Aust. Met. Mag. 45 (1996) 49–53

Confidence intervals for some performance measures of yes-no forecasts

R. Seaman
Bureau of Meteorology Research Centre, Australia
I. Mason
Department of Administrative Services, Canberra, Australia
and
F. Woodcock
Bureau of Meteorology Research Centre, Australia

(Manuscript received August 1995; revised November 1995)

The purpose of this paper is to encourage authors to include confidence intervals with some common performance measures for yes-no forecasts, when the assumption of independent forecast outcomes is not too unrealistic. Relevant theory is summarised, and a few examples are presented.

Introduction

Many forecasts and warnings issued by operational meteorological centres depend upon yes-no decisions about the occurrence of an event. Examples are rain-no rain forecasts, and warnings of severe storms, extreme fire danger, and so on. The quantitative verification of such forecasts has a history of more than 100 years (e.g. Finley 1884; Peirce 1884). Good overviews and extensive references appear in relevant chapters of Murphy and Katz (1985) and Wilks (1995).

Certain quantitative performance measures that are used to assess the skill of yes-no forecasts are simple proportions. With reference to the contingency matrix of Table 1, such measures include:

- probability of detection (POD) = $a/(a+b)$;
- false alarm ratio (FAR) = $c/(a+c)$;
- critical success index (CSI) = $a/(a+b+c)$;

and, perhaps less commonly:

- per cent correct (PC) = $100(a+d)/(a+b+c+d)$;
- probability of false detection (POFD) = $c/(c+d)$.

Corresponding author address: Mr R. Seaman, Bureau of Meteorology Research Centre, GPO Box 1289K Melbourne, Vic. 3001 Australia.

Table 1. Contingency table for definition of performance measures (see text).

		Forecast	
		Yes	*No*
Observed	*Yes*	a	b
	No	c	d

The Hanssen-Kuipers (HK) score, which has some advantages over the scores listed above (Woodcock 1976), is equivalent to (POD-POFD), and is therefore not a simple proportion like the others.

Clearly the reliability of performance measures that are simple proportions depends upon the number of decisions upon which the particular measure is based, namely, the denominator of the relevant ratio. The number of decisions is sometimes, but not always, reported together with the performance measure itself. But the number of decisions alone gives little *quantitative* idea of the reliability of the performance measure.

To the extent that a sample of yes-no forecast outcomes can be considered as a series of independent trials (more on this important proviso later), confidence intervals (CIs) are easily obtainable for the measures listed above. This paper elaborates on the preceding point, with the aim of encouraging authors to report CIs along with PODs, FARs etc. The use of CIs with such measures appears to be rare in the meteorological literature, although the use of CIs with measures analogous to POD etc. is strongly encouraged in other scientific disciplines such as medicine (e.g. Gardner and Altman 1989). Admittedly, the assumption of independent trials (discussed later) may be more problematical in meteorology.

CIs add value to a performance measure, for the forecast consumer and producer alike. The consumer is better informed about the reliability of the measure as an indicator of the underlying skill. Likewise, a knowledge of the uncertainty in a performance measure, perhaps based upon a small sample of forecasts, enables the forecast producer to more meaningfully monitor forecast performance against prespecified performance goals.

Theory

The CI for a proportion is based upon the binomial distribution. Binomial theory is extensively discussed in many statistical textbooks. The following is a summary of the relevant theory, as it applies to the POD. A similar discussion is applicable to the other measures mentioned in the introduction, with the exception of the HK score.

For the POD, we focus upon those cases where the event in question (rain, severe storm etc.) was observed. The population POD (p) is defined as the probability of correctly forecasting the event, when it occurs. From combinatorial theory, the probability of *exactly* k correct forecasts, from a sample of n independent outcomes, is given by the binomial frequency function.

$$f(k) = \frac{n!}{k!\,(n-k)!} \cdot p^k(1-p)^{n-k} \qquad \ldots 1$$

In practice, k and n are observed, and we can calculate a *sample* POD, namely $\hat{p}=k/n$, which is an estimate of the population value p. The problem is to specify a 100(1-α)%, two-sided CI, bounded by p_L and p_U, such that the probability of the interval (p_L, p_U) containing p, is 100(1-α)%. The CI is a measure of the uncertainty arising from sampling error alone. The probability associated with a CI is usually expressed as a percentage; for example the commonly used 95% CI corresponds to $\alpha = 0.05$. The so-called Clopper-Pearson CIs (Clopper and Pearson 1934) follow from Eqn 1, p_L and p_U being the solutions of the equations

$$\sum_{x=k}^{n} \frac{n!}{x!\,(n-x)!} \cdot p_L^x\,(1-p_L)^{n-x} = \alpha/2 \qquad \ldots 2$$

and

$$\sum_{x=0}^{k} \frac{n!}{x!\,(n-x)!} \cdot p_U^x\,(1-p_U)^{n-x} = \alpha/2 \qquad \ldots 3$$

The left side of Eqn 2 is the probability of k or more correct forecasts when $p=p_L$, and the left side of Eqn 3 is the probability of k or less correct forecasts when $p = p_U$. Because the binomial distribution is discrete rather than continuous, the interval (p_L, p_U), defined by Eqns 2 and 3, is conservative, in the sense that the probability of the interval containing p is equal to or greater than 100(1-α)%; see Sterne (1954) for details.

The solutions of Eqns 2 and 3, given k, n and α, for the commonly used CIs of 95% and 99%, are available in tabular or nomogram form in many textbooks (e.g. Pearson and Hartley 1966). Figure 1 is a nomogram for the Clopper-Pearson 95% CIs, constructed from Eqns 2 and 3, that provides similar information to corresponding nomograms in Pearson and Hartley, and elsewhere. The CIs corresponding to any value of α can be computed from commercially available software packages, such as IMSL (1987). A personal computer implementation for the same purpose is available from the first author.

When n is not too small, and p not too close to 0 or 1, CIs calculated from an asymptotic normal approximation to the binomial, with mean $\hat{p}$ and variance $\hat{p}(1-\hat{p})/n$, may be adequate. A criterion given in many textbooks (e.g. Hoel 1962) is that the normal approximation is good when

$$\min\,(np,\ n(1-p)) > 5 \qquad \ldots 4$$

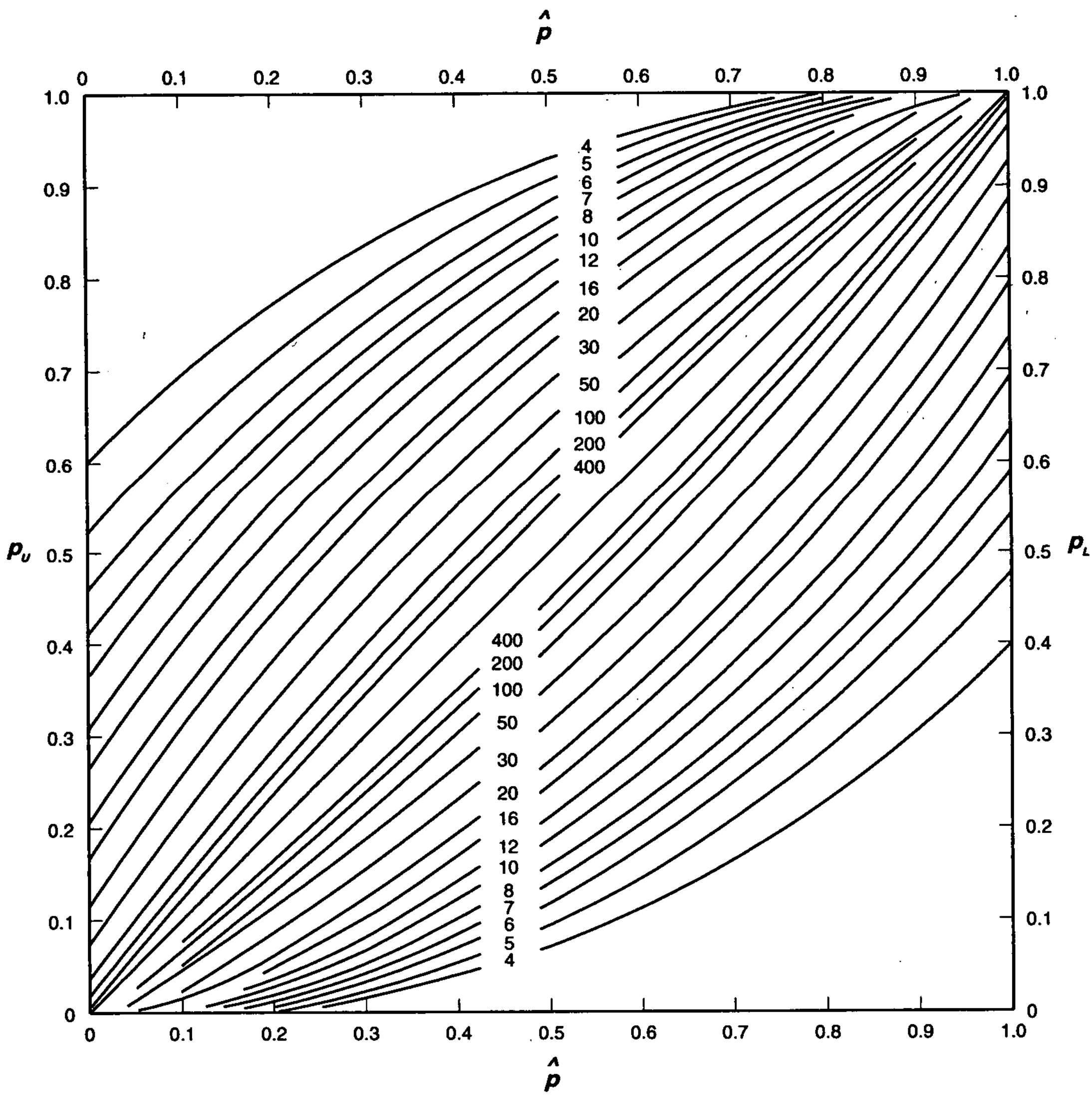
The criterion shown in Eqn 4 corresponds to the area near the middle of Fig. 1, where the lines are almost straight and inclined at about 45 degrees to the axes.

For simple proportions, there would appear to be little point in calculating approximate CIs when nomograms like Fig. 1 or their tabular counterparts are available. But the normal approximation is valuable when considering CIs for *differences* in proportions, as discussed in a following section. Also, when the normal approximation is valid, an asymptotic variance (σ^2_{HK}), and hence a CI, for the Hanssen-Kuipers score is readily calculated by making use of the fact that HK = (POD-POFD). In terms of quantities already defined

$$\sigma^2_{HK} = \text{(POD)}\,(1\text{-POD})/(a+b) + \text{(POFD)}\,(1\text{-POFD})/(c+d) \qquad \ldots 5$$

Asymptotically equivalent expressions for (σ^2_{HK}) are given by Hanssen and Kuipers (1965) and by Woodcock (1976).

Fig. 1 **Nomogram of the Clopper-Pearson 95% confidence intervals (p_L, p_U) for a proportion. Given the sample proportion $\hat{p}$ (abscissae), and the sample size n (the numbers on the curves), p_L is found using the lower set of curves with the bottom abscissa and the right ordinate. Likewise p_U is found using the upper set of curves with the top abscissa and the left ordinate.**



The assumption of independence

The theory in the preceding section is strictly valid only if the yes-no decision outcomes, upon which a performance measure is based, are independent. It is emphasised that the assumption of independence applies to the forecast outcomes; it does *not* apply to the occurrences of the meteorological event (rain, severe storm) itself, which would be a more stringent and often untenable assumption. In a yes-no forecast setting, the technical meaning of independence is that the probability of an outcome, conditional upon any other outcomes, is equal to the unconditional probability of that outcome. The assumption of indpendence would be violated if, in the case of the POD measure, a correct forecast of rain was more likely following a previous correct forecast of rain, than following a failure to forecast rain when it occurred.

The indpendence condition seems unlikely to be completely satisfied in practice. For example, the outcomes of rain-no rain forecasts on successive occasions when rain occurred arguably may be positively associated, because of the persistence of easy or difficult to forecast synoptic regimes. Quantitative evidence would be welcome. Rather more certainly, forecast outcomes in adjacent forecast districts are unlikely to be independent. But despite these considerations, CIs calculated on the assumption of independence are still useful. The presence of positive serial correlation simply means that the *effective* sample size is less than the actual sample size. For a quantitative discussion of effective sample size, see Brooks and Carruthers (1953, p. 326). When positive serial correlation may exist, a CI will be *at least as wide* as that calculated on the assumption of independence. So long as CIs are reported and interpreted in this way, they provide valuable information about the uncertainty of PODs, etc.

An obvious refinement to account for serial correlation is to attempt to estimate the effective sample size, but to do so may be easier said than done (Thiébaux and Zwiers 1984). A pragmatic approach is to exclude sufficient cases so that one is reasonably sure the remaining cases are independent. Or, in the case of spatial dependence, one could combine the results from small adjoining districts. Two other ways to account for serial correlation, both based upon statistical simulation, are described in Seaman (1992) and in Zwiers and Von Storch (1995). Overall, however, the presentation advocated in the last two sentences of the preceding paragraph seems simplest and possibly most useful.

Confidence intervals for differences in sample scores

A question often arises about the reliability of the difference between the performance measures for two samples. For example, it may be asked how reliable is an apparent improvement or deterioration of performance indicated by the PODs in successive years. For this purpose, we require the CI for a difference, rather than the CIs for performance measures themselves. Provided that the sample sizes n_1 and n_2 both satisfy Eqn 4, and there is no pairwise correlation between the samples, use of the normal approximation results in a CI for a difference in proportions bounded by d_l and d_U, where (e.g. Hoel 1962)

$$d_L = (\hat{p}_1\text{-}\hat{p}_2)\text{-}\ z_c\ (\hat{p}_1(1\text{-}\hat{p}_1)/n_1 + \hat{p}_2(1\text{-}\hat{p}_2)/n_2)^{1/2} \quad ...6$$

$$d_U = (\hat{p}_1\text{-}\hat{p}_2) + z_c\ (\hat{p}(1\text{-}\hat{p}_1)/n_1 + \hat{p}_2(1\text{-}\hat{p}_2)/n_2)^{1/2} \quad ...7$$

and z_c is the number of standard deviations (from tables of the normal distribution) corresponding to the required CI. For the 95% CI, $z_c = 1.96$.

For small samples (n_1 and/or n_2 not satisfying Eqn 4), there appears to be no simple analytic way to calculate a CI for a difference in proportions, and statistical simulation (Monte Carlo methods) may be necessary.

Examples

Rain-no rain forecasts

Table 2 shows the outcomes of rain-no rain forecasts during one year for an Australian city. The forecasts were made at 2000 local time for the 24-hour period commencing at 0900 the following day. Corresponding sample performance measures, with 95% CIs based upon independent outcomes in parentheses, are:

POD = 0.80 (0.72, 0.86),
FAR = 0.37 (0.29, 0.44),
CSI = 0.54 (0.47, 0.62).

Were there any reason to suspect lack of independence in outcomes, for example from historical forecast data for the location, the above CIs should be qualified as indicated in the section before last.

Table 2. Outcomes of rain-no rain forecasts at an Australian city (see text for further detail).

		Forecast	
		Yes	*No*
Observed	*Yes*	113	29
	No	65	158

Severe weather warnings

Table 3 shows the outcomes of severe weather warnings issued for the greater metropolitan area of an Australian city during one year. Note the absence of the 'not forecast — not observed' category; this arises because severe weather warnings are issued as needed. Note also the small number of cases in each category. Corresponding sample performance measures, with 95% CIs based upon independent forecast outcomes, are:

POD = 0.87 (0.47,1.00),
FAR = 0.42 (0.15, 0.73),
CSI = 0.55 (0.25, 0.81).

The upper bound of 1.00 for the POD should be interpreted as greater than 0.99; upper bounds are rounded upwards and lower bounds downwards.

Table 3. Outcomes of severe weather warnings for the greater metropolitan area of an Australian city. n/a=not applicable. See text for further detail.

		Warning	
		Yes	*No*
Observed	*Yes*	7	1
	No	5	n/a

Since severe weather warnings tend to be associated with events that are well separated in time, the assumption of independent outcomes may be quite good for a single district. The points to note in this example are the wide CIs, associated with small sample sizes. In particular, one would be wise not to become too euphoric about the POD result.

Year-to-year differences

Table 4 shows outcomes of rain-no rain forecasts for successive years at an Australian city. The forecasts were of a similar nature to those in the first example. In year 1, $(\mathrm{POD})_1 = 0.73$ and $(\mathrm{FAR})_1 = 0.41$; in year 2, $(\mathrm{POD})_2 = 0.74$ and $(\mathrm{FAR})_2 = 0.36$. The question is asked whether the improvement suggested by the lower FAR in year 2 can be ascribed to sampling variations.

From Eqns 6 and 7, the 95% CI for $(\mathrm{FAR})_2 - (\mathrm{FAR})_1$ is $(-0.16, 0.05)$. Since the CI encloses zero, it would be premature to conclude that the apparent improvement in FAR was due to anything other than sampling variability.

Concluding remarks

The inclusion of CIs along with yes-no forecast performance measures such as POD, FAR and CSI is to be encouraged in circumstances where the assumption of largely independent forecast outcomes is not too unrealistic. Whether or not the assumption of independent outcomes is reasonable is probably best gauged by a study of the outcomes of past forecasts made in comparable circumstances. As a general principle, forecasts that are based on much the same information are unlikely to have independent outcomes, while forecasts that are well separated in space and time are much more likely to have independent outcomes.

The use of CIs is particularly appropriate when the forecast sample is small, in order to ensure that the reader is warned of the unreliability of the corresponding performance measure in an explicit and quantitative way. But even when the sample size is large, as in the first example of the preceding section, CIs still add value to the presentation, since a narrow CI confirms the reliability of performance measure.

References

Brooks, C.E.P. and Carruthers, N. 1953. *Handbook of statistical methods in meteorology*. London, H.M.S.O.

Clopper, C.J. and Pearson, E.S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26*, 404–13.

Finley, J.P. 1884. Tornado predictions. *Amer. Met. Journ., 1*, 85–8.

Gardner, M.J. and Altman, D.G. 1989. *Statistics with confidence: Confidence intervals and statistical guidelines*. Universities Press, Belfast.

Hanssen, A.W. and Kuipers, W.J.A. 1965. On the relationship between the frequency of rain and various meteorological parameters. *Meded. Verhand., 81*, 2–15.

Hoel, P.G. 1962. *Introduction to mathematical statistics*. Third edition. John Wiley and Sons.

IMSL 1987. FORTRAN subroutines for statistical analysis. User's Manual. Version 1.0. IMSL, Houston, U.S.A.

Murphy, A.H. and Katz, R.W. (eds) 1985. *Probability, statistics and decision making in the atmospheric sciences*. Westview Press.

Pearson, E.S. and Hartley, H.O. (eds) 1966. *Biometrika tables for statisticians*. Vol 1. Third edition. Cambridge University Press.

Peirce, C.S. 1884. The numerical measure of the success of predictions. *Science, 4*, 453–4.

Seaman, R.S. 1992. Serial correlation considerations when considering differences in prediction skill. *Aust. Met. Mag., 40*, 227–38.

Sterne, T.E. 1954. Some remarks on confidence or fiducial limits. *Biometrika, 37*, 275–8.

Thiébaux, H.J. and Zwiers, F.W. 1984. The interpretation and estimation of effective sample size. *Jnl. Clim. appl. Met., 23*, 800–11.

Wilks, D.S. 1995. *Statistical methods in the atmospheric sciences. An introduction*. Academic Press.

Woodcock, F. 1976. The evaluation of yes-no forecasts for scientific and administrative purposes. *Mon. Weath. Rev., 104*, 1209–14.

Zwiers, F.W. and Von Storch, H. 1995. Taking serial correlation into account in tests of the mean. *Jnl climate, 8*, 336–51.

Table 4. Outcomes of rain-no rain forecasts for successive years at an Australian city (see text for details).

		Year 1 Forecast		*Year 2 Forecast*	
		Yes	*No*	*Yes*	*No*
Observed	*Yes*	107	40	108	37
	No	75	143	60	161