Uncertainty and Inference for Verification Measures

IAN T. JOLLIFFE

Department of Meteorology, University of Reading, Reading, United Kingdom

(Manuscript received 6 December 2005, in final form 12 July 2006)

ABSTRACT

When a forecast is assessed, a single value for a verification measure is often quoted. This is of limited use, as it needs to be complemented by some idea of the uncertainty associated with the value. If this uncertainty can be quantified, it is then possible to make statistical inferences based on the value observed. There are two main types of inference: confidence intervals can be constructed for an underlying "population" value of the measure, or hypotheses can be tested regarding the underlying value. This paper will review the main ideas of confidence intervals and hypothesis tests, together with the less well known "prediction intervals," concentrating on aspects that are often poorly understood. Comparisons will be made between different methods of constructing confidence intervals—exact, asymptotic, bootstrap, and Bayesian—and the difference between prediction intervals and confidence intervals will be explained. For hypothesis testing, multiple testing will be briefly discussed, together with connections between hypothesis testing, prediction intervals, and confidence intervals.

1. Introduction

When forecasts are made, it is generally desirable to assess how good they are. Often this is done by calculating the value of a verification measure or score, but too often this is all that is done. As noted by a reviewer, decisions concerning numerical model configurations, as well as changes in operational models have historically been based on simple comparisons of a few verification statistics, without regard for the uncertainty associated with the verification measures. Wiser and more meaningful selections and decisions would result from application of appropriate statistical theory incorporating this uncertainty.

Although uncertainty associated with the observed value of the measure is often not considered, there are some notable exceptions, for example, Bradley et al. (2003), Briggs et al. (2005), Ferro (2006, manuscript submitted to *Wea. Forecasting*), Hamill (1999), Raftery et al. (2005), Seaman et al. (1996), Stephenson (2000), Thornes and Stephenson (2001), and Woodcock (1976), together with some other references that are cited later in the paper. The value of a verification measure on its own is of little use; it also needs some quantification of the uncertainty associated with the observed value. Given this quantification, it is then possible to make statistical inferences using the observed value. For many verification measures, it is convenient to assume that what we have observed is a data-based version, $\hat{\theta}$, of an "underlying" value, θ, of the measure, and it is this underlying value that is the subject of our inferences.

One possibility is to use $\hat{\theta}$ to construct a confidence interval for θ. Often there are different ways of constructing an interval:

- based on exact probabilities,
- using asymptotic approximations,
- using resampling schemes such the bootstrap, and
- using a Bayesian rather than frequentist approach.

A simple example is introduced in section 2, and the various options for constructing confidence intervals are described and compared, in the context of two parameters of interest in this example, in section 3.

Sometimes the *difference* between the population values for two sets of forecasts, $\theta_1 - \theta_2$, is of interest. Confidence intervals in this case are discussed in section 4. Prediction intervals are closely related to, but fundamentally different from, confidence intervals. Their role in inference is discussed in section 5.

Hypothesis testing is an alternative approach to in-

Corresponding author address: Dr. Ian T Jolliffe, Dept. of Meteorology, University of Reading, Earley Gate, P.O. Box 243, Reading RG6 6BB, United Kingdom.
E-mail: i.t.jolliffe@reading.ac.uk

DOI: 10.1175/WAF989.1

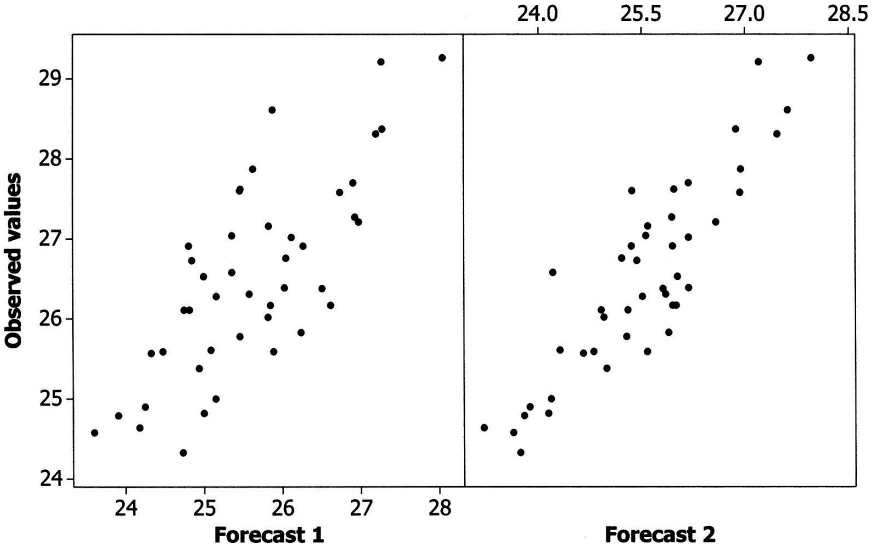

FIG. 1. Scatterplot of observed values vs forecast 1 and forecast 2. Correlations are 0.767 and 0.891, respectively.

ference about θ or $(\theta_1 - \theta_2)$. Here, there are specific hypotheses of interest. The most common are that $\theta = 0$, where zero represents a "no skill" value of the verification measure of interest, or $\theta_1 = \theta_2$ when it is required to determine whether a new forecasting system improves upon an old one with respect to the measure. The basic ideas of standard hypothesis testing (frequentist and Bayesian) are described in section 6, with discussion of links to confidence intervals and prediction intervals; the examples of sections 3 and 4 are revisited. Much of standard hypothesis testing relies on strong assumptions regarding probability distributions. Section 6 also describes permutation and randomization tests, for which the assumptions are much weaker.

The final section of the paper includes a brief discussion of some additional topics, in particular multiple testing, power, and assumptions underlying various inferential procedures, as well as some more general concluding remarks.

Although the paper contains little that is completely new, it brings together related material that is often poorly understood. It is couched in the context of forecast verification, but much of the discussion is relevant to other circumstances when uncertainty is present and inference required for weather or climate data.

2. Example

The example consists of hindcasts of sea surface temperature (SST) for the Niño-3–4 region for each of 44 yr (1958–2001). There are nine hindcasts for each year, corresponding to members of an ensemble of forecasts generated by a European Centre for Medium-Range Weather Forecasts (ECMWF) coupled ocean–atmosphere climate model with slightly different initial conditions for each member of the ensemble. Although not overly exciting in its own right, the dataset allows us to illustrate nicely the construction of confidence intervals for two different types of verification measures. Although we are actually dealing with hindcasts, for convenience we refer to them as forecasts in what follows. Nine time series of forecasts have been constructed by arbitrarily labeling the ensemble members 1–9 for each year, and constructing time series of all the members labeled 1, all the members labeled 2, and so on.

The verification measures that we use are two of the best known; similar considerations will hold for other measures. The first is the correlation coefficient between the forecast values of SST for each year and the values actually observed. This measures the strength of the linear relationship between forecast and observed values. Figure 1 shows scatterplots for the two time series of the nine that have the smallest and largest sample correlation coefficients with the observed data, namely, 0.767 and 0.891. These time series are labeled forecast 1 and forecast 2. A confidence interval is required for the underlying "population" correlation coefficient in each case.

For our second measure, we simply consider forecasts of whether the SST is above or below its mean value, and assess these binary forecasts using the measure commonly known as the hit rate, but which is also referred to as the probability of detection or sensitivity

TABLE 1. Cross classification of numbers of forecasts and observations below and above average for forecasts 1 and 3.

		Obs	Obs
		Below avg	Above avg
Forecast 1	Below avg	16	8
	Above avg	7	13
Forecast 3	Below avg	18	2
	Above avg	5	19

in other contexts. Here, the hit rate is the proportion of occurrences of above-mean SST that were successfully forecast. Table 1 gives the data for the two time series with the lowest (13/21 = 0.619) and highest (19/21 = 0.905) hit rates. The first series is forecast 1 in Fig. 1, but the second is different from forecast 2; we refer to it as forecast 3. For convenience, the sample mean has been used as the threshold converting the continuous temperature forecast into a binary forecast. In practice, thresholds will be determined in other ways, often as quantiles from large historical datasets.

Hit rate and correlation are used here mainly for illustrative purposes. It is not recommended that either be used as a sole verification measure; their limitations are discussed by Mason (2003) and Déqué (2003), respectively.

3. Confidence intervals

In the widely used frequentist approach to inference, it is supposed that a sample of data is taken from an underlying population, and some parameter, θ, of this population is of interest. In our example, θ could be the mean of a particular variable, or the population correlation coefficient between two variables, or the hit rate in a population of binary forecasts. From the data an estimate, $\hat{\theta}$, of θ is calculated. Often, when verification measures are calculated, they are simply quoted on their own, without recognizing that there is uncertainty associated with the cited value. Viewing the value as an estimate of some underlying population quantity allows us to quantify the uncertainty and to recognize that the estimate on its own is of little use. One possible way to supplement an estimate is to quote its standard error, SE($\hat{\theta}$). However, a better way is to calculate confidence intervals for θ, which (in a sense to be defined below) have a prespecified probability of enclosing the true value of θ. This probability, the confidence coefficient, is often taken to be 0.95 (95% interval) or 0.99 (99% interval), but in theory any value between 0 and 1 can be used. The interval may be of the simple symmetric form $\hat{\theta} \pm c\mathrm{SE}(\hat{\theta})$ for some constant c, but it can take other forms.

The interpretation of confidence intervals is often misunderstood. For a 95% interval it is not strictly correct to say that θ lies within the interval with probability 0.95, or that a particular interval includes θ with probability 0.95. The problem with the first statement is that it implies that θ is a random variable, which in the commonly used frequentist approach to statistical inference it is not. The parameter θ is fixed, but unknown: the interval is random. This contrasts with the Bayesian approach, discussed more fully below, where θ *is* a random variable.

The second statement is flawed because it refers to a single interval, which must necessarily include θ with a probability of either 0 or 1. The correct interpretation of 95% (99%) confidence intervals is that if we construct an infinite number of such intervals, then 95% (99%) of them will include the parameter of interest.

a. *Example 3.1: Hit rates*

Consider forecasts 1 and 3 described in the previous section, and suppose that the data used to construct them can be considered as a random sample from some larger population. Given the two observed hit rates of 0.619 and 0.905 for forecasts of above-average SST, find 95% confidence intervals for the hit rates in the underlying populations.

Like several other verification measures, the hit rate is the proportion of times that something occurs—in this case the proportion of occurrences of the event of interest that were successfully forecast. Denote such a proportion by $\hat{p}$. A confidence interval can be found for the underlying probability p of a correct forecast given that the event occurred. The situation is then the standard one of finding a confidence interval for the "probability of success" (p), in a binomial distribution, and there are various fairly well-known ways of tackling this (Agresti 2002; Garthwaite et al. 2002; Hogg and Tanis 2001), as follows.

1) APPROXIMATE (ASYMPTOTIC) INTERVALS

If the number of trials, n, is large, then $\hat{p}$ has approximately a Gaussian distribution with mean p and variance $p(1 - p)/n$. Using this distribution leads to a confidence interval whose endpoints are the roots of the quadratic equation (Garthwaite et al. 2002, section 5.2.2):

$$(n + z_{\alpha/2}^2)p^2 - (2n\hat{p} + z_{\alpha/2}^2)p + n\hat{p}^2 = 0. \qquad (1)$$

Here, $(1 - \alpha)$ is the confidence coefficient, and $z_{\alpha/2}$ is the quantile of a standard Gaussian distribution (mean, 0; variance, 1) that is exceeded with probability $\alpha/2$. For a 95% (99%) interval, $z_{\alpha/2}$ = 1.960 (2.576). For our

TABLE 2. The 95% confidence intervals, including Bayesian "credible" intervals, for underlying "population" hit rates for forecasts 1 and 3. The sample estimates $\hat{p}$ are 0.619 and 0.905 for forecasts 1 and 3, respectively.

	Forecast 1	Forecast 3
Crude approximation	(0.41, 0.83)	(0.78, 1.00)
Improved approximation	(0.39, 0.81)	(0.71, 0.97)
"Exact" interval	(0.41, 0.82)	(0.68, 0.98)
Bayes interval: uniform prior	(0.41, 0.79)	(0.71, 0.97)
Bayes interval: beta (10, 5) prior	(0.48, 0.79)	(0.66, 0.92)

example, 95% confidence intervals based on this approximation are given for the two forecasts in Table 2.

2) A CRUDER APPROXIMATION

A further approximation replaces the variance of $\hat{p}$ by the approximation $\hat{p}(1 - \hat{p})/n$, and this leads to the most widely used confidence interval for p, with endpoints

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1 - \hat{p})/n}. \tag{2}$$

This interval is symmetric about $\hat{p}$, whereas that given by Eq. (1) is not. The two approximations do not look too different in Table 2. However, Eq. (2) actually gives an upper limit of 1.03 for forecast 3 but, as this is nonsensical, it has been replaced by 1.00 in Table 2. This nonsense upper limit suggests that the approximation may not be too good here. A commonly quoted condition for using the crude approximation is that Min$[n\hat{p}, n(1 - \hat{p})]$ should exceed 5. This is comfortably satisfied for forecast 1, but not for forecast 3. There is less published advice about when to use the first approximation, but the case of forecast 3 is probably near the boundary of validity. In general, as p or $\hat{p}$ get closer to 0 or 1, larger values of n are needed for the approximations to be valid.

3) AN "EXACT" INTERVAL

The two intervals above approximate the discrete binomial distribution by the continuous Gaussian distribution, and will only be valid for "large" n. For small n, we can find an interval based on the binomial distribution itself rather than a Gaussian approximation. Such intervals are sometimes called "exact," though their coverage probability (confidence coefficient) is generally not exactly that specified, because of the discreteness of the distribution. Details are not given here (see Garthwaite et al. 2002, section 5.2.3), but charts are available for finding such intervals; see, for example, Pearson and Hartley (1970, Table 41). From our Table 2 we see that the exact 95% intervals are not too different from the improved approximation, indicating that n is large enough for the approximation to be reasonably good.

4) BAYES INTERVALS

The Bayesian approach to inference is different from the more usual frequentist approach. Here, a parameter such as p is assumed to have a probability distribution so that it now makes sense to say that the probability of p falling within a given interval is 95%. To implement the Bayesian approach, it is necessary to have a prior distribution for p, which quantifies knowledge of, and more specifically uncertainty about, p before the data are observed. Denote this prior distribution by $f(p)$. The prior distribution is multiplied by the likelihood function (and normalized by the marginal distribution of the data) to give a posterior distribution. The likelihood function is simply the probability distribution for the observed data viewed as a function of p. In our example the likelihood function is the binomial distribution

$$\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}, \tag{3}$$

where x $(=n\hat{p})$ is the number of successes in n trials. The posterior distribution $f(p|x)$ is proportional to $f(p)p^x(1 - p)^{\mathrm{n}-x}$, with a constant of proportionality that ensures that $f(p|x)$ integrates to 1 over the range (0, 1). Having found $f(p|x)$, values p_L and p_U can be determined such that the probability of p falling in the interval (p_L, p_U) is some predetermined value $(1 - \alpha)$. The interval (p_L, p_U) is then a $100(1 - \alpha)$% Bayes interval (sometimes called a credible interval). Usually, p_L and p_U are chosen symmetrically, so that there is probability $\alpha/2$ of p falling below p_L or above p_U, but this is not necessary.

Bayes intervals are appealing because they have an easier interpretation than the usual frequentist confidence intervals. However, their main drawback is that it is necessary to provide a prior distribution for p. If good, quantifiable, prior knowledge is available, then Bayes intervals are a sound choice; otherwise, different researchers may come up with different prior distributions. The fact that these different prior distributions may lead to different intervals, as we see below, can be a serious drawback. Prior ignorance is often represented by an uninformative prior, but even here different choices are possible for what is meant by "uninformative," as we also see below. Table 2 gives 95% symmetric (with respect to probability) intervals corresponding to the two different prior distributions shown in Fig. 2. One of these prior distributions, labeled $f(1, 1)$, is one type of uninformative prior; it has a constant

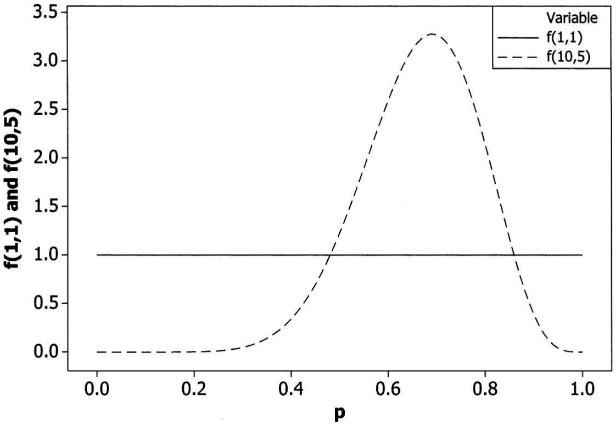

FIG. 2. A uniform prior distribution and a more informative beta prior distribution for a binomial parameter, *p*, such as hit rate.

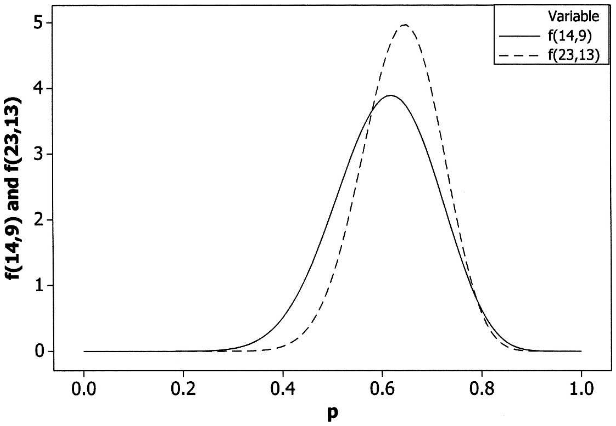

FIG. 3. Posterior distributions for *p* based on the prior distributions in Fig. 2 and the data for forecast 1.

(flat) probability density in the interval (0, 1). Although uninformative priors are often used for convenience and for mathematical reasons, there is frequently genuine prior information available concerning a parameter. In such cases, it should be used. Turning prior information into a prior distribution is not the simplest of tasks, but there are ways of doing so. For example, the second prior distribution in Fig. 2, labeled $f(10, 5)$, would be suitable if there is good prior information that the hit rate is most likely to be close to 0.67 and is unlikely to be below 0.4 or above 0.9.

The labeling of the prior distributions refers to membership of a family of distributions called beta distributions. This family is indexed by two parameters. Thus, the first distribution in Fig. 2 is a beta distribution with parameters (1, 1), and the second prior is a beta distribution with parameters (10, 5). The family can flexibly represent a wide range of probability distributions on the interval (0, 1). It also has the advantage that if a prior distribution is chosen from this family and combined with binomial data, then the posterior distribution is also a beta distribution (Epstein 1985). This property means that beta distributions give so-called conjugate prior distributions for binomial data.

Figure 3 shows the posterior distributions corresponding to the prior distributions in Fig. 2, when combined with the data from forecast 1. It can be seen that both posterior distributions are centered close to the observed proportion of hits (0.619). The posterior distribution corresponding to the flat prior distribution reflects the data, as expressed by the likelihood function. The informative prior is reinforced by the data and is narrower than either its prior distribution or the posterior distribution based on the flat prior.

Table 2 gives 95% intervals based on these two posterior distributions, for forecasts 1 and 3. It is seen that, for both forecasts, the intervals based on the flat prior are similar to other intervals in the table, but the intervals based on the informative prior distribution are narrower in the case of forecast 1, and shifted slightly to the left for forecast 3.

The flat prior $f(1, 1)$ has been described above as uninformative. A mathematical definition of uninformative leads to the distribution $f(0.5, 0.5)$: the so-called Jeffreys prior (Garthwaite et al. 2002, section 6.4). A referee has expressed a strong preference for this prior distribution, but it is U-shaped over the range (0, 1), and the present author prefers the "flat" $f(1, 1)$ as a more intuitive representation of "uninformative." Here, it makes little practical difference which is used in our example. To two decimal places, the lower limit of the 95% credible interval is the same for the two priors, and the upper limit increases from 0.79 to 0.80 for the Jeffreys prior.

The Bayes intervals in Table 2 are not very different for the two priors, because the flat prior does not contradict the message conveyed by the data and the informative prior reinforces it. If instead we consider the prior distribution $f(5, 10)$, which is the mirror image of $f(10, 5)$ and contradicts the data, the 95% Bayes interval for *p* for forecast 3 becomes (0.51, 0.81), which is quite different from any of the intervals in Table 2. Although this is a somewhat strange prior, there will be circumstances where apparent prior knowledge is misleading, and a contradiction between the messages coming from prior information and data may highlight nonstationarity, for example.

5) OTHER INTERVALS

Yet another approach is to use a resampling strategy such as the bootstrap to construct intervals. We provide

TABLE 3. The 95% confidence intervals for the correlation coefficient between observed data and forecast 1. The sample correlation coefficient r is 0.767.

Normal approx	Fisher transform	Percentile bootstrap	"Basic" bootstrap	Fisher-transformed basic bootstrap
(0.65, 0.89)	(0.61, 0.88)	(0.54, 0.90)	(0.63, 0.99)	(0.49, 0.89)

discussion and illustration of this approach below, in the context of finding confidence intervals for a correlation coefficient. Kane and Brown (2000) included bootstrap intervals in a comparison of different confidence intervals for the probability of detection (another name for hit rate) and for the probability of false detection. Their study includes intervals 2 and 3 above, but not 1 or 4. Some heavy mathematical theory is available for bootstrap intervals (Zheng and Loh 1995).

Further approaches to finding confidence intervals for a binomial parameter are described by Agresti (2002, section 1.4), with some theory in Brown et al. (2002). One assumption made in all the analyses of the example data is that the observations in different years are independent. Examination of the autocorrelation functions for the time series analyzed provides no evidence against this assumption. Miao and Gastwirth (2004) discuss intervals for the case where the observations are not independent.

b. *Example 3.2: Correlation coefficients*

Suppose we have n pairs of observations (x_1, y_1), $(\mathrm{x}_2, y_2), \ldots, (x_n, y_n)$, from which we calculate a (Pearson) correlation coefficient r between the x's and y's. It is required to find a confidence interval for the underlying population correlation coefficient ρ. As in example 1, we compare a number of ways for constructing such an interval.

1) CRUDE APPROXIMATION

The distribution of r can be approximated by a Gaussian distribution with mean ρ and variance $(1 - \rho^2)^2/n$. From this, using the same reasoning as for the cruder approximation in example 1, an approximate $100(1 - \alpha)\%$ confidence interval with endpoints

$$r \pm z_{\alpha/2}(1 - r^2)/\sqrt{n} \qquad (4)$$

can be constructed. The 95% intervals based on this expression are given in Table 3 for forecast 1.

2) BETTER APPROXIMATION: FISHER'S z TRANSFORMATION

The approximation above is generally poor unless n is very large. A much better approximation can be found using Fisher's z transformation. This transforms r to $z = ½ \ln[(1 + r)/(1 - r)]$. Then, z is approximately Gaussian with mean $½ \ln[(1 + \rho)/(1 - \rho)]$ and variance $(n - 3)^{-1}$. From this we can find a confidence interval for the transformed version of ρ, and then use the inverse transformation to obtain a confidence interval for ρ itself. The 95% confidence intervals based on forecast 1 are given in Table 3. The difference from the interval based on the crude approximation is small in this example, but the Fisher transformation often gives a better approximation. Note that all the intervals in Table 3 exclude zero by a wide margin, providing strong evidence that $\rho = 0$ is not true, and hence there is a (linear) relationship between forecasts and observations.

3) BOOTSTRAP INTERVALS

The idea behind the bootstrap is to avoid making assumptions about the distribution of R (the random variable of which r is a particular value), by repeatedly resampling from the data points themselves, calculating r for each sample, and hence building up a distribution for R empirically using these repeated values. Suppose that we take B random samples of size n *with replacement* from the n pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, and calculate r for each sample. The fact that the sampling is done with replacement means that usually some pairs of (x, y) values will be repeated in a sample, whereas others will not be included. The B calculated values of r are then ranked. For a confidence interval with confidence coefficient $(1 - \alpha)$, find the $(B\alpha/2)$th smallest and the $(B\alpha/2)$th largest values among the B ranked values of r. Call these values l and u. The most intuitive type of bootstrap interval, called a percentile interval, is simply (l, u). However, there are several other varieties of bootstrap intervals. Another simple one, sometimes known as the basic bootstrap, constructs the interval as $[r - (u - r), r + (r - \mathrm{l})]$, where r denotes the actual observed value of R. Other bootstrap intervals, such as the BC_α interval, are more complicated and aim to improve the properties of the basic and percentile intervals. Further information on these intervals can be found in Efron and Tibshirani (1993) or Garthwaite et al. (2002).

For forecast 1, in which $r = 0.767$ with $n = 44$, $B = 80$ bootstrap samples are taken. The values of r for each sample are displayed in Fig. 4. For a 95% interval and $B = 80$, we find that l equals the second smallest and u

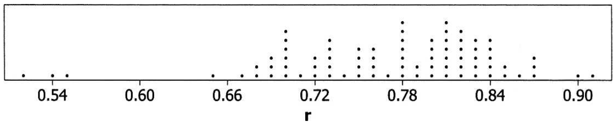

FIG. 4. Dot plot of values of the correlation coefficient r for 80 bootstrap samples for forecast 1.

the second largest of the 80 values of r. The percentile interval is (0.54, 0.90) and the basic bootstrap interval is (0.63, 0.99). These are quite different.

It is usual, and recommended, to use larger values of B, typically $B = 1000$. An advantage of using a small value in this illustration is that all the individual bootstrapped values of r can be seen in Fig. 4.

The basic bootstrap confidence interval can perform poorly (Efron and Tibshirani 1993, section 13.4) but is more likely to be satisfactory if the statistic on which it is based has an approximately Gaussian distribution. Thus, for r it is advisable to use Fisher's z transformation on the B values of r, calculate the basic bootstrap interval based on the transformed data, and then use the inverse transformation. The interval thus found is included in Table 3.

4) WHICH INTERVAL IS BEST?

The intervals given in Table 3 have much greater differences than those in Table 2: the question naturally arises, Which of the five intervals is best? First, we need to define what we mean by "best." Usually for confidence intervals this means how close is the actual coverage of the interval to the nominal value, for example, 95%. The Fisher transformed confidence interval is likely to be better than the approximate interval for untransformed data; similarly, the basic bootstrap derived from transformed data should be better than the untransformed interval. The percentile bootstrap interval, unlike the basic bootstrap, is invariant under monotone transformations, and the interval should be reasonably good if the distribution of a transformed version of the statistic is approximately Gaussian distributed (Efron and Tibshirani 1993, section 13.4), as it is here. The percentile interval and the two intervals based on Fisher's transformation all have similar upper limits, but quite different lower limits. Which is preferred depends on whether the outliers in Fig. 4 are thought to represent the behavior of r when sampling from the underlying population, in which case use one of the bootstrap intervals, probably the percentile interval, or whether they reflect an oddity of the particular dataset, in which case choose the z-transformation approximation. With a larger value of B, the believability or otherwise of the outlying data might have become clearer, but using the smallish value of $B = 80$ helps to illustrate that the choice between intervals may not be clear cut.

As in other parts of this review, there are additional techniques that could be mentioned, but space precludes detailed discussion. What has been discussed here is the *nonparametric* bootstrap in which the cumulative distribution function (c.d.f.) from which the data are drawn is estimated by the empirical c.d.f., which is a step function with step size $1/n$ at each of n data points arranged in ascending order. The *parametric* bootstrap assumes some distributional form for the c.d.f. but with unknown parameters. The parameters are estimated from the data and B samples are drawn from the resulting estimated c.d.f., in the same way that samples are drawn from the empirical c.d.f. in the nonparametric bootstrap; see Efron and Tibshirani (1993, chapter 21).

4. Confidence intervals for differences

Suppose that two forecasting systems, A and B, are to be compared with respect to their hit rates. It is of interest to discover whether any difference between the two hit rates $\hat{p}_1, \hat{p}_2$ is "statistically significant." In other words, is there sufficient evidence to declare that the underlying population hit rates p_1, p_2 are different? This can be tackled using hypothesis testing (see section 6), but we can also use confidence intervals to decide. One approach that is often used is to find confidence intervals for p_1 and p_2, based on $\hat{p}_1, \hat{p}_2$, respectively, and declare that p_1 and p_2 are different only if the intervals do not overlap. As we shall see shortly, this is not a powerful way of detecting differences between p_1 and p_2.

A better way to proceed is to construct a confidence interval for $p_1 - p_2$ directly. Assume for the moment that $\hat{p}_1$, $\hat{p}_2$ are independent of each other. We revisit this assumption later, but if it holds one possibility [for others see Agresti (2002), sections 3.1, 3.6, and 10.1] is to use the fact that for large n_1, n_2, the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Gaussian with mean $p_1 - p_2$ and variance $[p_1(1 - p_1)/n_1] + [p_2(1 - p_2)/n_2]$, where n_1, n_2 are the numbers of forecasts made with the two systems. This leads, after further approximation of p's by $\hat{p}$'s in the variance, to a confidence interval for $p_1 - p_2$ with endpoints

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}. \qquad (5)$$

In our example, if forecast 3 (1) corresponds to A (B), then $\hat{p}_1 = 0.905, \hat{p}_2 = 0.619, n_1 = n_2 = 21$, so that a 95%

interval has endpoints 0.29 ± 0.24, and the interval is (0.05, 0.53). The interval does not include zero, implying that there is evidence of a difference between the two underlying hit rates. Note, however, that in Table 2 the upper limit for p_2 exceeds the lower limit for p_1 for all the intervals presented. This overlap between the intervals gives the misleading impression that there is no evidence to declare a difference between p_1 and p_2.

It is almost always the case that if two parameters are to be compared using confidence intervals, then it is better to construct a single interval for the difference between the parameters than to construct separate intervals for each parameter.

Lanzante (2005), citing Schenker and Gentleman (2001), makes this point in the context of comparing means, but restricts attention to the case where the two samples used to estimate the parameters to be compared are independent. However, in many situations, forecasts will be made by the two systems for the same set of days or seasons, so that there will almost certainly be a positive association between the two forecasts on the same day (season). In this case the interval for the difference calculated above is too wide: to achieve a coverage of 95%, the interval for $p_1 - p_2$ should be narrower. Another way of saying the same thing is that the confidence coefficient of the interval above is actually greater than 95%, so that the evidence is even stronger that $p_1 - p_2$ is not zero.

This problem is present in the example above. The two forecasts were made for the same set of years; they have a correlation of 0.600 between them and are therefore certainly not independent. To allow for this, the expression for the variance of $\hat{p}_1 - \hat{p}_2$ needs to be reduced by twice the covariance between $\hat{p}_1$ and $\hat{p}_2$. Details will not be given here, but when this is done the 95% interval becomes (0.12, 0.46), narrower than before.

The example illustrates nicely that the overlap or otherwise of separate confidence intervals for two parameters should not be used to judge whether the parameters could be equal. Overlap here misleadingly suggests that they might be, but an interval for their difference clearly demonstrates otherwise. However, there is a final twist to this specific example.

Observant readers will have noticed that finding a significance difference between $\hat{p}_1$ and $\hat{p}_2$ contradicts the reality that the two "forecasts" were made up of ensemble members generated in the same way; hence, $p_1 = p_2$. This contradiction illustrates another potential pitfall when comparing verification measures for more than two forecasts. Suppose we have $n > 2$ forecasts and that two are to be compared. If those two are chosen randomly from the n, then the theory above is valid. However, here we deliberately chose the two with the largest and smallest $\hat{p}$'s. Hence, the difference is necessarily bigger than for a random choice of two forecasts, and the confidence interval is inappropriate; it is too narrow. This problem is related to that of multiple testing, which is discussed in section 7.

Although this section deals only with frequentist intervals, Bayes intervals could also be found and similar considerations hold in this case.

5. Prediction intervals

Prediction intervals, sometimes known as probability intervals, are different from (and simpler in concept than) confidence intervals, but are often erroneously referred to as confidence intervals. Recall that with confidence intervals, the intervals are constructed with the aim of including an unknown parameter. The parameter is fixed (in the frequentist approach) but unknown, and the intervals are random. A prediction interval is constructed so that it includes the value of a random variable with a given probability. Thus, the interval is not random; it is the variable (that may or may not fall in the interval) that is the random quantity. For example, if a random variable X has a Gaussian distribution with mean μ and variance σ^2, then a prediction interval for a value of X has endpoints

$$\mu \pm z_{\alpha/2}\sigma. \tag{6}$$

Correlation example

Given a value r of the sample correlation coefficient R between a set of forecasts and the corresponding observations, we saw in section 3 that we can construct a confidence interval for the underlying population correlation coefficient ρ. If the interval does not include zero, there is evidence of a (linear) relationship between forecasts and observations. Prediction intervals can be used to address the same question in a different way.

Assume that $\rho = 0$ and find the probability distribution for R under this assumption. We can then find an interval (r_L, r_U) such that R has a predetermined probability $(1 - \alpha)$ of falling within the interval, under the assumption that $\rho = 0$. This is a $100(1 - \alpha)\%$ prediction interval for R. The endpoints of the interval r_L, r_U are usually chosen symmetrically, so that the probabilities of R falling below r_L or above r_U are both equal to $\alpha/2$, though this is not necessary. If the observed value r lies outside the interval, we can conclude that there is evidence to contradict the assumption $\rho = 0$.

For example, for large n, R has approximately a Gaussian distribution interval with mean zero and variance n^{-1}, assuming that $\rho = 0$. This leads to a symmetric $100(1 - \alpha)\%$ prediction interval for R with endpoints

$$\pm z_{\alpha/2}/\sqrt{n}. \tag{7}$$

For our example with $n = 44$, this gives ± 0.295 for a 95% interval and ± 0.388 for a 99% interval. The observed values of r are well outside these intervals, providing convincing evidence that the assumption $\rho = 0$ is not tenable, as did all the confidence intervals in Table 3.

The above discussion is in terms of frequentist prediction intervals. Bayesian intervals are different. Suppose that $f(\rho|r)$ is the posterior distribution for ρ, and $f(r_{\text{new}}|\rho, r)$ is the probability density for a new value of r (based on a dataset of the same size as before). Then the predictive distribution of r_{new} is

$$f(r_{\text{new}}|r) = \int f(r_{\text{new}}\,|\rho, r) f(\rho|r)\, d\rho. \tag{8}$$

In many circumstances r and r_{new} are based on independent data, so $f(\text{r}_{\text{new}}|\rho, r)$ reduces to $f(r_{\text{new}}|\rho)$. Given the predictive distribution in (8), a Bayesian prediction interval can be constructed in the same way that a Bayes interval is constructed from a posterior distribution.

6. Hypothesis testing

Some types of uncertainty lead us to hypothesis testing. For example, we may be uncertain whether an observed value of a verification measure is compatible with there being no underlying skill for a set of forecasts, or we may be uncertain whether the observed difference between the values of a verification measure for two forecasting systems could have arisen by chance if the two systems were equally skillful. In both these cases we can address the question using either confidence intervals or prediction intervals, but a more direct, though often equivalent, approach is via hypothesis testing. The frequentist approach is now demonstrated using two examples, and links between hypothesis testing and prediction intervals become apparent. Following that, connections between hypothesis testing and confidence intervals are described and the section is completed by a discussion of Bayesian approaches to hypothesis testing.

a. *Example 6.1: Is $\rho = 0$?*

In most hypothesis testing settings there is a null hypothesis, often denoted H_0, and the objective is to assess the strength of evidence in the data against H_0. Consider the null hypothesis H_0: $\rho = 0$. To decide whether or not H_0 is true, a test statistic (a function of the data) is calculated whose value is likely to be different when H_0 is true than when it is false. In this example the obvious test statistic is R. There are (at least) two possible approaches to assessing the evidence against H_0. One is to calculate a p value, which is the probability of getting a value of R at least extreme as that observed, r, if H_0 is true. Although p values are commonly used and arguably provide a more informative approach to hypothesis testing than that discussed below, there is little discussion of them in this paper; see Jolliffe (2004) for an explanation of the use and subtleties of p values.

Here, we concentrate on a formulation in which we also specify an alternative hypothesis H_1, sometimes known as the research hypothesis, and use the observed value of R to decide whether or not to reject H_0 in favor of H_1. To do this we specify a desired probability of type I error, that is, the probability of rejecting H_0 when H_0 is actually true. This is most often set at 0.05 (5%) or 0.01 (1%), although it could be any value between 0 and 1, depending on how important avoiding a type I error is deemed to be.

A probability of type I error (sometimes referred to as the significance level of the test) of α can be achieved if H_0 is rejected if and only if r lies outside an interval that includes R with probability $(1 - \alpha)$ under H_0, in other words a $100(1 - \alpha)\%$ prediction interval for R. Thus, there is a direct link between prediction intervals and hypothesis testing.

If the alternative hypothesis is the symmetric (two sided) H_1: $\rho \neq 0$, then the appropriate prediction interval is also symmetric. Using the same approximation to the distribution of R as in the previous section, we reject H_0 at significance level α if and only if $|r| > z_{\alpha/2}/\sqrt{n}$. For $\alpha = 0.05$ and $n = 44$, we reject H_0 for absolute values of r greater than 0.295; for $\alpha = 0.01$, absolute values of r greater than 0.388 lead to the rejection of H_0. On a point of terminology, α is very widely known as the significance level of the test. In the atmospheric sciences, but not in other disciplines, $(1 - \alpha)$ is often quoted instead so that, for example, the terminology "significant at the 95% level" is used when most others outside the discipline would say "significant at the 5% level." It is clearly desirable that such potentially confusing usage is avoided.

The alternative hypothesis above was H_1: $\rho \neq 0$ or two sided. If, before collecting the data, it was thought inconceivable that the forecasts would be worse than chance, then it is realistic to use instead the so-called one-sided alternative hypothesis H_1: $\rho > 0$. For this one-sided test, with significance level α, H_0 will be rejected in favor of H_1 if and only if $r > z_\alpha/\sqrt{n}$. In our example, if $\alpha = 0.05$, then we reject H_0 in favor of our one-sided H_1 if $r > 0.248$. This is equivalent to seeing whether r falls outside an asymmetric 95% prediction interval with $r_L = -\infty$ and $r_U = z_\alpha/\sqrt{n}$. As well as links between hypothesis testing and prediction intervals, there are also connections with confidence intervals. These will be discussed after the second example below.

1) Permutation and randomization tests

The test above made an approximate distributional assumption about R. As with bootstrap procedures for confidence intervals, there are procedures for testing hypotheses that avoid any distributional assumptions. We illustrate the idea of permutation and randomization tests for testing H_0: $\rho = 0$. Denote the forecasts and observed data by (f_i, o_i), $i = 1, 2, \ldots n$, fix the f_i's, and consider all possible permutations of the o_i's. The correlation between the f_i's and permuted o_i's is calculated for each permutation. Under H_0, all permutations are equally likely, and a permutation test with significance level α will reject H_0 if the proportion of all calculated correlations that are at least as extreme as the observed value is less than α. Here "at least as extreme" means "at least as large" when the alternative hypothesis is H_1: $\rho > 0$ and "at least as large in absolute value" for H_1: $\rho \neq 0$.

As n increases, the number of possible permutations increases rapidly and it may not be computationally feasible to calculate r for them all. In this case a randomization test can be used, which uses the same basic idea as a permutation test: calculate r for a large number of datasets that are equally likely under H_0 and see how many values are more extreme than that actually observed. However, it uses only a randomly chosen subset of all possible permutations.

Bootstrap ideas can also be applied to hypothesis testing, as well as confidence intervals. Wilks (1997) discusses a number of bootstrap-based tests, going beyond simply resampling individual data points with replacement in order to account for temporal and spatial correlations in the data.

b. Example 6.2: Is $p_1 = p_2$?

Returning to the example of section 4, where two hit rates were compared, we can formally define a null hypothesis H_0: $p_1 = p_2$, where p_1, p_2 are underlying population hit rates corresponding to observed hit rates $\hat{p}_1, \hat{p}_2$. Similar ideas hold to those of the earlier example in this section. For large n_1, n_2, under H_0 $\hat{p}_1 - \hat{p}_2$ has an approximate Gaussian distribution with zero mean and variance $[p(1 - p)/n_1] + [p(1 - p)/n_2]$, where p is the unknown common value of p_1, p_2. From this, a symmetric $100(1 - \alpha)\%$ prediction interval for $\hat{p}_1 - \hat{p}_2$ can be constructed with endpoints

$$\pm z_{\alpha/2}\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \tag{9}$$

where $\hat{p}$ is the overall hit rate when the two samples are pooled. A test of H_0 against the two-sided alternative H_1: $p_1 \neq p_2$ at significance level α will reject H_0 in favor of H_1 if and only if $\hat{p}_1 - \hat{p}_2$ falls outside this interval. Modifications to the test and prediction interval for a one-sided H_1 are fairly obvious.

For a permutation test, pool the two samples into one large sample of size $n = n_1 + n_2$. Now, consider all possible partitions of the n observations into two subsets of size n_1, n_2. For each such partition, calculate the value of $\hat{p}_1 - \hat{p}_2$. A permutation test at significance level α will reject H_0 if the proportion of all calculated values of $\hat{p}_1 - \hat{p}_2$ that are at least as extreme as that actually observed is less than α. A randomization test is constructed similarly, but based only on a randomly chosen subset of all possible partitions.

As in the first part of section 4, the methodology discussed in this example assumes that the two samples leading to $\hat{p}_1$ and $\hat{p}_2$ are independent. Adjustments incorporating the covariance between $\hat{p}_1$ and $\hat{p}_2$ will be necessary if they are not.

A general point concerning hypothesis testing is the distinction between statistical significance and practical significance. When testing a simple null hypothesis H_0 (one that specifies only a single value for a parameter, or difference between parameters), H_0 is never *exactly* true. Hence, with very large datasets it becomes inevitable that H_0 will be rejected, even when the deviation from the null value of the parameter, or difference between parameters, is too small to be of any practical significance. Conversely, with very small sample sizes, an apparently large, and practically important, deviation from the null hypothesis may not be statistically significant because large variations are possible by chance for small samples.

c. Hypothesis testing and confidence intervals

We have seen a direct connection between hypothesis testing and prediction intervals. In many situations there is also a link between hypothesis testing and confidence intervals. Suppose that θ is a parameter and we wish to find a confidence interval for θ, or test the null hypothesis H_0: $\theta = \theta_0$ for some specified value θ_0. Often, there is a close connection between these two types of inference. Suppose that $\hat{\theta}$ is an estimator for θ whose probability distribution is known; then that distribution can be used to construct a confidence interval for θ. For example, suppose that $\hat{\theta}$ has a Gaussian distribution with mean θ and known variance σ^2. Then, a $100(1 - \alpha)\%$ confidence interval for θ has endpoints

$$\hat{\theta} \pm z_{\alpha/2}\sigma, \tag{10}$$

and a test of H_0 at significance level α against the two-sided alternative H_1: $\theta \neq \theta_0$ rejects H_0 in favor of H_1 if and only if this interval does not include θ_0. Thus, there is an exact equivalence between the test of the hypothesis and the confidence interval. A similar equivalence holds if H_1 is one sided, but now, as with prediction intervals in this case, the corresponding confidence interval is one sided.

The equivalence between standard confidence intervals and tests of the hypothesis is not always exact. In particular, it does not hold exactly in the two examples of this section. Here, the distributions of the estimators R and $\hat{p}_1 - \hat{p}_2$ are approximated, and different approximations are used for hypothesis testing and confidence intervals. For hypothesis testing, the approximations assume that H_0 holds, whereas for confidence intervals the approximations are more general. Thus, under H_0: $p_1 = p_2$, the variance of $\hat{p}_1 - \hat{p}_2$ is approximated by $[\hat{p}(1 - \hat{p})/n_1] + [\hat{p}(1 - \hat{p})/n_2]$, whereas when constructing a confidence interval, the approximation is $[\hat{p}_1(1 - \hat{p}_1)/n_1] + [\hat{p}_2(1 - \hat{p}_2)/n_2]$. Inferences based on the two approximations will often be very similar, but not identical.

d. Bayesian approaches to hypothesis testing

The main idea behind Bayesian inference is that everything you need to know about a parameter is provided by its posterior distribution. Hence, if it is required to compare two hypotheses, the way to do so is to compare the posterior probabilities or distributions for parameter values corresponding to those two hypotheses. Exactly how this is done depends on whether the hypotheses of interest are simple (the parameter takes only one value) or composite (the parameter takes more than one value).

Consider the hit rate example and suppose that we want to test H_0: $p = 0.5$ against H_1: $p = 0.7$. Both hypotheses are simple. Suppose they have prior probabilities $P[H_0]$ and $P[H_1] = 1 - P[H_0]$. If $f(x|p)$ is the likelihood function and $f(x)$ the marginal distribution for the data, then

$$P[H_0|x] = \frac{P[H_0]f(x|H_0)}{f(x)}, \tag{11}$$

with a similar expression for $P[H_1|x]$. The ratio $(P[H_0|x]/P[H_1|x])$, the posterior odds, can be written

$$\frac{P[H_0|x]}{P[H_1|x]} = \frac{P[H_0]}{P[H_1]}\frac{f(x|H_0)}{f(x|H_1)}. \tag{12}$$

The first term in this factorization is the prior odds and the second is a ratio of likelihoods, sometimes known as the Bayes factor. Suppose that a priori the two hypotheses above are considered to be equally plausible. Then, the posterior odds are equal to the Bayes factor. Substituting $p = 0.5$ and $p = 0.7$ into the likelihood function (3), and using $n = 21$ and $x = 13$ as for forecast 1 gives posterior odds of 0.750. For forecast 3 with $n = 21$, $x = 19$ the posterior odds are 0.005. For forecast 1, the observed value of $\hat{p}$ is between the null and alternative values for p, so the posterior odds are not too different from 1. However, for forecast 3, H_1 is much more plausible than H_0, given the observed value of $\hat{p}$, and this is reflected in the posterior odds. If the prior odds had been different, the posterior odds would have changed correspondingly. For example, suppose that a priori we believed H_0 to be 3 times as likely as H_1 (i.e., $P[H_0] = 0.75$, $P[H_1] = 0.25$), then the posterior odds would be multiplied by 3, to give 2.250, 0.014 in the two cases above. In the first case, where the Bayes factor is close to 1, the prior information in favor of H_0 outweighs the sample information in favor of H_1. However, in the second case the sample evidence in favor of H_1 is so strong that the prior information in favor of H_0 has relatively little effect.

Posterior information regarding hypotheses is often expressed as odds, but it can equally well be expressed as posterior probabilities. For example, posterior odds of 0.750 for H_0 are equivalent to posterior probabilities $P[H_0|x] = 0.429$, $[H_1|x] = 0.571$. Although relatively little is said in this paper regarding p values, it is important to note that a p value is often misinterpreted as $P[H_0|x]$, something which it certainly is not.

It is more usual for hypotheses to be composite than simple. Suppose that we wish to test H_0: $\theta \in \omega$ against H_1: $\theta \in \Omega - \omega$, where Ω consists of all possible values of a parameter θ and ω is a subset of Ω. Then, the posterior odds of H_0 against H_1 are

$$\frac{\displaystyle\int_{\omega} f(\theta|x)\,d\theta}{\displaystyle\int_{\Omega-\omega} f(\theta|x)\,d\theta}, \tag{13}$$

where $f(\theta|x)$ is the posterior distribution of θ. Summations replace the integrals in (13) if x is discrete rather than continuous.

As an example, consider the posterior distribution $f(14, 9)$ for p in example 3.1, for forecast 1 with the uniform prior $f(1, 1)$, and suppose we wish to test

H_0: $p \leq 0.5$ against H_1: $p > 0.5$. Then the numerator in (13) is 0.143, and the denominator is 0.857; hence, the odds for H_0 versus H_1 are 0.167.

It is common to have a null hypothesis that is simple and an alternative that is composite. For example, the hypotheses might be H_0: $p = 0.5$ and H_1: $p > 0.5$. The main reason for such hypotheses is that it is convenient to have a simple null hypothesis in the frequentist approach to inference. However, it is usually more realistic to treat both hypotheses as composite, as in the example immediately above. It seems unnatural for one hypothesis to be simple and the other composite, but the Bayesian approach can deal with this case too; see, for example, Garthwaite et al. (2002, section 7.4).

7. Discussion

The paper has described the assessment of uncertainty for verification measures using confidence intervals, prediction intervals, and hypothesis testing. Little of the material is completely new, but the topics collected together are often neglected or misunderstood.

The discussion has not been exhaustive. Only two verification measures have been considered, but a variety of types of inference have been illustrated. For most other verification measures, the general principles underlying one or more of the techniques described herein will be relevant, but details of implementation will vary from measure to measure.

The inferences depend to varying degrees on assumptions about the distribution of the data from which the measures are derived, and on the independence and stationarity of forecasts (and observations) made at different times. It is not possible to give definitive advice on how important each assumption will be in individual cases. However, it will usually be the case that as samples get larger, distributional assumptions become less important (distributions of verification measures often become approximately Gaussian, regardless of the distribution of the data) but, at the same time, nonstationarity is increasingly likely to cause problems. The temptation to increase sample size may also introduce other types of inhomogeneity into the data, spatial as well as temporal, and this should be avoided. If assumptions are unlikely to be satisfied, there will often be a nonparametric approach, such as the bootstrap, that avoids them. Even if this is not the case, it is better to have some crude estimate of uncertainty than none at all.

Inference has only been considered for one verification measure at a time. There are circumstances where it will be desirable to simultaneously consider more than one measure, for example the hit rate and false alarm rate, which together define the relative operating characteristics (ROC) curve (Mason 2003). Relatively little has been published on this, but see Pepe (2003).

A problem arising in hypothesis testing when many hypotheses are tested is that some "true" null hypotheses will almost inevitably be rejected. For example, if tests are carried out at the 5% significance level, and all null hypotheses are true, then approximately 1 in 20 of these hypotheses will be wrongly rejected. The most common situation in which this occurs in meteorology or climatology is when tests of a hypothesis are conducted at a large number of grid points, and maps are drawn indicating which grid points are "significant" at various significance levels, such as 10%, 5%, and/or 1%. This occurs in studies of teleconnections, where some index is correlated with the value of an atmospheric variable at various grid points, in order to see where significant correlations occur. Suppose that the null hypothesis of interest is that the index and variable are uncorrelated at *all* grid points. In other words, we wish to assess the "overall" or "field" significance. Although this paragraph, and the next four, are couched in more general terms, this problem is certainly relevant to verification measures. We may compute a large number of verification measures, at different stations or grid points, at different times, or for different weather or climate variables, and wish to decide which, if any, values represent significant skill.

A key reference on the topic of field significance is Livezey and Chen (1983); a summary can be found in Wilks (2006a, section 5.4). Livezey and Chen (1983) note that when a set of tests are independent, a procedure based on the binomial distribution can be used to test for field significance, using the significance or otherwise of individual tests. However, tests at nearby grid points are usually far from independent, and the only way of assessing field significance is to use a Monte Carlo approach, in which many simulated datasets are generated for which the field null hypothesis is known to be true. The significance levels for the observed data at individual grid points are then compared to those for the simulated data in order to see whether those observed are extreme compared to the simulated values.

There is a large amount of statistical literature on so-called multiple testing: whole books have been written on the subject (Hochberg and Tamhane 1987; Westfall and Young 1993). The two main features of the latter are its preference for p values (see Jolliffe 2004) over fixed significance levels, a topic largely ignored in the present paper, and its advocacy of "resampling methods," in particular the bootstrap, to adjust p values obtained from individual tests to give adjusted p values that are relevant to assessing "overall significance."

Elmore et al. (2006) use a moving blocks bootstrap to establish "significance" at individual grid points, followed by Livezey and Chen's (1983) procedure to establish field significance, given the results for individual grid points.

It may be that there is interest not simply in whether the overall null hypothesis can be rejected, but in *how many* of the individual tests are significant. Westfall and Young (1993) discuss strategies to tackle this issue with a "step-down" procedure, which adjusts the smallest p value, then the second smallest, and so on. An alternative approach that has received a lot of attention recently is based on the so-called false discovery rate (FDR), an idea that was introduced by Benjamini and Hochberg (1995). The FDR is the proportion of rejected hypotheses that are actually true. Benjamini and Hochberg (1995) described a procedure for deciding how many of the null hypotheses should be rejected if the FDR is to be controlled at some chosen level q.

Subsequently, there has been much discussion of alternative procedures for controlling the FDR, and their properties, and a few relevant references are now noted. The article by Ventura et al. (2004) is of particular interest. It gives a nice description of FDR-based procedures and argues that they provide a more meaningful way of dealing with multiple testing than techniques based on field significance. Ventura et al. (2004) compare the performance, for simulated data, of the basic Benjamini and Hochberg procedure, which assumes independence of the null hypotheses, and two alternatives due to Yekutieli and Benjamini (1999) and Benjamini and Yekutieli (2001), which allow dependence. Despite the fact that climatological data are typically spatial and have dependent null hypotheses, Ventura et al. (2004) find that the alternative procedures have no advantage (in fact one is far too conservative) compared to the basic procedure. Finally, Ventura et al. (2004) note that most procedures are conservative, in that the quoted controlled level of the FDR is an upper bound, and it may not be a very strong bound. In other words, they lack the power to detect genuinely false null hypotheses. Ventura et al. (2004) describe a modification, due to Genovese and Wasserman (2004), that improves power, while still controlling the FDR. Wilks (2006b) advocates the use of the FDR in an atmospheric science context.

Other recent references related to the topic include Black (2004), Cox and Wong (2004), Efron (2004), Perone Pacifico et al. (2004), Storey (2002), and Storey et al. (2004)

Another aspect of hypothesis testing that has not been much discussed in the atmospheric science literature [Ventura et al. (2004) and Wilks (1997) are notable exceptions] is the power of a test. Very often a significance level or probability of type I error is specified, but little attention is paid to the dual probability of type II error, that is, failing to reject the null hypothesis H_0 when we should do so. The power of a test is obtained by subtracting the probability of type II error from 1; thus, it is the probability of correctly rejecting H_0. Power is rarely mentioned for verification measures. An exception is an unpublished Ph.D. thesis by Potts (1991). She simulated data from two dependent fields, where the level of dependency could be varied and the fields roughly mimicked the patterns seen in sea level pressure fields over the North Atlantic region. The power of tests based on some well-known verification measures, including weighted and unweighted versions of mean square error, correlation, and anomaly correlation, was compared. As expected, the power increased with the degree of dependency and also with the size of the spatial area (number of grid points) studied. The differences in power between different measures was quite small, though different measures will, in general, be most powerful against different alternative hypotheses.

The relevance of the topics discussed in the paper is much wider than simply for verification measures. The key parts of the discussion are applicable whenever a statistic is calculated from a dataset—some measure of the uncertainty of that statistic will always enhance its value.

Acknowledgments. I am grateful to three anonymous reviewers, the editor, and Chris Ferro, for perceptive, stimulating, and constructive comments that helped to improve this paper. Caio Coelho provided the data used in the examples. Much of the work leading to the paper was supported by a NERC Discipline Bridging Award, NER/T/S/2003/0126.

REFERENCES

Agresti, A., 2002: *Categorical Data Analysis*. 2d ed. Wiley, 710 pp.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.,* **57B,** 289–300.

——, and D. Yekutieli, 2001: The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.,* **29,** 1165–1188.

Black, M. A., 2004: A note on the adaptive control of false discovery rates. *J. Roy. Stat. Soc.,* **66B,** 297–304.

Bradley, A. A., T. Hashino, and S. S. Schwartz, 2003: Distributions-oriented verification of probability forecasts for small data samples. *Wea. Forecasting,* **18,** 903–917.

Briggs, W., M. Pocernich, and D. Ruppert, 2005: Incorporating misclassification error in skill assessment. *Mon. Wea. Rev.,* **133,** 3382–3392.

Brown, L. D., T. T. Cai, and A. DasGupta, 2002: Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Stat.,* **30,** 160–201.

Cox, D. R., and M. Y. Wong, 2004: A simple procedure for the selection of significant effects. *J. Roy. Stat. Soc.,* **66B,** 395–400.

Déqué, M., 2003: Continuous variables. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 97–119.

Efron, B., 2004: Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Stat. Assoc.,* **99,** 96–104.

——, and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall, 436 pp.

Elmore, K. L., M. E. Baldwin, and D. M. Schultz, 2006: Field significance revisited: Spatial bias errors in forecasts as applied to the Eta Model. *Mon. Wea. Rev.,* **134,** 519–531.

Epstein, E. S., 1985: *Statistical Inference and Prediction in Climatology: A Bayesian Approach. Meteor. Monogr.,* Amer. Meteor. Soc., No. 42, 199 pp.

Garthwaite, P. H., I. T. Jolliffe, and B. Jones, 2002: *Statistical Inference*. 2d ed. Oxford University Press, 328 pp.

Genovese, C., and L. Wasserman, 2004: A stochastic process approach to false discovery rates. *Ann. Stat.,* **32,** 1035–1061.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

Hochberg, Y., and A. C. Tamhane, 1987: *Multiple Comparison Procedures*. Wiley, 450 pp.

Hogg, R. V., and E. A. Tanis, 2001. *Probability and Statistical Inference*. 6th ed. Prentice Hall, 704 pp.

Jolliffe, I. T., 2004: P stands for . . . . *Weather,* **59,** 77–79.

Kane, T. L., and B. G. Brown, 2000: Confidence intervals for some verification measures—A survey of several methods. Preprints, *15th Conf. on Probability and Statistics in the Atmospheric Sciences,* Asheville, NC, Amer. Meteor. Soc., 46–49.

Lanzante, J. R., 2005: A cautionary note on the use of error bars. *J. Climate,* **18,** 3699–3703.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.,* **111,** 46–59.

Mason, I. B., 2003: Binary events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 37–76.

Miao, W., and J. L. Gastwirth, 2004: The effect of dependence on confidence intervals for a population proportion. *Amer. Stat.,* **58,** 124–130.

Pearson, E. S., and H. O. Hartley, 1970: *Biometrika Tables for Statisticians.* Vol. 1. 3d ed. Cambridge University Press, 270 pp.

Pepe, M. S., 2003: *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, 302 pp.

Perone Pacifico, M., C. Genovese, I. Verdinelli, and L. Wasserman, 2004: False discovery rate control for random fields. *J. Amer. Stat. Assoc.,* **99,** 1002–1014.

Potts, J. M., 1991: Statistical methods for the comparison of spatial patterns in meteorological variables. Ph.D. thesis, University of Kent at Canterbury, Canterbury, United Kingdom, 271 pp.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Schenker, N., and J. Gentleman, 2001: On judging the significance of differences by examining the overlap between confidence intervals. *Amer. Stat.,* **55,** 182–186.

Seaman, R., I. Mason, and F. Woodcock, 1996: Confidence intervals for some performance measures of yes–no forecasts. *Aust. Meteor. Mag.,* **45,** 49–53.

Stephenson, D. B., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting,* **15,** 221–232.

Storey, J. D., 2002: A direct approach to false discovery rates. *J. Roy. Stat. Soc.,* **64B,** 479–498.

——, J. E. Taylor, and D. Siegmund, 2004: Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Roy. Stat. Soc.,* **66B,** 187–205.

Thornes, J. E., and D. B. Stephenson, 2001: How to judge the quality and value of weather forecast products. *Meteor. Appl.,* **8,** 307–314.

Ventura, V., C. J. Paciorek, and J. S. Risbey, 2004: Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *J. Climate,* **17,** 4343–4356.

Westfall, P. H., and S. S. Young, 1993: *Resampling-Based Multiple Testing*. Wiley, 340 pp.

Wilks, D. S., 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate,* **10,** 65–82.

——, 2006a: *Statistical Methods in the Atmospheric Sciences*. 2d ed. Academic Press, 627 pp.

——, 2006b: On "field significance" and the false discovery rate. *J. Appl. Meteor. Climatol.,* **45,** 1181–1189.

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.,* **104,** 1209–1214.

Yekutieli, D., and Y. Benjamini, 1999: Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan. Info.,* **82,** 171–196.

Zheng, X., and W.-Y. Loh, 1995: Bootstrapping binomial confidence intervals. *J. Plan. Stat. Info.,* **43,** 355–380.