Perception & Psychophysics
1996, 58 (1), 65–72

The sampling distribution of d'

JEFF MILLER
University of Otago, Dunedin, New Zealand

The distribution of sample $\hat{d}'$s, although mathematically intractable, can be tabulated readily by computer. Such tabulations reveal a number of interesting properties of this distribution, including: (1) sample $\hat{d}'$s are biased, with an expected value that can be higher or lower than the true value, depending on the sample size, the true value itself, and the convention adopted for handling cases in which the sample $\hat{d}'$ is undefined; (2) the variance of $\hat{d}'$ also depends on the convention adopted for handling cases in which the sample $\hat{d}'$ is undefined and is in some cases poorly approximated by the standard approximation formula, (3) the standard formula for a confidence interval for $\hat{d}'$ is quite accurate with at least 50–100 trials per condition, but more accurate intervals can be obtained by direct computation with smaller samples.

The theory of signal detection (TSD; e.g., Green & Swets, 1966) provides a measure of discriminative sensitivity, d', that has been used to study a wide range of discriminative abilities, including those underlying sensation and perception, recognition memory, and social comparison. Unfortunately, values of d' are theoretical quantities that cannot be measured directly. Researchers must approximate them with estimates, $\hat{d}'$, computed from observed discrimination responses.[1] Because of the randomness inherent in discrimination responses, any estimated value, $\hat{d}'$, is likely to be somewhat different from the true value, d'.

The statistical properties of the random variable, $\hat{d}'$, are not well understood for two reasons. First, as is elaborated in the next section, $\hat{d}'$ is ill defined; that is, there is always some non-zero probability that an experiment will result in data for which $\hat{d}'$ cannot be computed. Second, even in the cases for which it is defined, the sampling distribution of $\hat{d}'$ is mathematically intractable, and there are no simple equations for its mean, variance, and so on. As a result, the properties of $\hat{d}'$ have generally been examined either by developing approximation formulas (e.g., Gourevitch & Galanter, 1967) or by Monte-Carlo simulation (e.g., Hautus, 1995).

This article shows how the statistical properties of $\hat{d}'$ can be ascertained by direct computation and presents some representative results concerning the sampling distribution of $\hat{d}'$ in the yes/no task. First, I describe how the exact sampling distribution of $\hat{d}'$ can be tabulated by computer after one adopts some convention to deal with the problematic cases in which $\hat{d}'$ is undefined. Second, the mean and variance of $\hat{d}'$ are computed directly from such tabulations, and the results indicate that they depend in a complex fashion on the true value of d', the number of trials in the experiment, and the convention adopted for handling problematic cases. The form of this dependence has several implications for the design of experiments in which $\hat{d}'$ will be measured. Third, the standard procedures for computing approximate confidence intervals for d' are evaluated numerically.

THE SAMPLING DISTRIBUTION OF $\hat{d}'$

In the yes/no task, $\hat{d}'$ is computed by using an inverse-normal transformation of two observed response probabilities, $\hat{H} = N_h/N_s$ and $\hat{F} = N_{fa}/N_n$, where N_s, N_n, N_h, and N_{fa} are the numbers of signal trials, noise trials, hits, and false alarms, respectively:

$$\hat{d}' = z(\hat{H}) - z(\hat{F}), \tag{1}$$

where $z(p)$ is the value of the standard normal distribution having cumulative probability p.

The only random variables entering into this formula are the observed response counts, N_h and N_{fa}, each of which follows a binomial distribution, given the standard assumption of independent trials. In an experiment with N_s signal trials, for example, the distribution of N_h is

$$\Pr(N_h = k) = \binom{N_s}{k} p_h^k (1 - p_h)^{N_s - k}, \quad k = 0, 1, \ldots, N_s, \tag{2}$$

where p_h is the true probability of a hit on a signal trial. Because $z(\hat{H})$ is obtained by a direct transformation of N_h, it follows that the distribution of $z(\hat{H})$ is simply

$$\Pr\left[z(\hat{H}) = z\left(\frac{k}{N_s}\right)\right] = \binom{N_s}{k} p_h^k (1 - p_h)^{N_s - k}, \quad k = 0, 1, \ldots, N_s. \tag{3}$$

The distribution of $z(\hat{F})$ is related analogously to the binomial distribution of N_{fa}, which depends on the true

This work was supported by an Otago Research Grant. I thank Stanley Klein, Neil Macmillan, Rolf Ulrich, and Patricia Haden for helpful comments on earlier versions of the article. Special thanks go to Howard Kaplan not only for providing useful suggestions but also for catching some computational errors. Correspondence concerning this article should be addressed to J. Miller, Department of Psychology, University of Otago, Dunedin, New Zealand (e-mail: miller@otago.ac.nz).

probability of a false alarm p_{fa}. Furthermore, the values of p_h and p_{fa} can be computed directly from assumptions about the underlying signal detection model. Assuming equal variance signal and noise distributions and an unbiased response criterion, for example, these probabilities can be determined directly from tables of the cumulative normal (z) distribution:

$$p_h = \Pr(z < d'/2), \tag{4}$$

$$p_{fa} = 1 - \Pr(z < d'/2). \tag{5}$$

Thus, the exact probability distributions of $z(\hat{H})$ and $z(\hat{F})$ can be computed by using Equation 3 and its false alarm counterpart, under any set of assumptions about the underlying signal and noise distributions and criterion location.

Assuming independence of N_h and N_{fa}, the sampling distribution of $\hat{d}'$ is the convolution of the distributions of $z(\hat{H})$ and $z(\hat{F})$. That is, the possible values of $\hat{d}'$ can be enumerated by subtracting each of the possible values of $z(\hat{F})$ from each of the possible values of $z(\hat{H})$, and the probability that $\hat{d}'$ is equal to a given difference is the product of the probabilities associated with the individual $z(\hat{H})$ and $z(\hat{F})$ values yielding that difference.[2] Thus, the distribution of $\hat{d}'$ is discrete and can be tabulated. Figure 1 shows two examples of these sampling distributions, computed with $d' = 0.5$ (upper panel) and $d' = 2.5$ (lower panel), assuming samples of eight signal trials and eight noise trials and using the 0.0001 convention discussed next. It is clear that the sampling distribution is not well approximated by the normal distribution with such small sample sizes.

A conceptual difficulty with the preceding characterization of the distribution of $\hat{d}'$ is that there is some probability that $\hat{d}'$ will be undefined (cf. Hautus, 1995). Specifically, this happens when the subject makes the same response in all trials within a condition, resulting in 100% hits or 100% misses on signal trials or in 100% false alarms or 100% correct rejections on noise trials. In these problematic cases, application of Equation 1 would require computation of the z scores corresponding to cumulative normal probabilities of 0 or 1, which are undefined. To examine the statistical properties of $\hat{d}'$, then, it is necessary to adopt some convention for dealing with these undefined cases. Intuitively, it seems clear that the convention will not be very important when there are many signal and noise trials, because the subject will then be very unlikely to give the same response on all of them. As will be seen, direct computation is helpful in determining how many trials are needed before the difficulty can be ignored.

In computations reported in this article, three different conventions for handling problematic cases were exam-

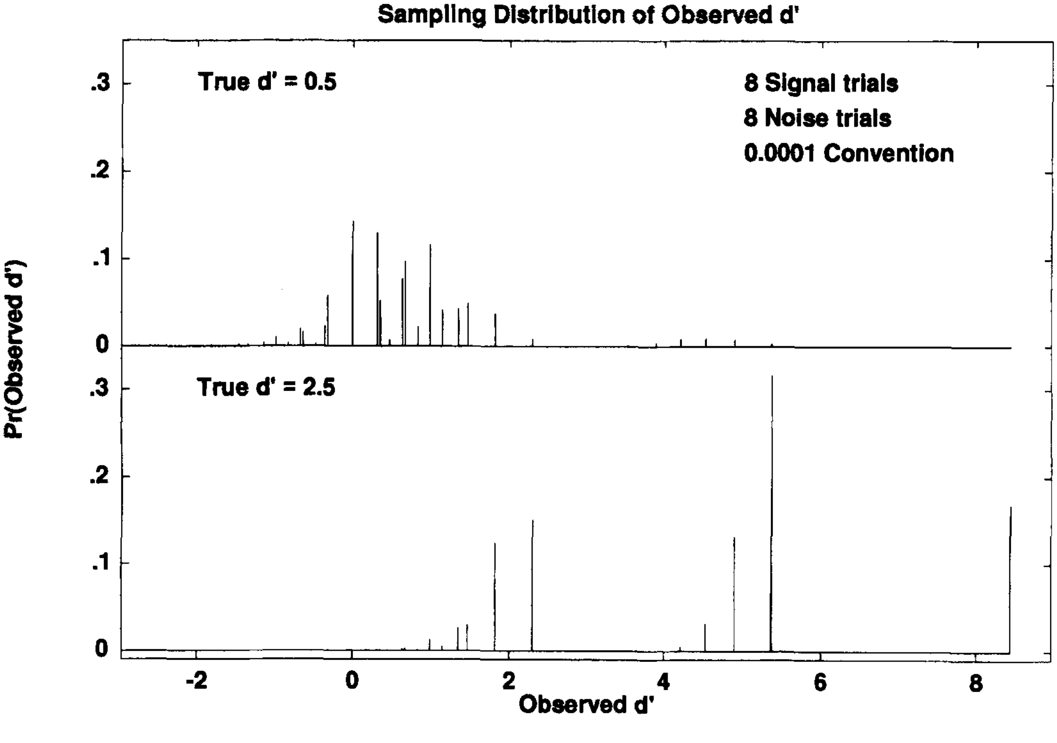

Figure 1. Sampling distributions of $\hat{d}'$ computed with samples of eight signal trials and eight noise trials. Each bar represents one possible discrete value of $\hat{d}'$, and the height of the bar represents the probability of that value under the assumption that the true d' equals 0.5 (upper panel) or 2.5 (lower panel). Values of $\hat{d}'$ less than −2 are also possible, but they are not shown because their probabilities are all less than 0.001. The 0.0001 convention was used to correct observations of zero or eight hits or false alarms (see text).

ined. The first was to replace observed values of 0 or N_s hits with 0.5 or $N_s - 0.5$, respectively, as recommended by Murdock and Ogilvie (1968), and to make the analogous correction with observed values of 0 or N_n false alarms. Unfortunately, the number 0.5 is an arbitrary constant in this procedure. To examine the effect of this arbitrary choice, the second convention was to replace observed values of 0, N_s, or N_n with 0.0001, $N_s - 0.0001$, or $N_n - 0.0001$, respectively, producing much more extreme z scores than those obtained using 0.5. The third convention was to eliminate the problematic cases rather than adjusting them, as would correspond to the experimental practice of rerunning any condition in which problematic results were obtained. In statistical terms, this involved computing the conditional distributions of $z(\hat{H})$ and $z(\hat{F})$, conditional on observing numbers of hits and false alarms between 1 and $N - 1$. Computation of the conditional distribution by enumeration is quite easy to do: The cases with 0 and N hits or false alarms are simply omitted, and the probabilities of the remaining cases are normalized to sum to one.

THE MEAN OF $\hat{d}'$

The mean or "expected value" of $\hat{d}'$, E[$\hat{d}'$], can be obtained from E[$z(\hat{H})$] and E[$z(\hat{F})$], without constructing the full distribution of $\hat{d}'$. Specifically, the mean is E[$\hat{d}'$] = E[$z(\hat{H})$] − E[$z(\hat{F})$] in the yes/no task. The expected values of $z(\hat{H})$ and $z(\hat{F})$ can be computed numerically, given a sample size N and a hit or false-alarm probability p, using

$$E[z(\hat{p})] = \sum_{k=0}^{N} z\left(\frac{k}{N}\right)\binom{N}{k}p^k(1-p)^{N-k}, \quad k = 0,1,\ldots,N. \tag{6}$$

Table 1 shows the mean of yes/no task $\hat{d}'$ values as a function of the true d' in the experiment, sample size ($N_s = N_n$), and convention for handling problematic cases (i.e., $k = 0$ and $k = N$), computed assuming unbiased responding (i.e., $p_h = 1 - p_{fa}$). Three important points are apparent from these values.

First, none of the conventions works very well with small samples. For reasons discussed in the next two paragraphs, each yields biased estimates for virtually all of the different possible values of d'. That is, for any convention, the average of many small-sample $\hat{d}'$s will generally be a biased estimate of the true d'. It is noteworthy that the convention of excluding problematic cases introduces bias just like the other two conventions, although it may seem intuitively less arbitrary.

Second, the arbitrary choice of a convention for handling problematic cases makes a big difference with small samples. The different conventions yield different means, because at each sample size the maximum attainable $\hat{d}'$ depends on the convention. With eight trials per condition, for example, the maximum attainable $\hat{d}'$s are 3.1, 8.4, and 2.3 for the 0.5 adjustment, 0.0001 adjustment, and conditional distributions, respectively. In

Table 1
Mean of $\hat{d}'$ in the Yes/No Task as a Function of d', Number of Trials, and Convention for Handling Problematic Cases

	True	Number of Signal and Noise Trials							
Convention	d'	8	16	32	64	128	256	512	1,024
0.5	0.5	0.555	0.530	0.514	0.507	0.503	0.502	0.501	0.500
	1.0	1.096	1.064	1.029	1.014	1.007	1.003	1.002	1.001
	1.5	1.599	1.604	1.551	1.524	1.512	1.506	1.503	1.501
	2.0	2.036	2.133	2.084	2.039	2.019	2.009	2.004	2.002
	2.5	2.385	2.615	2.622	2.563	2.529	2.514	2.507	2.503
	3.0	2.641	3.008	3.135	3.101	3.048	3.023	3.011	3.006
	4.0	2.926	3.481	3.879	4.091	4.122	4.069	4.032	4.015
	5.0	3.030	3.660	4.191	4.624	4.936	5.096	5.106	5.057
0.0001	0.5	0.640	0.531	0.514	0.507	0.503	0.502	0.501	0.500
	1.0	1.376	1.078	1.030	1.014	1.007	1.003	1.002	1.001
	1.5	2.286	1.686	1.553	1.524	1.512	1.506	1.503	1.501
	2.0	3.383	2.449	2.102	2.039	2.019	2.009	2.004	2.002
	2.5	4.580	3.454	2.755	2.567	2.529	2.514	2.507	2.503
	3.0	5.725	4.666	3.653	3.155	3.049	3.023	3.011	3.006
	4.0	7.386	6.950	6.142	5.121	4.347	4.081	4.032	4.015
	5.0	8.131	8.196	8.064	7.637	6.867	5.930	5.268	5.063
Conditional	0.5	0.515	0.529	0.514	0.507	0.503	0.502	0.501	0.500
	1.0	0.988	1.057	1.029	1.014	1.007	1.003	1.002	1.001
	1.5	1.384	1.569	1.551	1.524	1.512	1.506	1.503	1.501
	2.0	1.691	2.026	2.075	2.039	2.019	2.009	2.004	2.002
	2.5	1.912	2.391	2.574	2.561	2.529	2.514	2.507	2.503
	3.0	2.063	2.653	2.991	3.080	3.048	3.023	3.011	3.006
	4.0	2.223	2.933	3.484	3.870	4.056	4.065	4.032	4.015
	5.0	2.280	3.032	3.661	4.192	4.621	4.923	5.059	5.054

Note—Convention 0.5 is to replace 0 and N with 0.5 and $N - 0.5$, respectively, and convention 0.0001 is to replace them with 0.0001 and $N - 0.0001$. The conditional convention is to exclude observations of 0 and N.

any case, since the choice of convention is arbitrary, investigators wishing to estimate d' values must endeavor to collect enough trials so that the arbitrary choice will have no effect. As is apparent from Table 1, the required sample size depends somewhat on the true value of d'; samples of 50–100 trials are probably large enough for $d' \leq 3$, but samples of up to 1,000 are needed as d' values approach five.

Third, regardless of the convention for handling problematic cases, the mean of $\hat{d}'$ depends on the sample size. Even with constant values of d' and the response criterion (and hence constant values of p_{h} and p_{fa}), the results indicate that mean $\hat{d}'$ can vary dramatically and nonmonotonically across the smaller sample sizes, although it does eventually converge to the true value when sample size is large enough. Nonmonotonic effects are strongest for the adjustment of 0.5 and the conditional distribution, and they increase with the true d'. These nonmonotonicities result from the interplay of two counteracting nonlinear effects: (1) the maximum attainable $\hat{d}'$ increases with sample size (e.g., values of 3.1, 3.7, and 4.3 for sample sizes of 8, 16, and 32 tiials, respectively, for the 0.5 adjustment); and (2) the probability of this maximum $\hat{d}'$ decreases with sample size. In any case, the dependence of mean $\hat{d}'$ on sample size clearly means that experimenters wishing to compare values of d' across conditions must either obtain equal numbers of observations in all conditions or else ensure that both conditions have enough observations for the mean of $\hat{d}'$ to be quite close to its asymptote.

Simulations indicate that the differential bias as a function of sample size is large enough to present a potential confound in comparisons between conditions having different numbers of observations, especially at more extreme values of d'. As an illustration, I simulated 5,000 yes/no experiments, each having 40 subjects divided equally among conditions with $N_{\text{s}} = N_{\text{n}} = 20$ and with $N_{\text{s}} = N_{\text{n}} = 80$, and the 0.5 adjustment was used for problematic samples. Simulated sensitivity was identical for all subjects, with $d' = 2.56$ (i.e., 90% correct). Nonetheless, between-subjects t tests with a .05 significance level indicated that mean $\hat{d}'$ was significantly larger in the condition with the smaller number of observations in 8.8% of the simulated experiments. This is far greater than the 2.5% rate of Type I errors in this direction ordinarily associated with such a test. Simulations with other levels of sensitivity and other sample sizes in a 2:1 or 4:1 ratio often gave Type I error rates that were similarly discrepant from the theoretical values.

The problem of differential bias as a function of sample size would likely be worse in multisubject than in single-subject designs, because researchers tend to compute values of $\hat{d}'$ from smaller Ns in between-subjects designs, and bias is large when sample size is small. This problem is not routinely addressed, however. For example, Reinitz (1990, Experiment 2) compared yes/no detection performance following valid versus invalid spatial cues. He computed values of $\hat{d}'$ from 64 and 32 trials per subject in the two cuing conditions, but did not discuss the fact that differential sample-size biases would be expected to modulate the effect of cuing on $\hat{d}'$; his experimental effects seem too large to be explainable by such bias, however.

Unfortunately, the biases in Table 1 do not immediately reveal how best to estimate the true d' from an observed value of $\hat{d}'$ obtained in an experiment. For example, mean $\hat{d}'$s observed with $N = 64$ tend to be a little larger than true d's, so it seems reasonable to estimate that a true value is slightly less than an observed one with this sample size. Exactly how much less is not clear, however, and further work is needed to identify the optimal estimator(s) of d'. In the meantime, investigators using sample $\hat{d}'$s as estimates of true d's should at least be aware of the bias inherent in their estimates.

THE VARIANCE OF $\hat{d}'$

When an experimental design yields values of $\hat{d}'$ for a number of subjects, and statistical analyses are conducted across subjects (e.g., t tests across subjects, using one or more values of $\hat{d}'$ from each subject), the random error of any individual $\hat{d}'$ can usually be ignored, because it is subsumed under intersubject variability. In single-subject designs, however, it is often desirable to know how much random error is associated with a given $\hat{d}'$ (cf. Macmillan & Creelman, 1991, chapter 11). In particular, when fitting a model to the data of an individual subject, one needs to know the predicted distribution of $\hat{d}'$, or at least its variance, in order to decide whether a given observed value of $\hat{d}'$ is discrepant enough that the model should be rejected.

The variance of $\hat{d}'$ can also be obtained without constructing its full distribution, given the standard assumption that $z(\hat{H})$ and $z(\hat{F})$ are independent. The variance, Var$[\hat{d}']$, is Var$[z(\hat{H})]$ + Var$[z(\hat{F})]$ in the yes/no task. To compute this variance, it is useful to obtain the second raw moment:

$$\mathrm{E}[z(\hat{p})^2] = \sum_{k=0}^{N}\left[z\left(\frac{k}{N}\right)\right]^2\binom{N}{k}p^k(1-p)^{N-k}, \quad k = 0,1,\ldots,N. \tag{7}$$

Then the variance of $z(\hat{p})$ is simply

$$\mathrm{Var}\,[z(\hat{p})] = \mathrm{E}\,[z(\hat{p})^2] - \mathrm{E}\,[z(\hat{p})]^2. \tag{8}$$

Table 2 shows the variance of $\hat{d}'$ in the yes/no task as a function of the true d', sample size ($N_{\text{s}} = N_{\text{n}}$), and convention for handling problematic cases, again computed assuming unbiased responding. Like the mean, the variance is heavily dependent on the convention when sample size is small. The adjustment of 0.0001 yields the largest variances because it produces the most extreme values of $\hat{d}'$, and the conditional distribution yields the smallest variances because it discards the most extreme values. To avoid an influence of the arbitrary choice on predicted variances, experimenters wishing to fit models to single-subject data should include at least 60 signal and noise trials per subject when expecting d's of 0–2, at

Table 2
Variance of $\hat{d}'$ in the Yes/No Task as a Function of d', Sample Size, and Method of Computation

Method of	True	Number of Signal and Noise Trials							
Computation	d'	8	16	32	64	128	256	512	1,024
0.5	0.5	0.4913	0.2289	0.1065	0.0516	0.0255	0.0126	0.00630	0.00314
	1.0	0.4823	0.2549	0.1167	0.0558	0.0274	0.0136	0.00675	0.00337
	1.5	0.4395	0.2908	0.1375	0.0639	0.0310	0.0153	0.00760	0.00379
	2.0	0.3583	0.3098	0.1733	0.0785	0.0372	0.0182	0.00900	0.00448
	2.5	0.2590	0.2836	0.2144	0.1051	0.0477	0.0229	0.01125	0.00558
	3.0	0.1682	0.2169	0.2245	0.1471	0.0671	0.0309	0.01496	0.00737
	4.0	0.0560	0.0820	0.1213	0.1567	0.1409	0.0758	0.03334	0.01568
	5.0	0.0149	0.0219	0.0346	0.0561	0.0873	0.1136	0.09804	0.05012
0.0001	0.5	0.9702	0.2367	0.1065	0.0516	0.0255	0.0126	0.00630	0.00314
	1.0	1.7498	0.3256	0.1169	0.0558	0.0274	0.0136	0.00675	0.00337
	1.5	3.0514	0.6672	0.1439	0.0639	0.0310	0.0153	0.00760	0.00379
	2.0	4.4496	1.5545	0.2594	0.0788	0.0372	0.0182	0.00900	0.00448
	2.5	5.2325	2.9679	0.7429	0.1211	0.0478	0.0229	0.01125	0.00558
	3.0	4.9965	4.1871	1.9200	0.3595	0.0698	0.0309	0.01496	0.00737
	4.0	2.7000	3.6052	3.8818	2.7035	0.8683	0.1184	0.03347	0.01568
	5.0	0.8750	1.3986	2.1417	2.9179	3.1027	2.0427	0.58394	0.07239
Conditional	0.5	0.4416	0.2275	0.1065	0.0516	0.0255	0.0126	0.00630	0.00314
	1.0	0.3950	0.2459	0.1167	0.0558	0.0274	0.0136	0.00675	0.00337
	1.5	0.3224	0.2576	0.1365	0.0639	0.0310	0.0153	0.00760	0.00379
	2.0	0.2401	0.2396	0.1641	0.0784	0.0372	0.0182	0.00900	0.00448
	2.5	0.1648	0.1915	0.1784	0.1032	0.0477	0.0229	0.01125	0.00558
	3.0	0.1055	0.1339	0.1573	0.1304	0.0667	0.0309	0.01496	0.00737
	4.0	0.0361	0.0492	0.0703	0.0965	0.1067	0.0720	0.03333	0.01568
	5.0	0.0098	0.0137	0.0204	0.0317	0.0489	0.0695	0.07568	0.04819
G & G	0.5	0.4017	0.2009	0.1004	0.0502	0.0251	0.0126	0.00628	0.00314
	1.0	0.4303	0.2151	0.1076	0.0538	0.0269	0.0134	0.00672	0.00336
	1.5	0.4832	0.2416	0.1208	0.0604	0.0302	0.0151	0.00755	0.00377
	2.0	0.5700	0.2850	0.1425	0.0712	0.0356	0.0178	0.00891	0.00445
	2.5	0.7081	0.3540	0.1770	0.0885	0.0443	0.0221	0.01106	0.00553
	3.0	0.9291	0.4646	0.2323	0.1161	0.0581	0.0290	0.01452	0.00726
	4.0	1.9067	0.9534	0.4767	0.2383	0.1192	0.0596	0.02979	0.01490
	5.0	5.0214	2.5107	1.2553	0.6277	0.3138	0.1569	0.07846	0.03923

Note—The first three methods of computation are the three conventions for handling problematic cases, as in Table 1. Method "G & G" is the approximation of Gourevitch and Galanter (1967).

least 100–200 trials for d's of 2–3, and correspondingly more trials for larger d's.[3]

Inspection of Table 2 reveals that the variance of $\hat{d}'$ depends on true d' and sample size in a complex fashion for both the 0.5 and 0.0001 conventions. With the 0.0001 convention and $N_s = N_n = 16$, for example, variance starts low at $d' = 0.5$, increases to a maximum at $d' \approx 3.0$, and then decreases again for larger d's. The nonmonotonic dependence on d' is also evident at $N_s = N_n = 8$ and 32, but for these two sample sizes the maximal variance occurs at $d' \approx 2.5$ and 4.0, respectively. To develop some intuition for this pattern, it is helpful to consider how the distribution of $\hat{d}'$ would change as d' changes. Figure 1 shows the distributions for $d' = 0.5$ and $d' = 2.5$, in the case where $N_s = N_n = 8$ and the 0.0001 convention is used. Note that for the larger d', the distribution of $\hat{d}'$ shifts to the right, because larger $\hat{d}'$ values become more probable. The larger d' also produces a distribution with more variance, because the numerical values of $\hat{d}'$ spread out as they depart more from zero (because of the stretching inherent in the inverse normal probability transformation). Although not shown in this figure, it is clear that as d' gets very large, the distribution must eventually pile up at its maximal value of 8.43, because the probability of perfect performance tends toward 1.0 as d' increases. As this happens, the variance of $\hat{d}'$ naturally decreases toward an asymptote of zero, because the mass of the distribution becomes concentrated in the maximal value.

Variances obtained by direct computation can be compared with those obtained from an approximation formula given by Gourevitch and Galanter (1967). Gourevitch and Galanter noted that

$$\mathrm{Var}\left[z\left(\frac{K}{N}\right)\right] \approx \frac{p(1-p)}{N\phi(p)^2}, \tag{9}$$

where $\phi(p)$ is the height of the normal density at $z(p)$. The variance of $\hat{d}'$, then, is approximately

$$\mathrm{Var}[\hat{d}'] \approx \frac{\hat{p}_h(1-\hat{p}_h)}{N_s\phi(\hat{p}_h)^2} + \frac{\hat{p}_{fa}(1-\hat{p}_{fa})}{N_n\phi(\hat{p}_{fa})^2}. \tag{10}$$

Table 2 also shows variances estimated with Gourevitch and Galanter's (1967) approximation (Equation 10). The approximation is excellent for $N > 100$ and

$d' \leq 1.0$, but may not be accurate enough for precise quantitative work with larger values of d'. Even at sample sizes large enough so that the convention for handling problematic cases does not influence the predicted variance, the approximation can underestimate the true variance by 10%.

It is also worth noting that straightforward use of Gourevitch and Galanter's (1967) approximation can also lead to overestimations of the true variance. For example, Bonnel and Miller (1994) chose parameter values to minimize the error score

$$\Delta = \sum_{i=1}^{I} \frac{(\hat{d}_i' - d_i')^2}{Var[\hat{d}_i']} \qquad (11)$$

across I conditions. Using Gourevitch and Galanter's approximation formula for Var$[\hat{d}']$, the best fit was often obtained with parameters that gave extremely large values of the $\hat{d}_i'$s, because such parameters also yield large predicted variances for the denominator of the overall error measure, Δ. For example, in a condition with $N_s = N_n = 200$ and $\hat{d}' = 2.5$, a bias-free model makes a smaller contribution to Δ if it predicts $d' = 7.8$ than if it predicts $d' = 2.8$. Even though $d' = 7.8$ is much more discrepant from the observed value than is $d' = 2.8$, the ratio of squared error to predicted variance is smaller in the former case because of the large predicted variance associated with $d' = 7.8$. This property of the approximation formula can easily cause automatic parameter-search programs to settle on parameter estimates yielding unreasonably large predicted d's. Fortunately, direct computation of variances allows one to avoid this problem, because for any convention the predicted variance decreases for values of d' larger than some fixed constant.[4] Naturally, it would be wise to fit a model by using several different conventions to show that the arbitrarily selected convention did not have a large influence on the fit.

CONFIDENCE INTERVALS FOR d'

The standard method for computing a 95% confidence interval for d' in the yes/no task is to use Gourevitch and Galanter's (1967) variance approximation formula (cf. Macmillan & Creelman, 1991) together with the assumption that the sampling distribution of $\hat{d}'$ is approximately normal. With direct computation, it is straightforward to check the accuracy of such confidence intervals.

According to the definition of a confidence interval (see, e.g., Kendall & Stuart, 1961), if, for example, 1.23 is the value of $\hat{d}'$ observed in a sample, then the lower and upper bounds, L and U, of a 95% confidence interval should be chosen so that

$$2.5\% = \Pr(\hat{d}' > 1.23 \mid d' = L) = \Pr(\hat{d}' < 1.23 \mid d' = U). \qquad (12)$$

That is, the observed value 1.23 should lie at the 2.5th percentile point in the bottom tail of the sampling distribution generated when $d' = U$, and it should lie at the 97.5th percentile point of the sampling distribution when $d' = L$. Intuitively, this means that 1.23 is almost discrepant enough from the lower (or upper) bound that we can reject the null hypothesis that the true value is equal to this bound, at the .05 significance level with a two-tailed test.

The appropriate values of L and U can be obtained by numerical search (see, e.g., Press, Flannery, Teukolsky, & Vetterling, 1986), although the procedure is somewhat computationally intensive. To find U, for example, one tries a series of candidate values of d'. For each candidate d' value, the associated values of p_{h} and p_{fa} are determined by using the cumulative normal distribution (cf. Equations 4 and 5). Then, the predicted sampling distributions of $z(\hat{H})$ and $z(\hat{F})$ are obtained from the underlying binomials N_{h} and N_{fa}, and the predicted sampling distribution of $\hat{d}'$ is tabulated by convoluting the predicted distributions of $z(\hat{H})$ and $z(\hat{F})$—that is, by generating the differences obtained with all possible combinations of the component $z(\hat{H})$ and $z(\hat{F})$ values (cf. Equation 3). Finally, the predicted sampling distribution is examined to see whether its 2.5th percentile is equal to the observed $\hat{d}'$ value, as desired. The search continues until a candidate d' is found for which the predicted 2.5th percentile equals the observed sample $\hat{d}'$, and this candidate is then taken as the value of U. The lower bound of the confidence interval is found similarly, except that the search stops when 97.5th percentile of the predicted distribution is equal to the observed $\hat{d}'$. As always, computing the sampling distribution of $\hat{d}'$ for each candidate d' requires adoption of some convention for handling problematic cases. It is therefore prudent to determine confidence intervals by using different conventions and to make sure that the results do not depend on the convention.

Table 3 shows a number of examples of confidence interval upper bounds computed with direct computation and with the approximation of Gourevitch and Galanter (1967); similar results were obtained in a comparison of lower bounds computed with the different methods. The approximation is quite accurate when $\hat{d}' \leq 2.5$ and there are more than 50–100 trials per condition. Examination of cases in which the approximation is not accurate suggests that the inaccuracies are due to a combination of factors, including the following: (1) The approximated variance differs from the true variance; (2) the sampling distribution of $\hat{d}'$ deviates somewhat from the normality assumed in the standard procedure for computing confidence intervals; and (3) in the standard procedure, there is only one variance, estimated by using $\hat{d}'$, whereas the true predicted variance differs somewhat for d's at the top and bottom of the confidence interval.

It is also interesting to note the virtual identity, with $\hat{d}' \leq 2.5$ and 16 or more trials per condition, of confidence bounds computed by using adjustments of 0.5 versus 0.0001 to handle problematic cases. Under these conditions, the problematic cases are rare enough that their adjustment influences only the outer 5% of the

Table 3
Upper Bounds of 95% Confidence Intervals for d' Obtained With Direct Computation or the Approximation of Gourevitch and Galanter (1967)

Method of Computation	Sample $\hat{d}'$	Number of Signal and Noise Trials 8	16	32	64	128	256	512	1,024
0.5	0.5	1.701	1.414	1.130	0.933	0.815	0.719	0.657	0.611
	1.0	2.377	1.943	1.652	1.451	1.320	1.229	1.162	1.114
	1.5	2.892	2.398	2.176	1.955	1.831	1.736	1.669	1.620
	2.0	3.490	2.889	2.688	2.504	2.366	2.256	2.184	2.130
	2.5		3.597	3.317	3.078	2.891	2.782	2.702	2.644
	3.0		4.117	3.819	3.665	3.446	3.323	3.233	3.165
	4.0				4.767	4.641	4.473	4.330	4.237
	5.0						5.695	5.537	5.379
0.0001	0.5	1.701	1.414	1.130	0.933	0.815	0.719	0.657	0.611
	1.0	2.376	1.943	1.652	1.451	1.320	1.229	1.162	1.114
	1.5	2.890	2.398	2.176	1.955	1.831	1.736	1.669	1.620
	2.0	3.418	2.889	2.688	2.504	2.366	2.256	2.184	2.130
	2.5		3.594	3.317	3.078	2.891	2.782	2.702	2.644
	3.0		4.042	3.819	3.665	3.446	3.323	3.233	3.165
	4.0				4.741	4.641	4.473	4.330	4.237
	5.0						5.692	5.537	5.379
Conditional	0.5	1.820	1.419	1.130	0.933	0.815	0.719	0.657	0.611
	1.0	2.755	1.968	1.653	1.451	1.320	1.229	1.162	1.114
	1.5	3.700	2.473	2.179	1.955	1.831	1.736	1.669	1.620
	2.0	5.375	3.098	2.706	2.504	2.366	2.256	2.184	2.130
	2.5		4.348	3.425	3.083	2.891	2.782	2.702	2.644
	3.0		5.865	4.150	3.707	3.447	3.323	3.233	3.165
	4.0				5.392	4.778	4.482	4.330	4.237
	5.0						6.266	5.627	5.384
G & G	0.5	1.742	1.378	1.120	0.938	0.810	0.719	0.654	0.609
	1.0	2.285	1.908	1.642	1.454	1.321	1.226	1.160	1.113
	1.5	2.862	2.463	2.181	1.981	1.840	1.740	1.670	1.620
	2.0	3.480	3.046	2.740	2.523	2.370	2.262	2.185	2.131
	2.5		3.667	3.325	3.084	2.913	2.792	2.707	2.646
	3.0		4.337	3.946	3.669	3.473	3.335	3.237	3.168
	4.0				4.959	4.678	4.480	4.340	4.240
	5.0						5.778	5.551	5.390

Note—Because N_h and N_{fa} must be integers, it is not possible to observe all listed values of $\hat{d}'$ for each number of trials. The upper bounds were nonetheless determined by finding the value of d' satisfying Equation 12 for each of the possible observed values listed in the column headed "Sample $\hat{d}'$." Empty cells indicate cases in which the indicated sample $\hat{d}'$ was larger than could be observed without perfect performance on signal or noise trials, making computation of an upper bound on d' unreasonably sensitive to the convention for handling perfect performance.

sampling distribution. This means that it is possible to compute convention-independent 95% confidence intervals with sample sizes too small for convention-independent variances (cf. Table 2), because variances are influenced by the full distribution whereas confidence intervals are influenced only by the middle 95%.

CONCLUSION

The method of direct computation can be used to investigate the distribution of $\hat{d}'$ in virtually any signal detection paradigm, because all paradigms yield values of $\hat{d}'$ computed from the observed values of underlying binomial or multinomial random variables (cf. Macmillan & Creelman, 1991). This method, while computationally intensive, is well within the capabilities of available desk-top computers. The method can be used to investigate the statistical properties of $\hat{d}'$ despite the mathematical intractability of this measure, and thereby to improve the accuracy of quantitative model tests and confidence-interval computation performed with data from individual subjects.

REFERENCES

Bonnel, A.-M., & Miller, J. (1994). Attentional effects on concurrent psychophysical discriminations: Investigations of a sample-size model. *Perception & Psychophysics*, **55**, 162-179.

Gourevitch, V., & Galanter, E. (1967). A significance test for one parameter isosensitivity functions. *Psychometrika*, **32**, 25-33.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, **27**, 46-51.

Kendall, M. G., & Stuart, A. (1961). *The advanced theory of statistics* (Vol. II). London: Griffin.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

Murdock, B. B., Jr., & Ogilvie, J. C. (1968). Binomial variability in short-term memory. *Psychological Bulletin*, **70**, 256-260.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes: The art of scientific computing*. Cambridge: Cambridge University Press.

Reinitz, M. T. (1990). Effects of spatially directed attention on visual encoding. *Perception & Psychophysics*, **47**, 497-505.

NOTES

[1] Throughout this article, symbols with "hats" denote estimates, computed from data, of the corresponding theoretical parameter values symbolized without hats (e.g., $\hat{d}'$ is an estimate of d').

[2] I ignore the fact that the same value of $\hat{d}'$ could arise from different combinations of $z(\hat{H})$ and $z(\hat{F})$. This presents no computational problems, because such equivalent combinations can be identified and their probabilities combined in the procedure described below.

[3] In principle, a researcher could decide in advance how to handle problematic cases and obtain predicted means and variances by direct computation, using that specific convention. This seems unsatisfactory, however, because it could lead to a scenario in which a model would be accepted or rejected depending on an experimenter's arbitrary choice of how to handle problematic data that never occurred in the actual experiment.

[4] As d' increases, the true probability of a hit approaches one and the true probability of a false alarm approaches zero. In the limit, then, performance is always perfect, so the variance of $\hat{d}'$ approaches an asymptote of zero.

(Manuscript received April 4, 1994;
revision accepted for publication May 20, 1995.)