Bayesian Probability and Scalar Performance Measures in Gaussian Models

Caren Marzban

National Severe Storms Laboratory, Norman, OK 73069

Cooperative Institute for Mesoscale and Meteorological Studies,
University of Oklahoma, Norman, OK 73019

Department of Physics, University of Oklahoma, Norman, OK 73019

Abstract

The transformation of a real, continuous variable into an event probability is reviewed from the Bayesian point of view, after which a Gaussian model is employed to derive an explicit expression for the probability. In turn, several scalar (one-dimensional) measures of performance quality, and reliability diagrams are computed. It is shown that if the optimization of scalar measures is of concern, then prior probabilities must be treated carefully, whereas no special care is required for reliability diagrams. Specifically, since a scalar measure gauges only one component of performance quality - a multidimensional entity - it is possible to find the critical value of prior probability that optimizes that scalar measure; this value of "prior probability" is often not equal to the "true" value as estimated from group sample sizes. Optimum reliability, however, is obtained when prior probability is equal to the estimate based on group sample sizes. Exact results are presented for the critical value of "prior probability" that optimize the Fraction Correct, the True Skill Statistic, and the Reliability Diagram, but the Critical Success Index and the Heidke Skill Statistic are treated only graphically. Finally, an example based on surface air pressure data is employed to illustrate the results in regards to precipitation forecasting.

1 Introduction

Frequently, one is faced with the task of "transforming" a variable (e.g. surface air pressure, gate-to-gate velocity difference. etc.) into a probability for a corresponding event (e.g. precipitation, tornado, etc.). A related problem is that of the performance of the forecaster, i.e. the accuracy of the forecasts or the reliability of the generated probabilities (Murphy 1993, 1996; Murphy, Brown, and Chen 1989; Murphy and Winkler 1987, 1992; Wilks 1995).

Several subtleties arise in both problems. For instance, in forming forecast probabilities, it is important to consider the correct conditional probability, namely the probability of an event, given the observation of a variable, and not the converse (Brooks and Doswell 1995; Murphy and Winkler 1987, 1992). The relationship among the various conditional probabilities is given by Bayes' theorem (Kendall and Stuart 1969; O'Hagan 1994). Also, in assessing the performance of the forecaster, not only the correct probabilities must be assessed, but also it is important to acknowledge the multidimensionality of performance itself. For instance, it is entirely possible that one forecaster will outperform another, in terms of a specific measure of performance, but not in terms of another measure. [1]

The "worse" measures are scalar (one-dimensional) and non-probabilistic, while the "best" are multi-dimensional and probabilistic, with admixtures also possible. However, sometimes the particular aspect of performance that is of interest is unambiguously specified (by funders, for example!), in which case one may concentrate on only the corresponding scalar measure, and effectively treat performance as a one-dimensional entity. There are also times when one has no choice but to appeal to a scalar measure; for instance, in deciding the winner of a forecasting contest, enforcing the multidimensionality of performance may lead to several winners - one for each dimension of performance. Of course, it is possible

[1] In this article, "performance" refers to a measure of forecast quality (Murphy 1993).

that a forecaster outperforms all others in terms of all the components of performance, but such situations are neither guaranteed nor likely. In this article, although the framework for forming forecasts is intrinsically probabilistic, primary consideration is given to scalar, non-probabilistic measures based on a contingency table (Wilks 1995). Reliability diagrams will also be considered, but a complete treatment of multidimensional and probabilistic measures will be postponed to a later article.

One non-probabilistic, scalar measure that appears to satisfy most, of the requirements that one could place on such measures (Marzban and Stumpf 1997) is Heidke's Skill Statistic (HSS). A measure that is intimately related (Doswell et al. 1990) to HSS is the True Skill Statistic (TSS). Another popular measure in meteorology is the Critical Success Index (CSI) (Donaldson et al. 1975a,b), with its popularity withstanding its "inequitability" (Gandin and Murphy 1992) in that its values for random guessing and persistence are unequal; in fact, technically, CSI is not even a measure of skill since it does not take into account either of these two factors. Schaefer (1990) also addresses CSI in the rare-event situation. And of course, there is the most notorious of measures, namely the Fraction Correct (FRC), which in spite of its numerous inadequacies is still in common use. Its inclusion in the present study serves only as a point of contrast.

As for probabilistic and multidimensional measures, by virtue of being multidimensional, they cannot be represented by a single number, and one must appeal to multidimensional means, e.g. 2-dimensional diagrams, to represent such quantities. One example is the reliability diagram (Wilks 1995) which is the one that will be discussed here as an example of a probabilistic, multidimensional measure. Multi-category reliability diagrams have also been considered (Hamill 1997).

We begin with a review of the Bayesian approach of transforming a real, continuous

variable into the posterior probability of an event, given an observation. Then, a Gaussian model is employed to allow for an explicit computation of the probabilities and several measures of accuracy. [2] It is shown that by virtue of being scalar measures and also depending on prior probability, it is possible to "free" the prior probabilities from their "true" values, as estimated from the group sample sizes, and instead set them to critical values that maximize a given measure. These critical values of "prior probability" will be computed for the above-mentioned scalar measures. Exact results are found for FRC, TSS, and reliability diagrams, but CSI and HSS lend themselves mostly to an approximate (graphic) treatment.

2 Probabilities and Bayes' Theorem

In a probabilistic approach to forecasting, conditional probabilities play an important role (Brooks and Doswell 1995; Murphy and Winkler 1987), and so, the conditions under which one obtains the probability of a given event must be carefully specified. For instance, consider the situation where there are only two possible hypotheses (e.g., tornado and nontornado, or rain and no-rain, etc.), generically labeled as "1" and "0" for the existence of event and no-event, respectively. Then, the probability of making an observation of a quantity x (e.g., wind-speed, temperature, etc.), given the hypothesis, is a completely different quantity than the probability of a hypothesis being in effect when x is measured. The former is sometimes called likelihood, and the latter is the posterior probability of an event, given the measurement x; the two probabilities are related through Bayes' theorem (Kendal and Stuart 1969; O'Hagan 1994). It is this posterior probability which is of interest when a

[2] Given the ubiquity of Gaussian distributions in meteorological data, these findings are valid in a wide range of applications. A Gaussian model has also been considered by Mason (1982).

forecast is made, since x is measured first and one is then interested in the probability of the hypothesis that gave rise to that value of x.

In the 2-group case, generally labeled "0" and "1" - Bayes' theorem states that

$$P_1(x) = \frac{p_1 L_1(x)}{p_0 L_0(x) + p_1 L_1(x)}, \tag{1}$$

where p_0, p_1 are the prior probabilities for the two groups, and $L_0(x), L_1(x)$ are the likelihood functions, while $P_1(x)$ is the posterior probability that the hypothesis "1" was in effect when x is measured. Of course, $p_0 + p_1 = 1$ and $P_0 + P_1 = 1$.

How does one compute these probabilities from data on x? First, one simply plots two histograms (i.e. a frequency plot) - one for the x values corresponding to the nonevents (0s), and another for the events (1s). These frequencies can be labeled as $N_0(x)$, and $N_1(x)$, respectively. The likelihoods are then defined as

$$L_i(x) = \frac{N_i(x)}{N_i},$$

where N_i $(i = 0, 1)$ are the respective sample sizes. Clearly, the sum of all the observations in each histogram is equal to the total number of non-events (N_0), and events (N_1), respectively, in the sample. Consequently, the sum of the observations under the $L_0(x)$, and $L_1(x)$ plots is equal to 1:

$$\int L_i(x)dx = 1, \quad (i = 0, 1) \ .$$

In other words, the likelihoods are simply normalized histograms. As for the prior probabilities, their estimates are given by the sample sizes as

$$p_i = \frac{N_i}{N},$$

where $N = N_0 + N_1$. The posterior probabilities are then computed from equation (1). This completes the transformation of an observation, x, into a posterior probability of a corresponding event.

As will be shown, the value of $p_1 = N_1/N$ does not necessarily imply optimum performance when performance is gauged in terms of scalar, categorical measures. One aim of this article is to derive the critical values of "prior probability", p_{1c}, that optimize a given measure.

3 A Gaussian Model

There are many reasons (Wilks 1995) for assuming a parametric form for the Likelihoods, $L_i(x)$, and proceed to estimate the parameters from the sample data. A most common ansatz is that of normality, i.e., that the single variable x is distributed in a Normal, or Gaussian, fashion:

$$L_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} \, exp^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \quad , \tag{2}$$

where μ_i and σ_i are the mean and the standard deviation for the variable x, in group i ($i = 0, 1$), all estimated from the sample data. Figures 1a and 1b show some generic, gaussian likelihood functions for two groups.

As stated in the Introduction, sometimes one is interested in non-probabilistic measures such as CSI, TSS, or HSS, which are meaningful only for categorical forecasts. Probabilistic forecasts can always be reduced to categorical ones by the introduction of one (or more) thresholds. For example, in regards to Figure 1a, a value of x can be assigned to (or forecast as) the group with the higher likelihood. Then, the value of x at which the two curves intersect would be the natural threshold marking the boundary between the two groups. However, as discussed in the Introduction, this is the "wrong" probability to consider. A measurement x must be classified into the group with the higher *posterior* probability. Figure 1 is still instructive in that it allows for the interpretation of prior probability as a "generalized" threshold. As will be shown below, the qualifier "generalized" serves two functions:

in the special case where the two groups have equal standard deviations, prior probability simply shifts the threshold away from the one at the crossing-point of the two curves in Figure 1a to the crossing point of the curves $p_0 L_0(x)$ and $p_1 L_1(x)$. The other sense in which prior probabilities represent generalized thresholds arises in the more general case where the group standard deviations are unequal; in that case, $p_0 L_0(x)$ and $p_1 L_1(x)$ cross not at one threshold but at two thresholds, marking the boundaries between the two groups. It is then said that the decision boundary is nonlinear. The effect follows simply from the relevant equations and will be shown below.

In obtaining measures of performance for categorical forecasts, as mentioned previously, the decision criterion must be based on $P_1(x)/P_0(x)$, or equivalently $\log(P_1(x)/P_0(x))$. In a gaussian model, it follows from (1) and (2) that

$$\log(P_1(x)/P_0(x)) = \frac{1}{2} D^2(x),$$

where the so-called discriminant function, $D^2(x)$, is given by

$$D^2(x) = (\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2})x^2 - 2(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2})x + (\frac{\mu_0^2}{\sigma_0^2} - \frac{\mu_1^2}{\sigma_1^2}) + 2\log(\frac{\sigma_0}{\sigma_1}) - 2\log(\frac{1-p_1}{p_1}). \qquad (3)$$

Recall that the means and the standard deviations are estimated from the sample data, and then an observation, x (either from the same data or an independent data), is assigned to group 1 if $D^2(x) > 0$ (i.e., $P_1(x) > P_0(x)$), otherwise it is classified (forecast) as a 0 (i.e., $P_0(x) > P_1(x)$). [3] Also note that this larger-posterior-probability rule implies the unique threshold of 0.5 for posterior probability, because $P_0(x) + P_1(x) = 1$. In other words, as far as posterior probability is concerned it makes no sense to consider a threshold other than 50%. Therefore, the root(s) of Equation (3) are the threshold(s), and they depend on p_1. (Recall that these roots correspond to the values of x where the quantities $p_0 L_0(x)$ and $p_1 L_1(x)$

[3] An x that yields $D^2(x) = 0$ can always be assigned to one of the groups on a random basis.

intersect.) It is clear that any scalar measure of performance, through its dependence on these root(s), also depends on the prior probability p_1. One can then find the critical value of p_1 that optimizes a given measure.

There are two cases that must be treated separately - the "linear" case, if $\sigma_0 = \sigma_1$, and the "quadratic" case where $\sigma_0 \neq \sigma_1$. The reason for these names will become clear, next.

Linear Discriminant Function

In the fortunate situation where $\sigma_0 = \sigma_1 = \sigma$, (i.e., if the data is so-called homoelastic) then the discriminant function becomes linear[4] in x, and there is, then, only the one root (threshold)

$$t_{linear} = (\frac{\mu_1 + \mu_0}{2}) + \frac{\sigma^2}{\mu_1 - \mu_0} \log(\frac{1 - p_1}{p_1}). \tag{4}$$

As mentioned previously, the number of crossing points (1 or 2; excluding the ones at $+\infty$ and $-\infty$) is determined by the relative size of σ_0 and σ_1 (i.e. equal or not) and not by p_0 or p_1. Therefore, in the linear case there is only one threshold regardless of the value of prior probability.

Henceforth, and without loss of generality, we will assume $\mu_1 > \mu_0$. This can be done simply by labeling the two groups appropriately. Then $x > t_{linear}$ implies that x belongs to group 1, and $x < t_{linear}$ implies that x belongs to group 0. It is then possible to calculate the False Alarm Rate and the Miss Rate, in terms of which the various measures can be written (next section). The former is simply the rate at which 0s are classified as 1s, and the latter is the rate of misclassifying 1s as 0s, i.e.,

$$c_{01} = \int_{t_{linear}}^{\infty} L_0(x)dx, \quad \text{and} \quad c_{10} = \int_{-\infty}^{t_{linear}} L_1(x)dx.$$

[4]This has great utility in the multivariate case, because then the coefficients of the various x-terms would represent the predictive strength of the respective independent variables.

Substitution of the gaussian expressions for $L_i(x)$ yields,

$$c_{01} = \frac{1}{2}(1 - erf(t_0)), \quad \text{and} \quad c_{10} = \frac{1}{2}(1 + erf(t_1)), \tag{5}$$

where $erf(x)$ is the gaussian error function, and $t_i, (i = 0, 1)$ are defined as

$$t_i \equiv \frac{t_{linear} - \mu_i}{\sqrt{2}\sigma} = \frac{1}{\sqrt{2}} \left(\pm \frac{\delta}{2} + \frac{1}{\delta} \log(\frac{1}{p_1} - 1) \right), \tag{6}$$

where the $+, -$ signs are for $i = 0, 1$, respectively, and we have defined the useful quantity

$$\delta \equiv \frac{\mu_1 - \mu_0}{\sigma}, \tag{7}$$

to be determined from sample data. The *number* of false alarms and misses is obtained by multiplication of the rates by the respective group sample sizes: number of false alarms $= N_0\, c_{01}$, and number of misses $= N_1\, c_{10}$.

Quadratic Discriminant Function

If, as is too-often the case, $\sigma_0 \neq \sigma_1$, then equation (3) in general has two roots, i.e., two thresholds. Writing the discriminant function as

$$D^2(x) = (\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2})(x - t^+)(x - t^-),$$

with the roots written as $t^\pm$, then the false alarm rate and the miss rate are given by

$$\begin{aligned} c_{01} &= \int_{t^-}^{t^+} L_0(x)\, dx = \frac{1}{2}[erf(t_0^+) - erf(t_0^-)], \\ c_{10} &= \int_{-\infty}^{t^-} L_1(x)\, dx + \int_{t^+}^{\infty} L_1(x)\, dx = 1 - \frac{1}{2}[erf(t_1^+) - erf(t_1^-)], \end{aligned} \tag{8}$$

if $\sigma_0^2 > \sigma_1^2$. By exchanging $0 \leftrightarrow 1$ in these equations, one obtains the respective rates, if $\sigma_1^2 > \sigma_0^2$. The $t_i^\pm, (i = 0, 1)$ are defined as

$$t_i^\pm \equiv \frac{t^\pm - \mu_i}{\sqrt{2}\sigma_i}.$$

Thus far, we have not written the expression for the roots $t^{\pm}$, simply because the relevant quantities in (8) depend on $t_0^{\pm}$ and $t_1^{\pm}$, and they can be found to be

$$t_{0,1}^{\pm} = \frac{\delta_{0,1}}{\sqrt{2}(\delta_0^2 - \delta_1^2)} \left(-\delta_{1,0}^2 \ \begin{matrix} \pm \\ \mp \end{matrix} \ \sqrt{\delta_0^2\delta_1^2 + (\delta_0^2 - \delta_1^2)[\log{(\frac{\delta_0}{\delta_1})}^2 + \log{(\frac{1}{p_1} - 1)}^2]} \right), \tag{9}$$

where

$$\delta_i \equiv \frac{\mu_1 - \mu_0}{\sigma_i}, \quad (i = 0, 1) \tag{10}$$

In equation (9) the upper signs $\pm$ apply when $\delta_0 > \delta_1$, and the lower signs $\mp$ apply when $\delta_1 > \delta_0$.

The two roots are Real (non-imaginary) if only if the expression under the square-root is non-negative, which translates into a constraint on p_1:

$$\text{If } \ \delta_0^2 \geq \delta_1^2, \ \text{ then } \ p_1 \leq p_{cc}, \ \text{ otherwise } \ p_1 \geq p_{cc},$$

where p_{cc} is defined as

$$p_{cc} \equiv \left(1 + \frac{\delta_1}{\delta_0} \, exp^{\frac{\delta_0^2\delta_1^2}{2(\delta_1^2 - \delta_0^2)}} \right)^{-1}. \tag{11}$$

In other words, there is a range of values for p_1 where there is no (real) threshold at all. For such values of p_1, there does not exist a discriminant function. It is easy to understand this phenomenon: recall that the two roots correspond to the two crossing points of $p_0 L_0(x)$ and $p_1 L_1(x)$, but there exists some value of p_1 for which one of the two quantities lies entirely below the other, and so there are no crossing points. In such a case, *all* observations of x are then persistently classified as belonging to the group with the higher $p_i L_i(x)$. Note that in the linear case when $\sigma_1 = \sigma_0$, none of the groups can have a $p_i L_i(x)$ that lies entirely below the other, and so p_{cc} is specific to the quadratic case.

The above results are relevant mostly for scalar measures of categorical forecasts, and the critical values of p_1 that optimize the measures will be presented in Section 5. At the

other extreme, if forecasts are probabilistic, and one has the full luxury of dealing with multidimensional, probabilistic measures, then a different approach must be adopted. Section 5 also presents the critical value of p_1 that yields the most reliable plot in a reliability diagram (in complete generality, i.e. without reference to a Gaussian model).

4 Measures

In a Bayesian approach to forecasting, the dependence of posterior probabilities on prior probabilities is transmitted to the performance measures - both the scalar and multidimensional measures, although in the latter case there is no reason to set the prior probabilities to a value different from that estimated from the group sample sizes. The dependence of the false alarm rate and the miss rate on p_1 is now explicit in eqs (5), (6), and (8), (9). We must now write the measures in terms of these rates. The four scalar measures FRC, CSI, TSS, and HSS can be defined in terms of the contingency table (otherwise known as the Confusion Matrix, or in short the C-matrix),

$$\text{C-matrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

where a and d are the number of correct forecasts of nonevents and events, respectively; b and c are the number of false alarms and misses, respectively, and are therefore given by $b = c_{01}N_0$ and $c = c_{10}N_1$. Note that $N_0 = a + b$ and $N_1 = c + d$. These four measures can be manipulated to depend on c_{01}, c_{10} and the ratio of the two group sample sizes only:

- Fraction Correct

$$\text{FRC} \equiv \frac{a+d}{N_0+N_1} = \frac{1-c_{01}}{1+\frac{N_1}{N_0}} + \frac{1-c_{10}}{1+\frac{N_0}{N_1}},$$

- Critical Success Index

$$\text{CSI} \equiv \frac{d}{b+N_1} = \frac{1-c_{10}}{1+\frac{N_0}{N_1}c_{01}},$$

ffl True Skill Statistic

$$\mathrm{TSS} \equiv \frac{ad - bc}{N_0 N_1} = 1 - c_{01} - c_{10} \ ,$$

ffl Heidke's Skill Statistic

$$\mathrm{HSS} \equiv \frac{2(ad - bc)}{N_0(b + d) + N_1(a + c)} = \frac{2(1 - c_{01} - c_{10})}{2 + (\frac{N_0}{N_1} - 1)c_{01} - (1 - \frac{N_1}{N_0})c_{10}} \ .$$

A reliability diagram is simply a graph of the observed ratio of event sample size to the total sample size, at a given value of forecast probability:

ffl Reliability diagram

$$\left.\frac{N_1}{N}\right|_{P_1} \quad \text{vs.} \quad P_1 \ .$$

where $|_{P_1}$ means "at a given value of P_1". This quantity will be derived in the next section.

Note that if $N_0 = N_1$, then HSS=TSS and FRC=(1+TSS)/2. It is interesting that TSS has no dependence (explicit or implicit) on the sample sizes, and consequently it is well-defined in the rare-event limit; see the Discussion section. As for the reliability diagram, a most reliable forecaster will produce a straight, diagonal line on that diagram, and points below (above) the diagonal line reflect over- (under-) forecasting. Recalling eqs (5) and (8) for c_{01}, c_{10}, the dependence of the scalar measures on prior probability becomes explicit. Figures 2a-d show four measures as a function of p_1, in the linear case, for several examples of N_0/N_1 and δ. Note that when $N_0 = N_1$ (Figures 2a,b) all four measures peak around $p_1 = 1/2$. The differences between the measures appear in full force, however, when $N_0 \neq N_1$ (Figures 2c,d); whereas TSS continues to peak at $p_1 = 1/2$, FRC appears to peak at $p_1 = N_1/N$, and CSI and HSS have other critical values. All of these values will be derived in the next section.

The significance of setting the prior probability at its optimal value can be seen in Figure 2c, for instance. Hastily setting p_1 at 0.5, would result in an HSS of 18%, while $p_1 = N_1/N$ would yield HSS=6%. However, HSS at its critical value is 25%. The improvement in HSS is even more significant for smaller values of δ (Figure 2c is for $\delta = 1$).

As will be shown below, it is important to point out that the critical value of p_1 that yields optimal results in a reliability diagram is exactly the "true" value as estimated from the group sample sizes, i.e. $p_1 = N_1/N$. Values of p_1 that are different from this true value appear to exist only for scalar, categorical measures.

For the quadratic case, the p_1-dependence of the four measures, for several different values of $N_0/N_1, \delta_0, \delta_1$ is illustrated in Figures 2e,f. These Figures show the effect of p_{cc} - the value of p_1 beyond which the roots of the discriminant function become imaginary (see equation (11)). As can be seen, the behavior of the curves is different depending on whether $\delta_0 > \delta_1$ or $\delta_1 > \delta_0$; in the former (Figure 2e) the curves behave similar to the linear case (Figure 2d), except that there is a forbidden region above p_{cc}. However, if $\delta_1 > \delta_0$ (Figure 2f), then not only there is a forbidden region, but also the "true" value of $p_1 = N_1/N$ falls in this forbidden region. In other words, it is possible that this natural choice of p_1 will in fact yield a classifier that has no skill at all, when performance is gauged with these measures. Also note that FRC reaches its maximum at p_{cc}.

Having obtained and illustrated the p_1-dependence of the measures, the task at hand becomes to differentiate these measures with respect to p_1 and find the roots of the resulting expressions. These critical values, p_{1c}, are the values at which the various measures are maximized. For the reliability diagram, this will not be necessary, for the value of p_1 corresponding to maximum reliability is given exactly by the "true" value (next section).

5 Some Exact Results

Given the nonlinear nature of the equations, analytic expressions for p_{1c} are difficult to derive. However, for FRC, TSS, and the reliability diagram exact results are possible. In FRC and TSS, the appearance of c_{01}, c_{10} in the numerator alone makes the calculation possible if one notes the identity

$$\frac{d}{dt} erf(t) = \frac{2}{\sqrt{\pi}} \exp^{-t^2} ,$$

which follows from the definition of the gaussian error function

$$erf(t) = \frac{2}{\sqrt{\pi}} \int_0^t \exp^{-x^2} \, dx .$$

The details of the calculation will not be presented here, but we find, in both the linear and the quadratic case,

$$\text{For FRC:} \quad p_{1c} = \frac{N_1}{N}, \text{ if } \delta_0 \geq \delta_1 \tag{12}$$

$$\text{For TSS:} \quad p_{1c} = \frac{1}{2} \ . \tag{13}$$

Similarly, for the reliability diagram

$$\text{For reliability diagram: } p_{1c} = \frac{N_1}{N} \ . \tag{14}$$

These values of p_{1c} will maximize the respective measures. Equations (12)-(14) are interesting in that they reproduce the two popular critical values - the one suggested by Bayes' Postulate (Kendall and Stuart 1969), i.e. 1/2, and the "true" value, i.e., N_1/N. So, TSS appears to have an affinity for $p_1 = 1/2$, while FRC and the reliability diagram have an affinity for N_1/N.

It is worthwhile to outline the derivation of p_{1c} for the reliability diagram found in equation (14), because it is true in general, i.e. not for Gaussian distributions only. Recall

from equation (1) that $P_1(x)$ can be written as

$$P_1(x) = \frac{1}{1 + (\frac{p_0 N_1}{p_1 N_0}) \frac{N_0(x)}{N_1(x)}} \quad ,$$

because $L_i(x) = N_i(x)/N_i$, where $N_i(x)$ is the sample size of the i^{th} class, at a given value of x. Meanwhile, the observed ratio of event sample size to the total sample size, also at a given value of x, is

$$\frac{N_1(x)}{N(x)} = \frac{N_1(x)}{N_0(x) + N_1(x)} = \frac{1}{1 + \frac{N_0(x)}{N_1(x)}}.$$

Both expressions have an x-dependence only through the ratio $N_0(x)/N_1(x)$, and therefore, eliminating this ratio from both equations yields the P_1-dependence of $N_1/(N_0 + N_1)$, i.e. a reliability diagram:

$$\frac{N_1}{N}|_{P_1} = \frac{aP_1(x)}{1 + (a-1)P_1(x)} \quad , \tag{15}$$

where $a = (p_0 N_1)/(p_1 N_0)$. Figure 3 shows the resulting reliability diagram for several values of a. Equation 15 implies that $a = 1$ yields maximum reliability; this value of a corresponds to $p_{1c} = N_1/N$. Note that this is also the value at which FRC - not the best of measures - is maximized.

The existence of c_{01}, c_{10} in the denominators of CSI and HSS renders the critical equations nonlinear, and as a result only limiting and numerical solutions for p_{1c} are possible. As examples of the former, it is possible to show that for $N_0 >> N_1$, CSI and HSS have the same p_{1c}, although that value itself can be found only numerically. The same holds if $\delta_0, \delta_1 >> 1$. And if $N_0 << N_1$, then the p_{1c} of CSI is equal to that of FRC, namely N_1/N. Also, for HSS, $p_{1c} \to N_1/N$, as $\delta \to \infty$; given that this value of p_1 is the "true" value optimizing a reliability diagram, then it is evident that a forecaster with an optimum HSS is apt to have a sub-optimum reliability plot unless the data set happens to have "large" δ or N_0/N_1.

6 Numerical Results

For the sake of brevity, in this article only HSS is considered, although the CSI results are available as well. It is important to note that all of the scalar measures are written in terms of only N_0/N_1, and δ_0, δ_1. As a result, the p_{1c} for HSS and the corresponding HSS itself (i.e., HSS(p_{1c})) can be tabulated in terms of these quantities. Again, recall that these quantities are obtained directly from the sample data.

For the linear case, $\delta_0 = \delta_1 = \delta$, and so the p_{1c} are determined from only two quantities - N_0/N_1 and δ. Figures 4a, b allow one to read-off p_{1c} and HSS(p_{1c}), respectively, given δ and N_0/N_1. For values of δ and N_0/N_1 not given in Figure 4, observe that the plots in Figure 4a asymptotically approach N_1/N for large δ, and large N_0/N_1, as was shown in the previous section. Then, one may also find HSS(p_{1c}) for large values of δ and N_0/N_1, from Figure 4b.

For the quadratic case, the critical values of p_1 can be read-off from Figures 5a-f, given $N_0/N_1, \delta_0, \delta_1$. It is sufficient to present only $N_0/N_1 = 2, 4, 8, 100, 500, 1000$, and $0 \leq \delta_0, \delta_1 \leq 7$ results, since the results for other values can be found by extrapolation. As in the linear case, $p_{1c} \to N_1/N$ for large δ_0 or δ_1. The $N_0/N_1 = 1$ case is not considered numerically, since in that case HSS=TSS, and from equation (13) we find $p_{1c} = 1/2$ for HSS, exactly.

Similarly, HSS(p_{1c}) (i.e., the maximum value of HSS) can be read off from Figures 6a-f. Again, for $N_0/N_1 = 1$, HSS($p_{1c} = 1/2$) is not presented, since it can be calculated exactly from eqs (8) through (10). Hence, Figures 5 and 6 allow one to obtain the critical value of p_1 and the corresponding HSS, given N_0/N_1, δ_0, and δ_1, all obtained from the sample data. It is evident from the defining equations of the measures in section 3, that whereas CSI is not symmetric under the exchange $0 \leftrightarrow 1$, FRC, TSS, and HSS are. Indeed, it is this symmetry that has justified the presentation of the results for only $N_0/N_1 \geq 1$, since the results for $N_0/N_1 < 1$ can be obtained by the exchange $0 \leftrightarrow 1$ everywhere.

7 Example

In order to illustrate the above methodology, an example will be considered in this section. The example is carefully selected in order to point out some of the subtleties.

The hourly surface air pressures from Syracuse, New York, for the year 1990, were considered. Each hourly observation was also accompanied by whether or not some form of precipitation - rain, various types of snow, hail, etc. - was observed. The above methodology can be applied to deduce the optimal value of prior probability for precipitation when surface air pressure is employed to forecast precipitation; performance is gauged in terms of Heidke's Skill Statistic or via a reliability diagram.

The number of precipitation and no-precipitation observations was 1,776 and 6,145, respectively; the mean pressures were 998.48 (mb) and 1002.86 (mb), and the corresponding standard deviations were 8.673 (mb) and 7.738 (mb). This is all that will be required. The actual frequency distributions and the corresponding gaussian fits are shown in Figure 7.

First, recall the condition $\mu_1 > \mu_0$ imposed at the outset of the analysis (section 2), implies that the group with the large mean must be labeled as group 1. Therefore, we have

$$\begin{aligned} N_0 &= 1776, \quad \mu_0 = 998.48, \quad \sigma_0 = 8.673, \\ N_1 &= 6145, \quad \mu_1 = 1002.86, \quad \sigma_1 = 7.738. \end{aligned}$$

Then,

$$N_0/N_1 = 0.289, \quad \delta_0 = 0.505, \quad \delta_1 = 0.566,$$

with δ_0, and δ_1 found from equation 10.

If we are willing to overlook the difference between σ_0 and σ_1, then we may work in the linear scheme with $\delta = (\delta_0 + \delta_1)/2 = 0.536$. Given δ and N_0/N_1, Figure 4a is to be consulted to obtain the desired quantity, however, the $N_0/N_1 = 0.289$ results do not appear in that

Figure. At this point, we simply recall the symmetry of the results under the exchange of the labels 0 and 1. Then, we simply look-up the $N_0/N_1 \rightarrow N_1/N_0 = 1/0.289 \sim 3.5$ results. Figure 4a suggests a value of $p_{1c} = 0.45$, but this is really the value of p_{0c} since we relabeled the groups. As a result, $p_{1c} = 1 - p_{0c} = 0.55$. Similarly, Figure 4b suggests a corresponding value of HSS$\sim$ 15%.

The single root of the linear discriminant function is 997.60(mb), computed from equation (4). This means that in this linear approximation pressures below this value are more likely to be associated with precipitation, while those above are more likely to represent no precipitation.

On the other hand, if one assures that σ_0 and σ_1 are statistically distinct (more on this, below), then the quadratic method must be employed. According to Figures 5a and 5b, since $N_0/N_1 = 3.5$ lies in between $N_0/N_1 = 2$ and 4, then p_{1c} lies in between 0.45 and 0.43. So, we may choose $p_{1c} = 0.44$, but again, this is really p_{0c} because of the relabeling. Therefore, $p_{1c} = 1 - p_{0c} = 0.56$. Similarly, HSS itself can be read-off from Figures 6a and 6b as HSS$\sim$ 20%.

This example was chosen for having a σ_0 and σ_1 that are relatively comparable in magnitude, in order to allow for the illustration of both the linear and quadratic methods. However, if the object is more than an illustration, then it behooves one to question whether or not the difference between σ_0 and σ_1 is statistically significant. This can be done by computing the confidence intervals on σ_0 and σ_1. Without going into the details, suffice it to say that in this example, although the means μ_0 and μ_1 are statistically distinct at the 99% level, σ_0 and σ_1 are statistically equivalent (at the 99% level), and so the linear method is quite adequate. See the Discussion section for an interesting consequence that would have ensued if σ_0 and σ_1 had turned out to be statistically distinct.

As for the reliability diagram, the exact results of the previous section indicate that the most reliable plot is obtained when p_1 takes the "true" value N_1/N, i.e. $p_{1c} = 0.78$. Then the reliability plot is a diagonal line. Note that this value of p_1 is different from the value that optimizes HSS in either the linear or the quadratic regime.

8 Discussion

In addition to allowing for improved (scalar) performance, there is one other reason for considering a value of p_1 that is different from N_1/N, and that arises if the single variable x is the output of some sort of a regression analysis. It is entirely possible that the group sample sizes in the training data may not be proportional to the climatological one. In such a situation, it is unclear which p_1 should be selected - the p_1 of the training data or that of the test data.

It is interesting that TSS does not depend on group sample sizes (section 3). This may seem contrary to what has previously been said about TSS in the rare-event limit: On one hand, if $N_0 >> N_1$, then any reasonable classifier will yield a C-matrix with $a >> b, c, d$, from which it follows that (Doswell et al. 1990)

$$\lim_{a>>b,c,d} \mathrm{TSS} = \lim_{a>>b,c,d} \frac{ad - bc}{(a+b)(c+d)} = \lim_{a>>b,c,d} \frac{d - \frac{bc}{a}}{(1+\frac{b}{a})(c+d)} = \frac{d}{c+d} = 1 - c_{10}.$$

The right-most term in this equation is the so-called Probability of Detection which by itself constitutes an improper measure since one can optimize it by persistently forecasting all observations as "1"s. On the other hand the independence of TSS from N_0, N_1, would imply that TSS is given by $1 - c_{01} - c_{10}$ regardless of whether or not events are rare. This may appear to be paradoxical. The "catch" is embedded in the expressions $b = c_{01}N_0$ and $c = c_{10}N_1$, with c_{01}, c_{10} given by the N-independent eqs (4)-(7), which together imply that

TSS is independent of N_0 and N_1. But this is true if only if there exist unique false alarm and miss rates (c_{01} and c_{10}) that are independent of the sample sizes. In other words, assuming that there exists unique, N-independent false alarm and miss rates, then TSS is in fact a well-defined measure even in the rare-event limit. TSS is still a pathological measure even in the case where there exist unique false alarm and miss rates, in that if the former is unusually small (i.e., $c_{01} << 1$), then TSS=$1 - c_{10}$; this time, however, the problem is peculiar to the classifier, and not any rare-event conditions present in the data.

In the example, above, it was mentioned that σ_0 and σ_1 are statistically equivalent with 99% confidence, and it was concluded that for all practical purposes the linear method would be adequate. There is an interesting effect that would have occurred *if* σ_0 and σ_1 were statistically distinct: That would have implied that there would be two crossing points between $p_0 L_0(x)$ and $p_1 L_1(x)$. Indeed, the two roots can be found to be 996.5(mb) and 1043.5(mb). This, in turn, would have implied that pressures below 996.5(mb) are more likely to be associated with precipitation, and those above 996.5(mb) are more likely to be associated with no precipitation. This is as in the linear scheme and is physically acceptable; however, the existence of the second root would have implied that pressures above 1043.5(mb) are more likely to be associated with precipitation! Although, in this case, the statistical equivalence of σ_0 and σ_1 precludes such nonlinear behavior in pressure, it is important to emphasize the "ease" with which such effects may occur; all that is required is that the two groups be normally distributed and have statistically distinct standard deviations! In such a case, the "wider" distribution is guaranteed to cross the "narrower" one twice, thereby giving rise to such nonlinear effects.

9 Summary

In this article, the Bayesian approach to the transformation of a single continuous variable to a posterior probability of a corresponding event is outlined. The issue of forecast quality is then addressed in terms of four scalar performance measures for categorical (reduced) forecasts, and reliability diagrams. It is shown that these measures may be optimized by setting the event prior probability at certain critical values as given by p_{1c}, where $p_{1c} = N_1/N$ (i.e. the "true" value) for Fraction Correct and reliability diagrams, and $p_{1c} = 1/2$ for the True Skill Statistic. The Critical Success Index and Heidke's Skill Statistic do not allow for exact results, and so their values of p_{1c} are presented graphically. The use of the graphs requires only a knowledge of the means, the standard deviations and the sample-size-ratio of the two groups.

Acknowledgements

The author is grateful to John Cortinas for providing the data set employed for illustrating the methodology, in the example section, and to Kim Elmore and Jeanne Schneider for a discussion of various aspects of the data set.

References

Brooks, H., and C. A. Doswell III, 1995: A Comparison of Measures-Oriented and Distributions-Oriented Approaches to Forecast Verification. *Wea. Forecasting*, **11**, 288-303.

Donaldson, R. J., R. M. Dyer, and M. J. Krauss, 1975a: Operational Benefits of Meteorological Doppler Radar. Report AFCRL-TR-75-0103, Air Force Cambridge Research

Laboratories, 26 pp.

Donaldson, R. J., R. M. Dyer, and M. J. Krauss, 1975b: An Objective Evaluator of Techniques for Predicting Severe Weather Events. Preprints, *9th Conference on Severe Local Stroms*, Norman, OK, Amer. Meteor. Soc. 321-326.

Doswell III, C. A., R. Davies-Jones, and D. Keller, 1990: On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables. *Wea. Forecasting*, **5**, 576-585.

Gandin, and L. S., A. Murphy, 1992: Equitable Skill Scores for Categorical Forecasts. *Mon. Wea. Rev.*, **120**, 361-370.

Hamill, T. M., 1997: Reliability Diagrams for Multi-Category Probabilistic Forecasts. To appear in *Wea. Forecasting*.

Kendall, and M. G., A. Stuart, 1969: *The Advanced Theory of Statistics*, Vol. 1. Hafner Publishing Company, NY. 439 pp.

Marzban, C., and G. Stumpf, 1997: Measures of Skill: Application to a Neural Network for Severe Weather Prediction. To appear in *Wea. Forecasting*.

Mason, I., 1982: A Model for Assessment of Weather Forecasts. *Australian Meteorological Magazine*, **30:4**, 291-303.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281-293.

Murphy, A. H., 1996: The Finley affair: A signal event in the history of forecast verification. *Wea. Forecasting*, **11**, 3-20.

Murphy, A. H., and B. G. Brown, Y-S. Chen, 1989: Diagnostic Verification of Temperature Forecasts. *Wea. Forecasting*, **4**, 485-501.

Murphy, A. H., and R. L. Winkler, 1987: A General Framework for Forecast Verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

Murphy, A. H., and R. L. Winkler, 1992: Diagnostic Verification of Probability Forecasts. *Int. J. Forecasting*, **7**, 435-455.

O'Hagan, A., 1994: *Kendall's Advanced Theory of Statistics, Vol. 2B: Bayesian Inference.* Halsted Press, and imprint of John Wiley & Sons, Inc., NY. 330 pp.

Schaefer, J. T., 1990: The Critical Success Index as an Indicator of Warning Skill. *Wea. Forecasting*, **5**, 570-575.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, NY. 467 pp.

Figure Captions

Figure 1. Gaussian likelihood functions for two groups with means at $\mu = 30$ and $\mu = 50$, and a) with equal variances and b) with unequal variances. In the former, the two curves cross at only one point, but in the latter they cross at two points.

Figure 2. The p_1-dependence of four measures for some examples: a) $N_0/N_1 = 1, \delta = 1$, b) $N_0/N_1 = 1, \delta = 2$, c) $N_0/N_1 = 10, \delta = 1$, and d) $N_0/N_1 = 10, \delta = 2$, in the linear case, and e) $N_0/N_1 = 100, \delta_0 = 5, \delta_1 = 1$, f) $N_0/N_1 = 100, \delta_0 = 1, \delta_1 = 5$, in the quadratic case. The vertical lines in each graph mark the "true" value of p_1, i.e. $p_1 = N_1/N$.

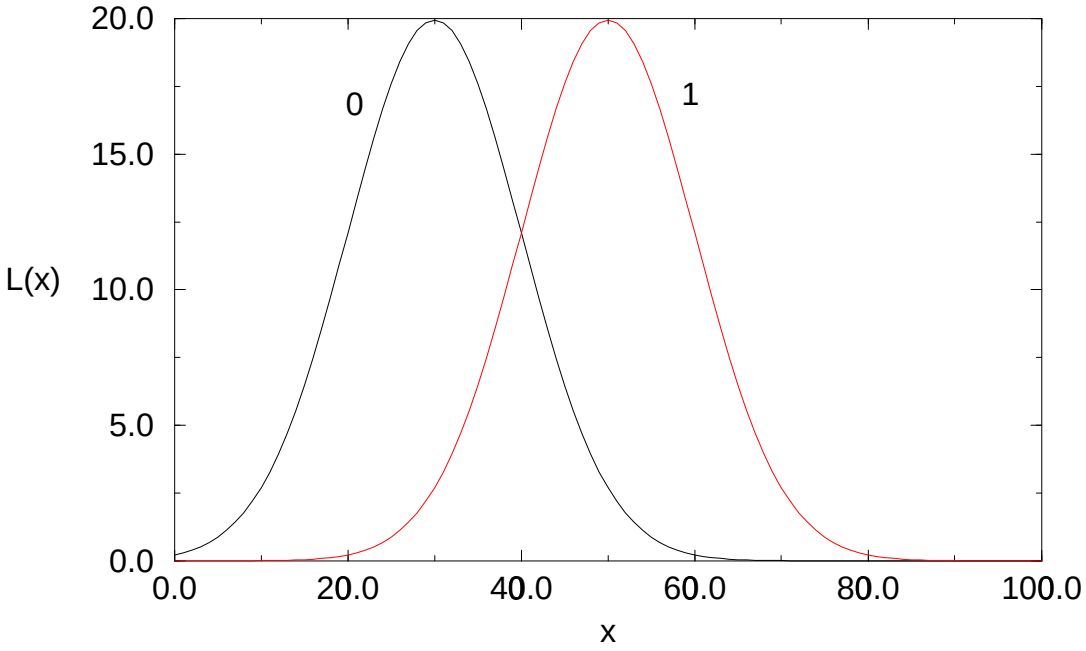
Figure 3. The reliability diagram for several values of $a = \frac{p_0 N_1}{p_1 N_0}$. Optimum reliability corresponds to $a = 1$, which translates to $p_1 = N_1/N$.

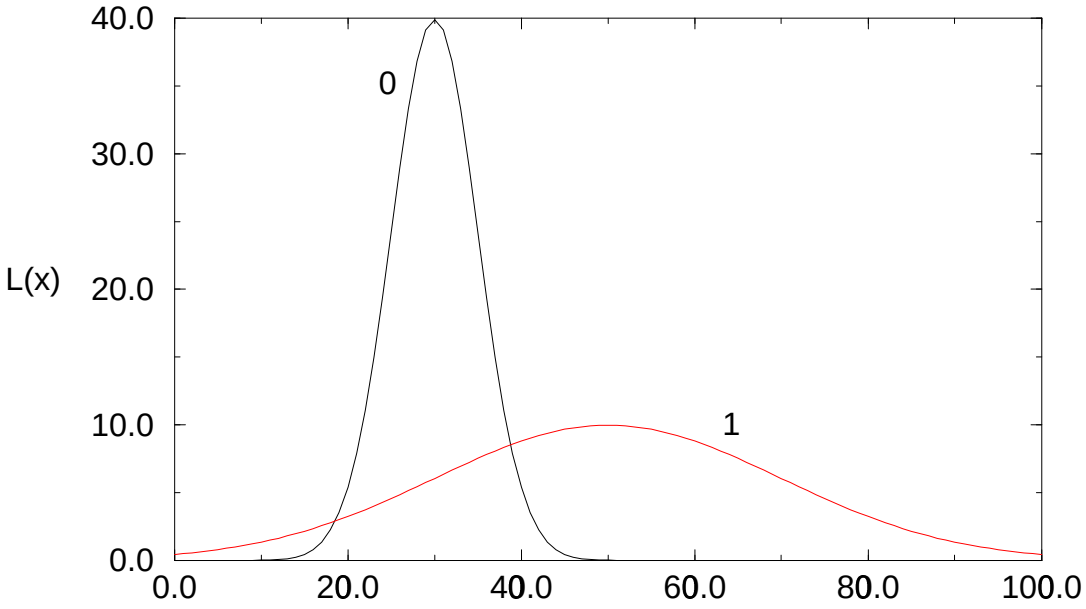
Figure 4. a) The values of p_{1c}, and b) the corresponding HSS, HSS(p_{1c}), in the linear case, as a function of δ and for 10 values of $N_0/N_1 = 1, 2, 4, 8, 16, 32, ..., 2^{10} = 1024$.

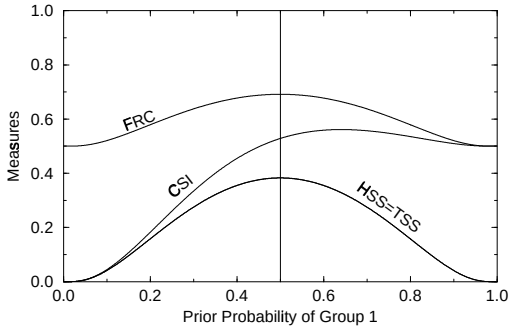
Figure 5. The values of p_{1c}, in the quadratic case, as a function of δ_0 and $\delta_1 = 0.1, 0.5, 1.0, 1.5, 2.0, ..., 7.0$, for $N_0/N_1 = 2, 4, 8, 100, 500, 1000$. (For $N_0/N_1 = 1, p_{1c} = 1/2$.)

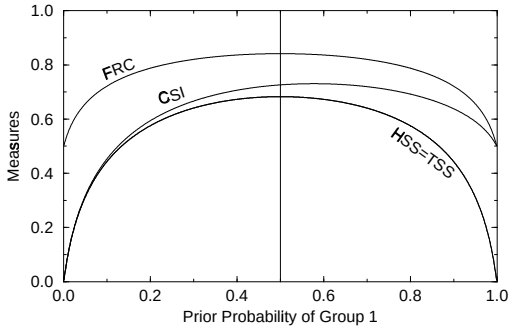
Figure 6. The values of HSS(p_{1c}), in the quadratic case, as a function of δ_0 and $\delta_1 = 0.1, 0.5, 1.0, 1.5, 2.0, ..., 7.0$, for $N_0/N_1 = 2, 4, 8, 100, 500, 1000$. (For $N_0/N_1 = 1$, HSS is found from eqs (8)-(10) with $p_{1c} = 1/2$.)

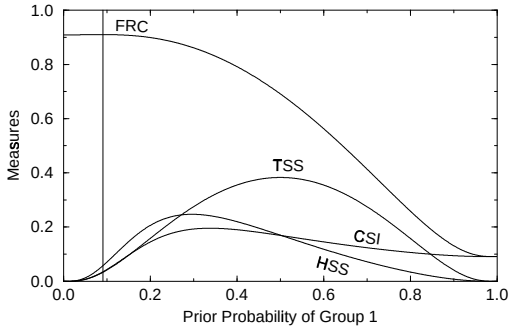
Figure 7. The distributions of hourly surface air pressures for Syracuse, NY, for the year 1990. The rough curves represent the data and the smooth curves are gaussian fits to the data. The "larger" curve is for hours when some form of precipitation occurred, and the "smaller" curve is for when there was no precipitation.

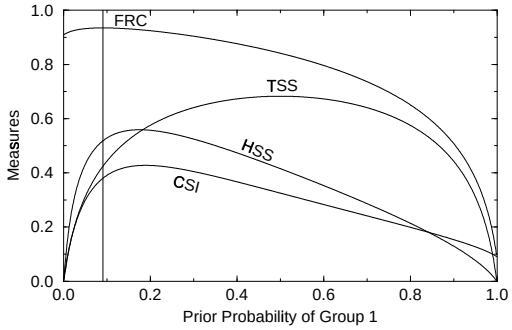

20.0
15.0
10.0
5.0
0.0
L(x)
0
1
0.0
20.0
40.0
60.0
80.0
100.0
x

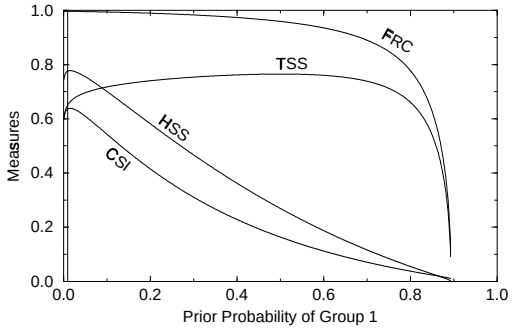

40.0
30.0
20.0
10.0
0.0
L(x)
0
1
0.0
20.0
40.0
60.0
80.0
100.0

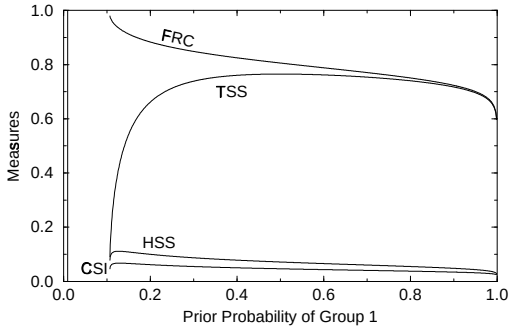

Measures
Prior Probability of Group 1
FRC
CSI
HSS=TSS

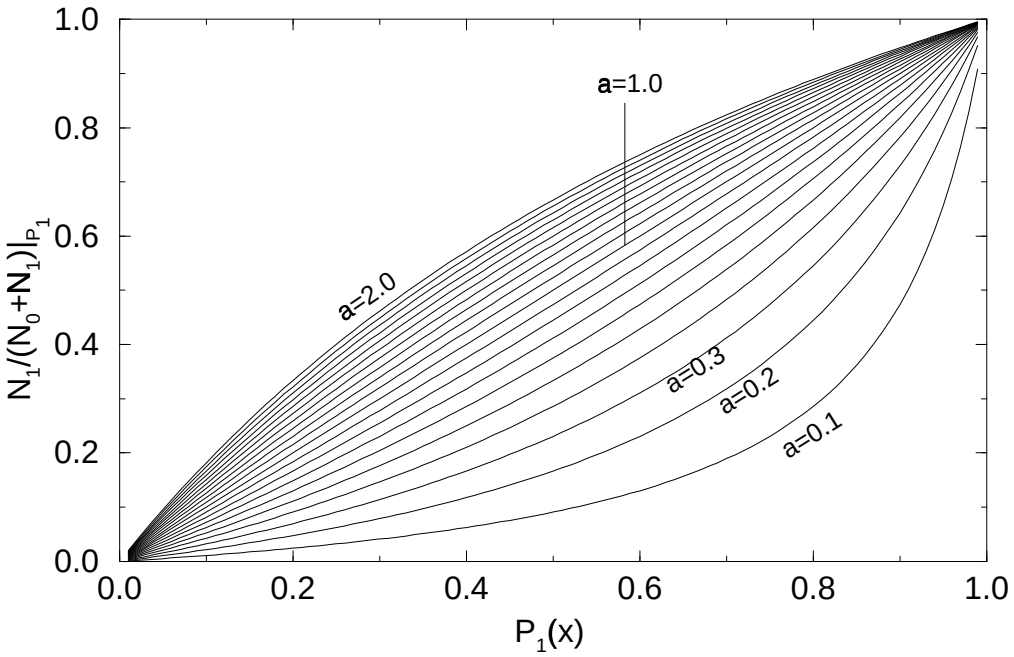

Measures
Prior Probability of Group 1
FRC
CSI
HSS=TSS

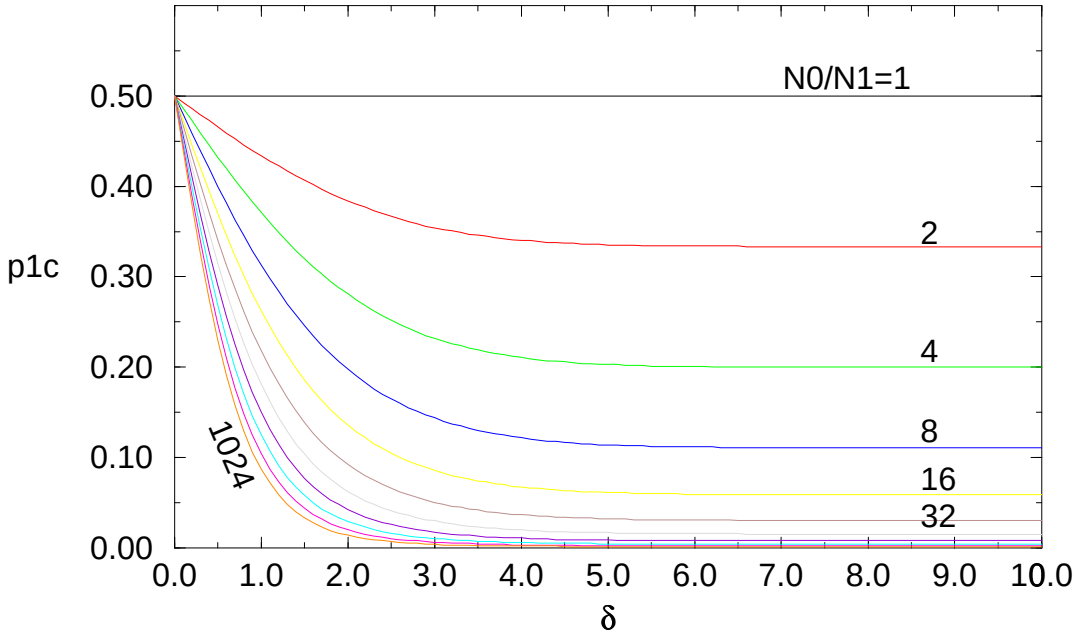

Measures
Prior Probability of Group 1
FRC
TSS
CSI
HSS

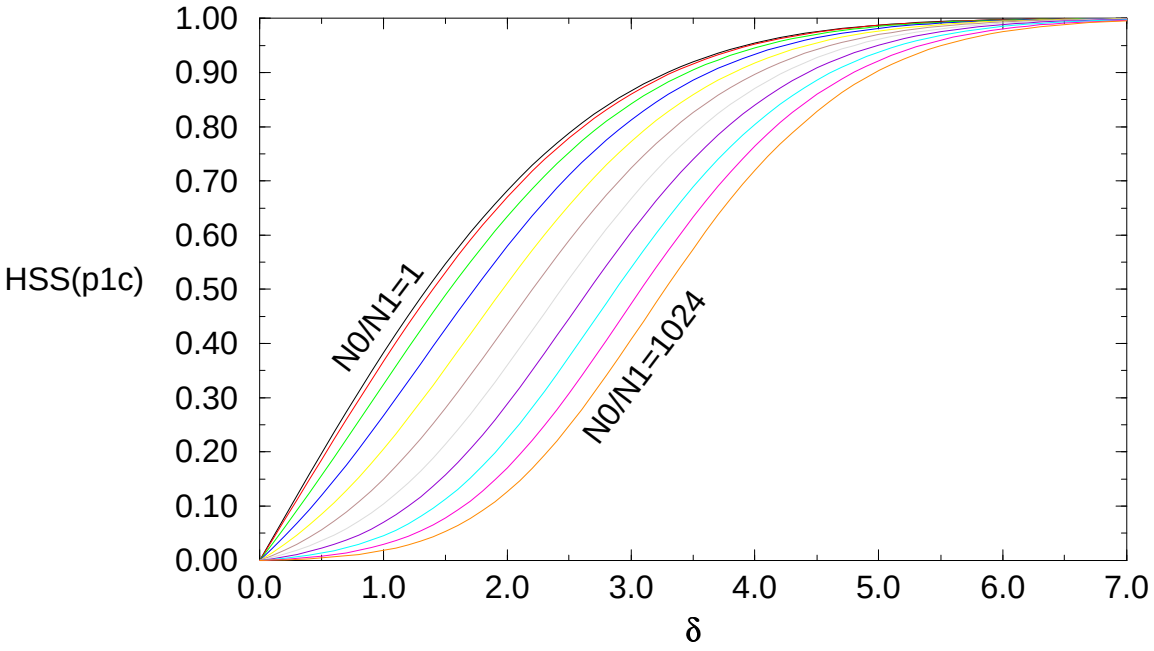

Measures
Prior Probability of Group 1
FRC
TSS
HSS
CSI

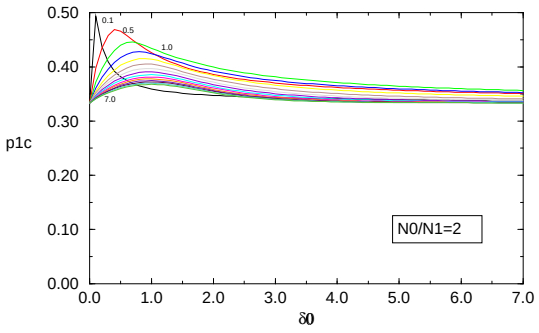

Measures
Prior Probability of Group 1
FRC
TSS
HSS
CSI

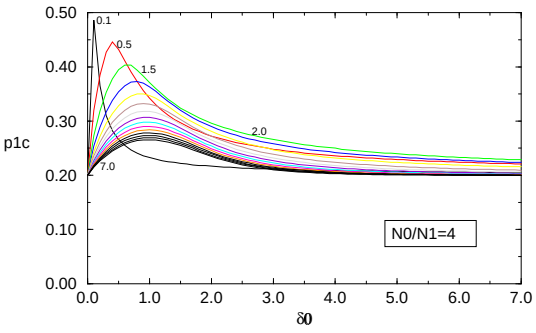

Measures
Prior Probability of Group 1
FRC
TSS
HSS
CSI

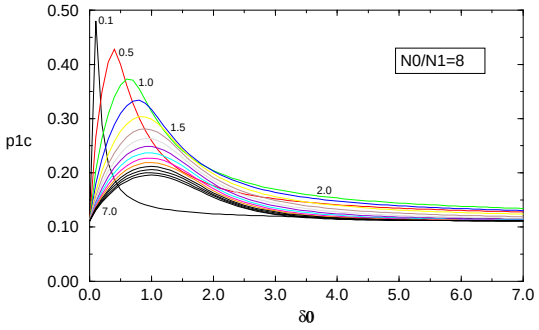

1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
$N_1/(N_0+N_1)|_{P_1}$
$P_1(x)$
a=1.0
a=2.0
a=0.3
a=0.2
a=0.1

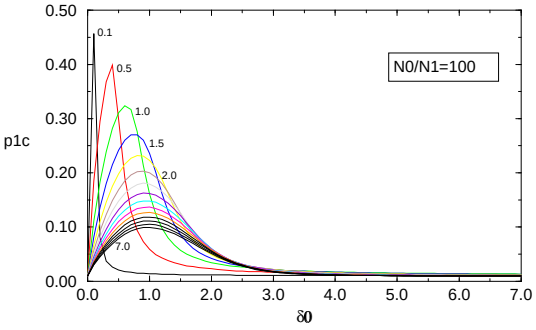

p1c
0.50
0.40
0.30
0.20
0.10
0.00
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
δ
N0/N1=1
2
4
8
16
32
1024

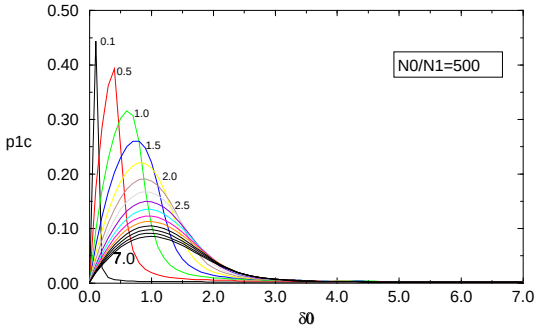

HSS(p1c)
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
δ
N0/N1=1
N0/N1=1024

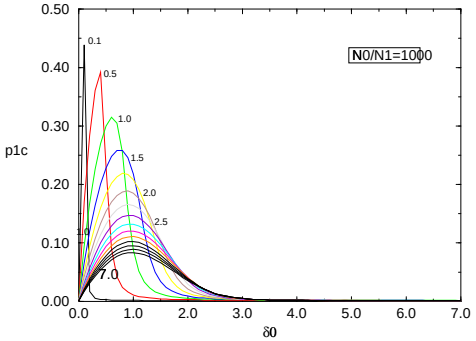

p1c
δ0
N0/N1=2
0.1
0.5
1.0
7.0

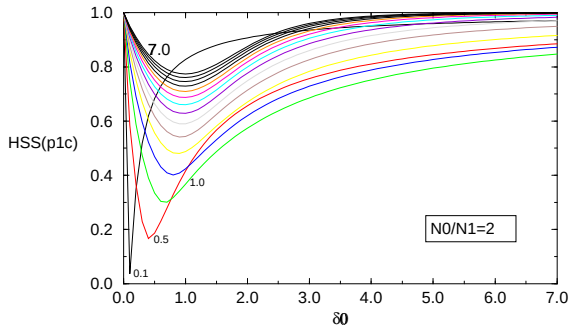

p1c
δ0
N0/N1=4
0.1
0.5
1.5
2.0
7.0

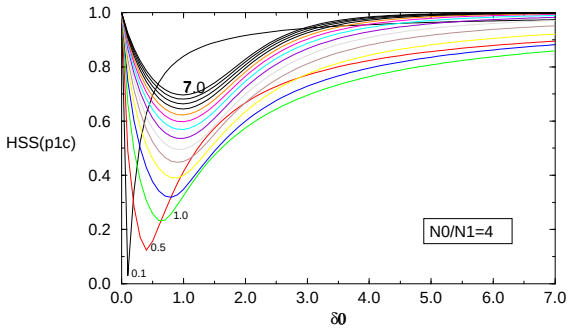

p1c
δ0
N0/N1=8
0.1
0.5
1.0
1.5
2.0
7.0

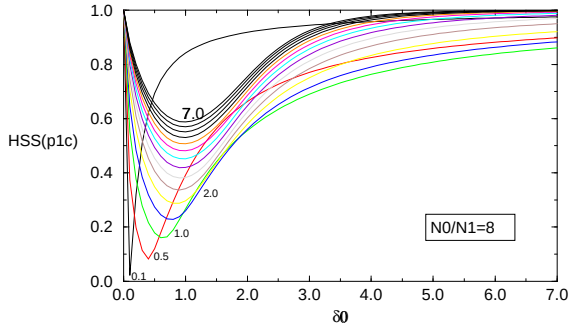

p1c
δ0
N0/N1=100
0.1
0.5
1.0
1.5
2.0
7.0

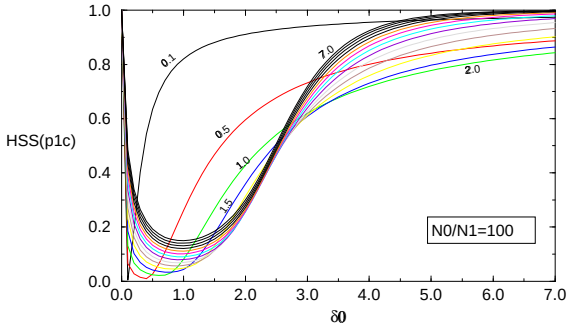

p1c
δ0
N0/N1=500
0.1
0.5
1.0
1.5
2.0
2.5
7.0

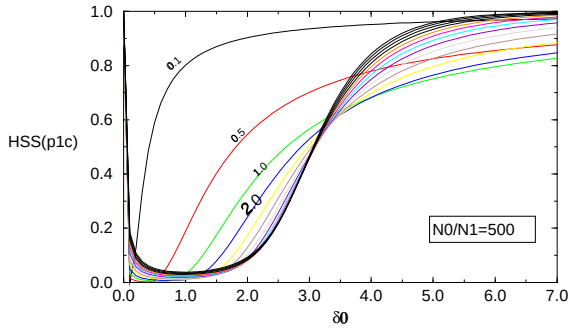

p1c
δ0
N0/N1=1000
0.1
0.5
1.0
1.5
2.0
2.5
7.0

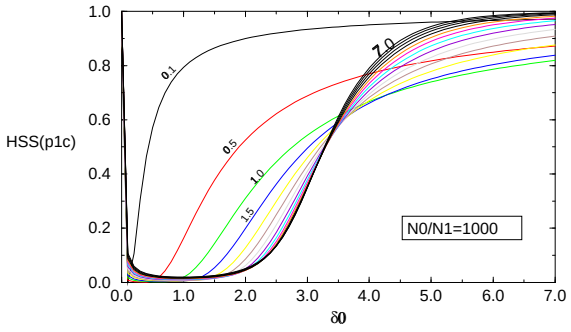

HSS(p1c)
δ0
N0/N1=2
7.0
1.0
0.5
0.1

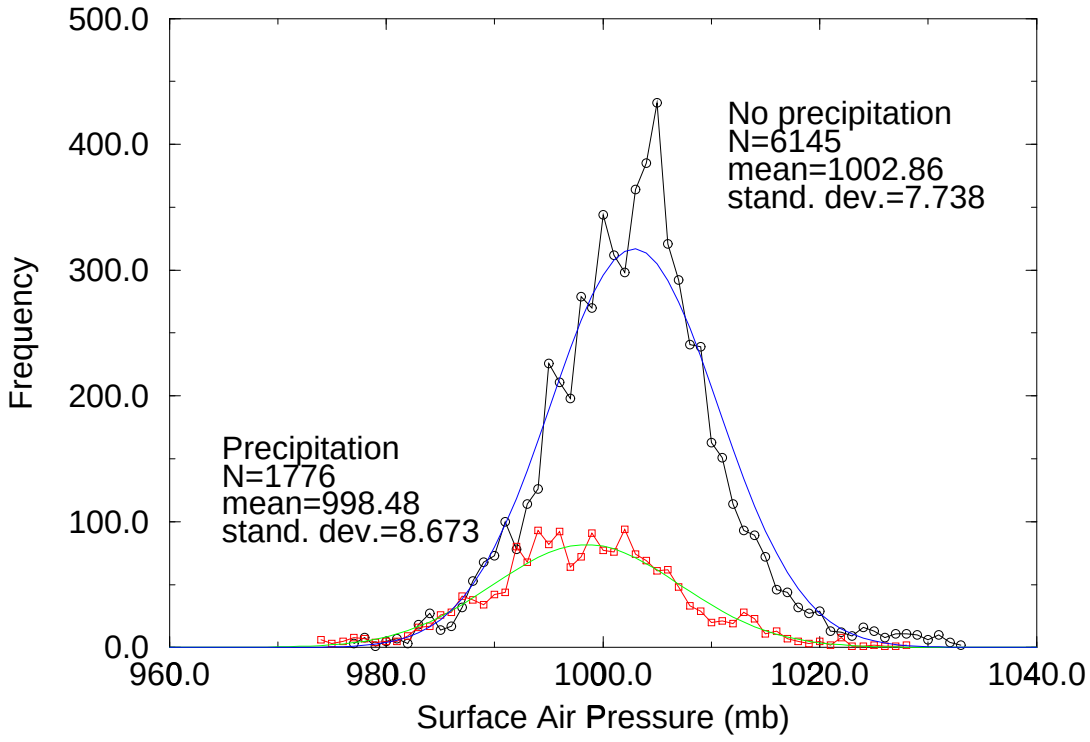

HSS(p1c)
δ0
N0/N1=4
7.0
1.0
0.5
0.1


HSS(p1c)
δ0
N0/N1=8
7.0
2.0
1.0
0.5
0.1


HSS(p1c)
δ0
N0/N1=100
0.1
7.0
2.0
0.5
1.0
1.5


HSS(p1c)
δ0
N0/N1=500
0.1
0.5
1.0
2.0


HSS(p1c)
δ0
N0/N1=1000
0.1
7.0
0.5
1.0
1.5


500.0
400.0
300.0
200.0
100.0
0.0
Frequency
960.0
980.0
1000.0
1020.0
1040.0
Surface Air Pressure (mb)
No precipitation
N=6145
mean=1002.86
stand. dev.=7.738
Precipitation
N=1776
mean=998.48
stand. dev.=8.673