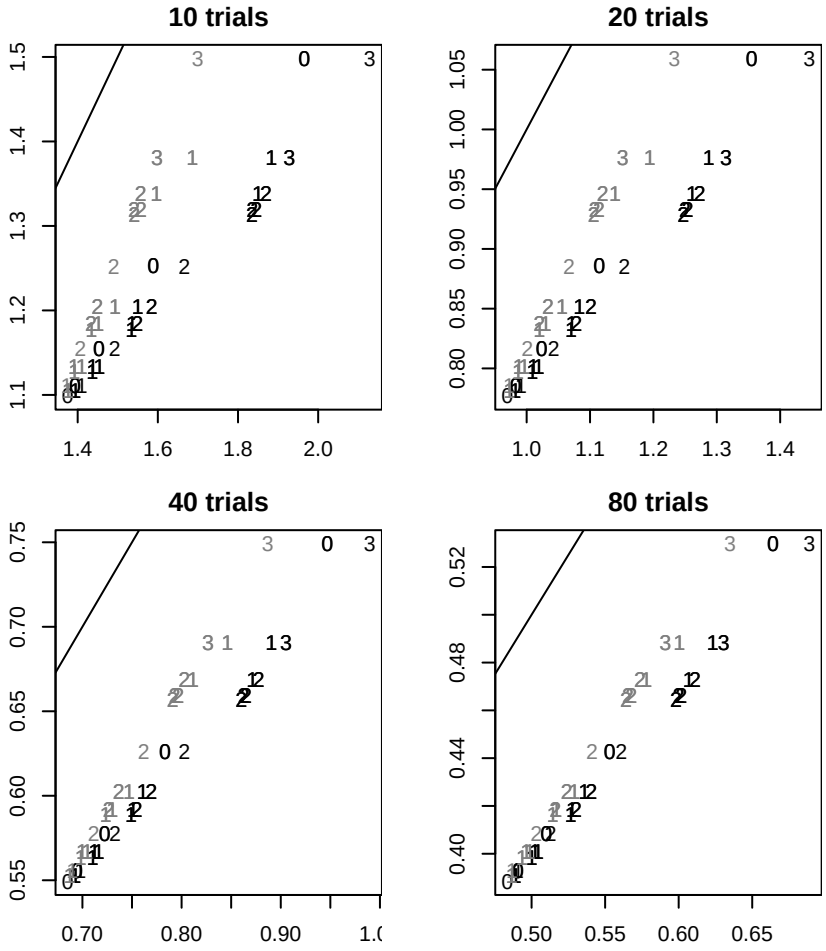
Gourevitch and Galanter's approximation and the maximum likelihood estimation method appear to be near-linear transforms of each other, although the slope of the relevant line is slightly different for upper and lower bounds, judging from the fact that the upper bounds for positive and negative $\hat{d}'$s appear lie on lines with slightly different slopes in figure 9. Note that the difference in slopes decreases as the number of trials increases, suggesting that with sufficiently large numbers of trials, the sampling distribution of d' does indeed approach a normal (or at least symmetrical) distribution, as assumed by Gourevitch and Galanter's approximation.

Neither Gourevitch and Galanter's approximation nor the maximum likelihood estimation method shows any signs of converging with the results of Miller's method as the number of trials increases (although the absolute differences between the locations of the relevant confidence interval boundaries will, of course, decrease as the sizes of the intervals themselves decrease).

6 Conclusion

At least two methods for calculating the confidence interval around a single $\hat{d}'$ have been proposed in the literature: the approximate analytical solution of Gourevitch and Galanter (1967) and the exact computation (with assumption of no response bias) described by Miller

Figure 9: Upper bounds of the 95% confidence intervals around various $\hat{d}'$s as computed by Gourevitch and Galanter's (1967) approximation (y-axis) compared to maximum likelihood estimation (x-axis). Symbols give the approximate value of the $\hat{d}'$ associated with each upper bound. Black symbols represent non-negative $\hat{d}'$s and gray symbols negative $\hat{d}'$s. Also included is the line $x = y$.

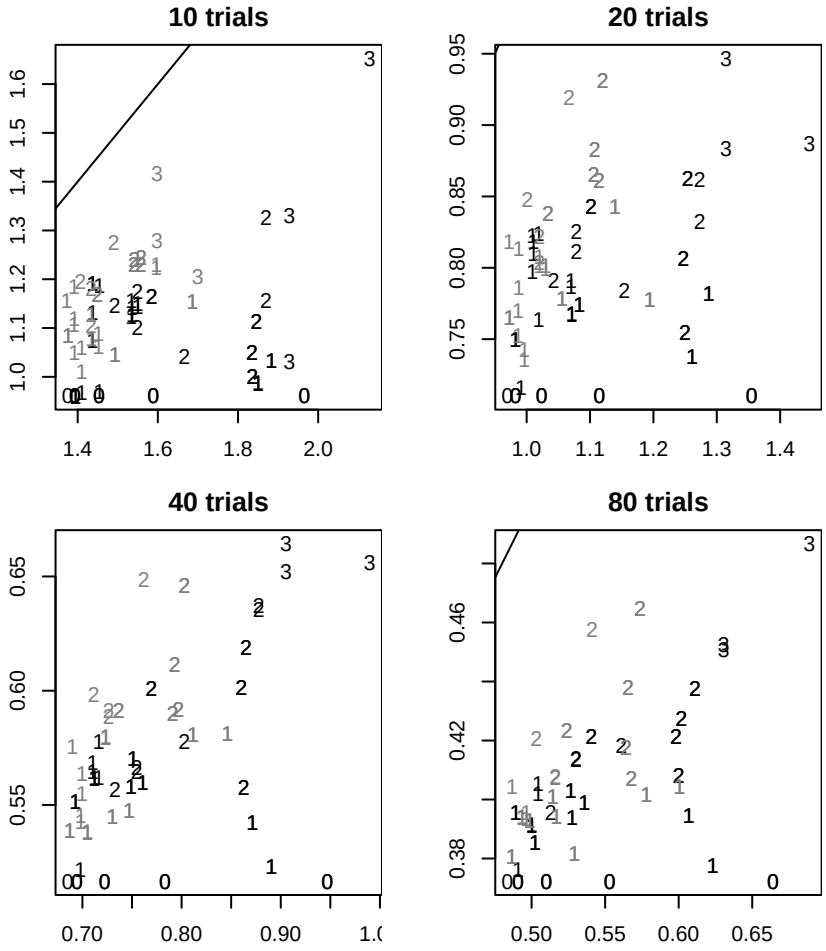


(1996).[5] A third method has not, to my knowledge, been previously proposed: that of maximum likelihood estimation.

Although it is the most assumption-laden method of the three, Gourevitch and Galanter's approximation has the very real advantage of being computationally simple and efficient. The other two methods involve much more intensive computations, although they are feasible for $\hat{d}'$s involving at least a few hundred observations (or more, for maximum likelihood estimation). It is impossible to say which of the two latter methods yields more accurate

[5]A third type of approach, involving Monte Carlo simulation (Kadlec 1999), is not discussed here.

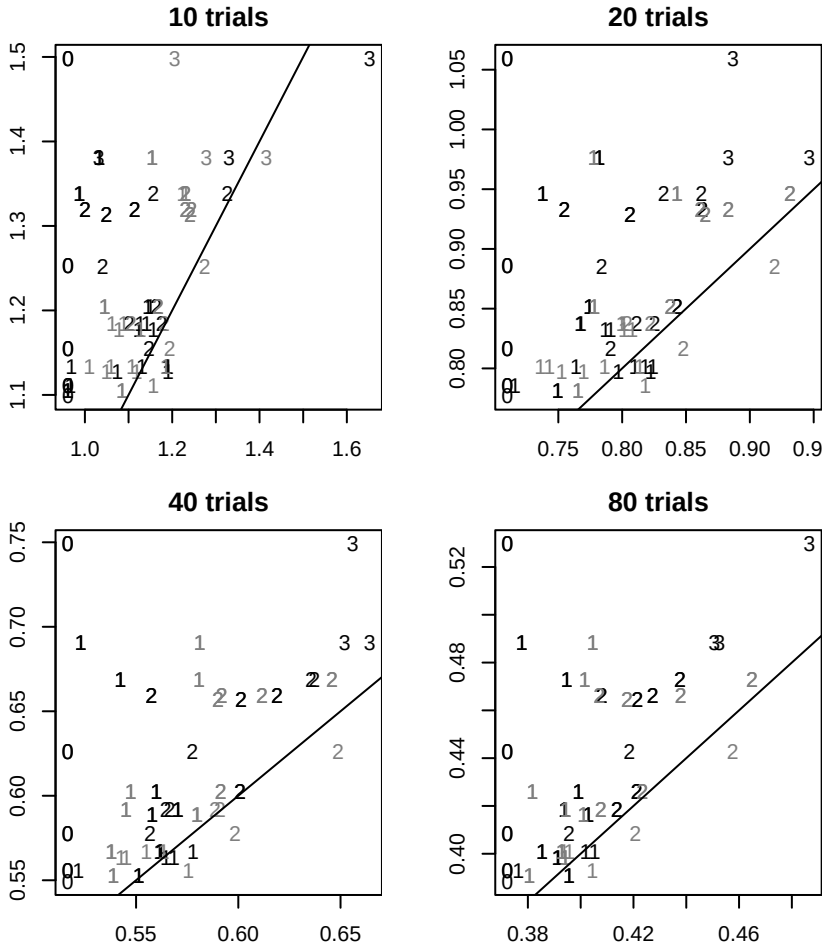
Figure 10: Upper bounds of the 95% confidence intervals around various $\hat{d}'$s as computed by Miller's (1996) method (y-axis) compared to maximum likelihood estimation (x-axis). Symbols give the approximate value of the $\hat{d}'$ associated with each upper bound. Black symbols represent non-negative $\hat{d}'$s and gray symbols negative $\hat{d}'$s. Also included is the line $x = y$.



results, since Miller's method considers only alternative hypotheses with no response bias and the maximum likelihood estimation method is only as good as maximum likelihood estimation itself.

Researchers working with data that seems to involve little or no response bias may find Miller's method to be the most desirable. Those working with extremely large datasets may find that only Gourevitch and Galanter's approximation is feasible. Those working with datasets with extreme values of $\hat{p}_h$ or $\hat{p}_f$ may find it desirable to use the maximum likelihood estimation method as the one whose assumptions are least violated by the data at hand. In

Figure 11: Upper bounds of the 95% confidence intervals around various $\hat{d}'$s as computed by Gourevitch and Galanter's (1967) approximation (y-axis) compared to Miller's (1996) method (x-axis). Symbols give the approximate value of the $\hat{d}'$ associated with each upper bound. Black symbols represent non-negative $\hat{d}'$s and gray symbols negative $\hat{d}'$s. Also included is the line $x = y$.



addition, the large confidence intervals produced by maximum likelihood estimation make it the most conservative of the three methods.

Appendix: R Code

This section describes several functions written in R (R Development Core Team 2007) that were used in calculating the confidence intervals in this paper, available from the author. Most of the functions described below come in two versions: one whose name ends in `count`

and one whose name ends in `list`. The former functions require as arguments integers representing the number of hits, misses, false alarms, and correct rejections observed. The latter require a single argument in place of those four consisting of a vector of strings, where each `"H"` in the vector corresponds to an observed hit, each `"M"` to a miss, each `"F"` to a false alarm, and each `"R"` to a correct rejection.

`dprime.count, dprime.list` These functions compute d' from the given number of hits, misses, false alarms, and correct rejections. Perfect hit and false alarm rates are adjusted upward (or downward) by `0.5`.

`G.int.count, G.int.list` These functions compute confidence intervals around the $\hat{d}'$ corresponding to the observed data according to Gourevitch and Galanter's (1967) approximation. There is no adjustment for perfect hit or false alarm rates. The functions compute `1 - a` confidence intervals, where `a` is specified in an additional argument; the default is `0.05`. The functions return a numeric vector of the lower and upper bounds of the confidence interval.

`dprime.pmf` This function computes the sampling distribution of a given d' (that is, its pmf); it takes as arguments values for p_h, n_S, p_f, and n_N. It returns a two-column matrix where the first column contains an ordered list of possible $\hat{d}'$s and the second column their corresponding probabilities.

`dprime.pmf.restricted` This function is the same as `dprime.pmf`, but it returns values only for $\hat{d}'$s above (or below) a certain value. The limit, and whether that limit is a maximum or a minimum (`"max"` or `"min"`), are specified in two additional arguments. Using this function rather than `dprime.pmf` makes Miller's method slightly more efficient.

`pmf.int.count, pmf.int.list` These functions compute confidence intervals around the $\hat{d}'$ corresponding to the observed data according to Miller's (1996) method. There is no adjustment for perfect hit or false alarm rates. The functions compute `1 - a` confidence intervals, where `a` is specified in an additional argument; the default is `0.05`. The functions return a numeric vector of the lower and upper bounds of the confidence interval.

`MLE.int.count, MLE.int.list` These functions compute confidence intervals around the $\hat{d}'$ corresponding to the observed data according to the maximum likelihood estimation method. Perfect hit and false alarm rates are adjusted upward or downward by `0.5`. The functions compute `1 - a` confidence intervals, where `a` is specified in an additional argument; the default is `0.05`. The difference between discrete values of $\hat{p}_h$ and $\hat{p}_f$ is given in an additional argument; the default is `0.001`. The functions return a numeric vector of the lower and upper bounds of the confidence interval.

References

George Casella and Roger L. Berger. *Statistical Inference*. Thompson Learning Asia and China Machine Press, second (international) edition, 2002.

Vivian Gourevitch and Eugene Galanter. A significance test for one parameter isosensitivity functions. *Psychometrika*, 32(1):25–33, 1967.

Helena Kadlec. Statistical properties of d' and β estimates of signal detection theory. *Psychological Methods*, 4:22–43, 1999.

Neil A. Macmillan and C. Douglas Creelman. *Signal Detection: A User's Guide*. Cambridge University Press, Cambridge, 1991.

Jeff Miller. The sampling distribution of d'. *Perception and Psychophysics*, 58(1):65–72, 1996.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, 2007.