Estimation of uncertainty in verification measures

Ian Jolliffe

Universities of Reading, Aberdeen
and Southampton

i.t.jolliffe@reading.ac.uk

Outline of talk

- Introduction
- Intervals expressing uncertainty
  - Confidence intervals
  - Prediction intervals
- Tests of hypothesis
  - Links with intervals

Introduction

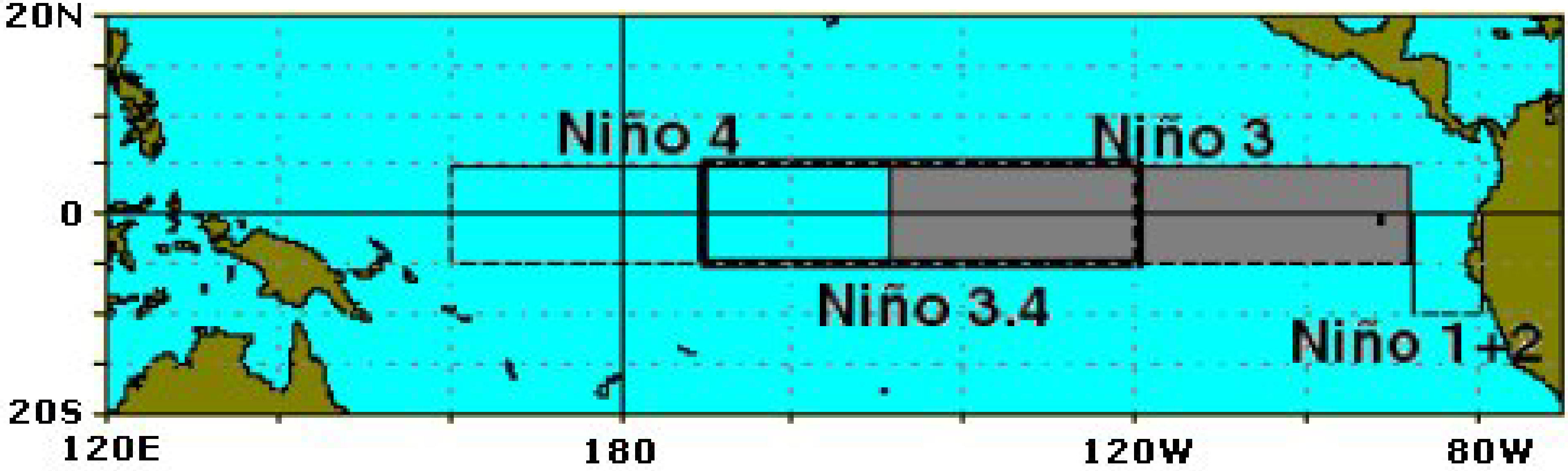
- Often values of verification measures are given without any mention of the uncertainty associated with them
- Could quote a standard error – OK, but a confidence interval is better, especially if the distribution of the measure is not close to Gaussian
- In comparing values of a measure at different times, hypothesis testing may be a good way of addressing the uncertainty associated with an improvement

Confidence intervals

- Given a sample value of a measure find an interval with a specified level of confidence (e.g 95%, 99%) of including the corresponding population value of the measure
- Note:
  - the interval is random; the population value is fixed
  - it is assumed that the data we have are a random sample from some larger (possibly hypothetical) population

Example

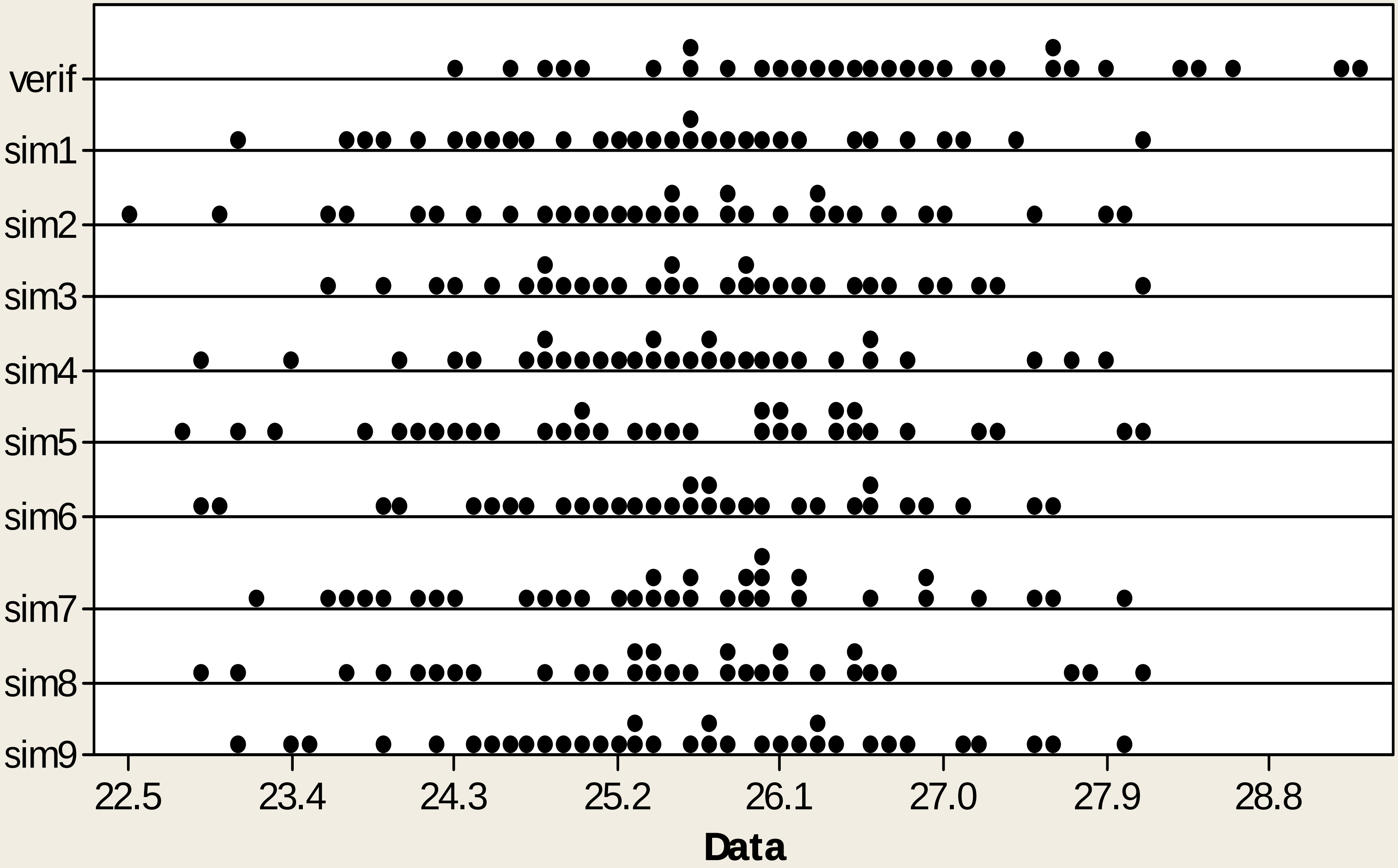
- Niño 3-4 1958-2001. Data + 9 hindcasts produced by a ECMWF coupled ocean-atmosphere climate model with slightly different initial conditions for each of the 9 members of this ensemble



An initial look at the data

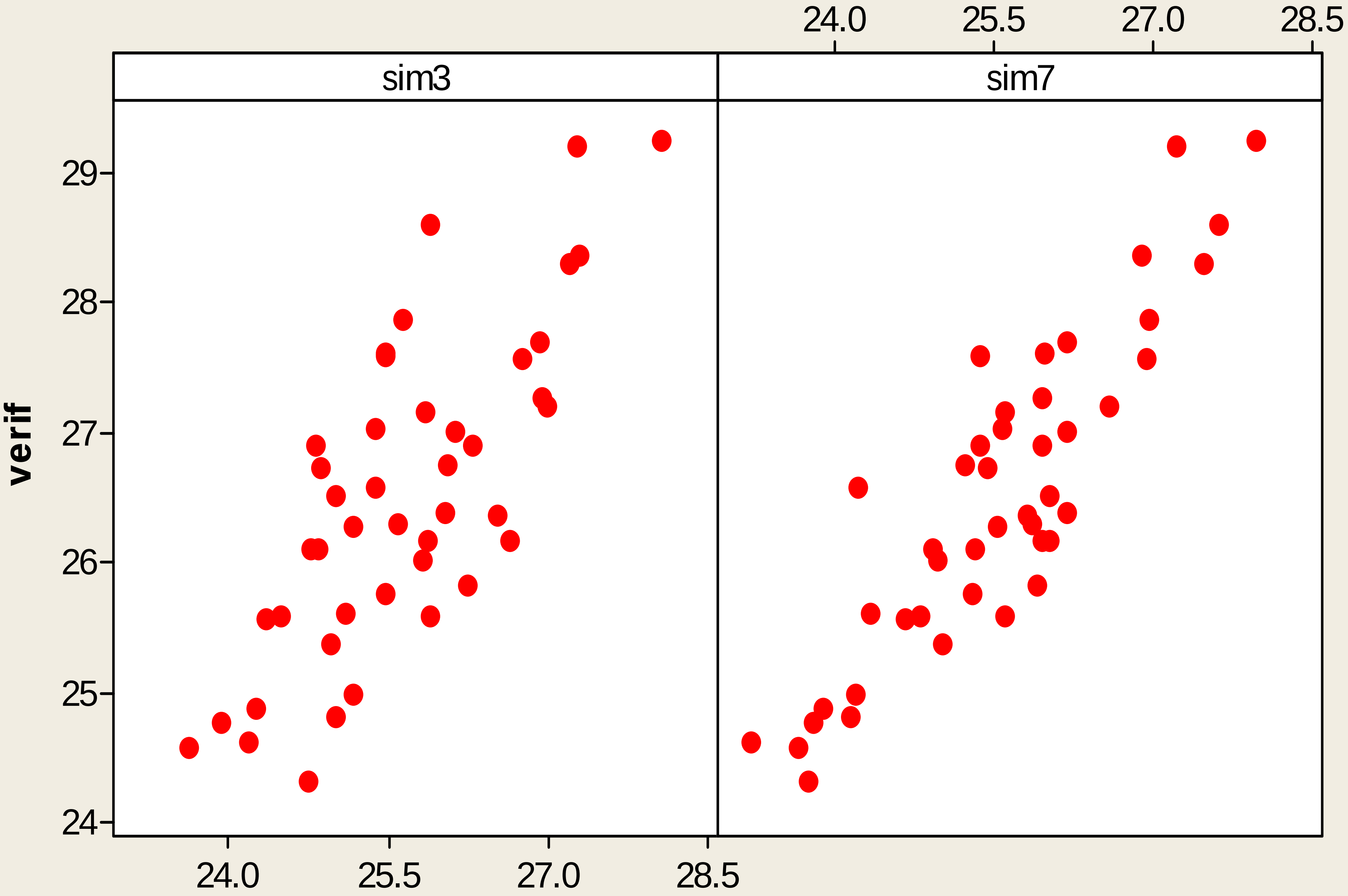
- We can look at the data in a number of ways, with a large number of possible verification measures – for illustration consider
  - Binary (value above/below mean); use hit rate as a verification measure
  - Continuous; use correlation as a verification measure
- The next three slides show
  - Dotplots – there is little to distinguish the 9 hindcasts, but they all seem to be too small on average (biased) relative to the actual values
  - Scatterplots of the data against the worst and best hindcasts with respect to correlation (r = 0.767, 0.891)
  - Data tabulated according to whether they are above or below average for the two hindcasts above

Dotplot of verif, sim1, sim2, sim3, sim4, sim5, sim6, sim7, sim8, sim9



Each symbol represents up to 2 observations.

Scatterplot of verif vs sim3, sim7



Binary data for two hindcasts (Hit rate 0.667, 0.762)

		Verif	
		Below	Above
Sim3	Below	16	7
	Above	8	13
Sim7	Below	16	7
	Above	5	16

Confidence interval for hit rate

- Like several other verification measures, hit rate is the proportion of times that something occurs – in this case the proportion of occurrences of the event of interest that were forecast. Denote such a proportion by p.
- A confidence interval can be found for the underlying probability of a correct hindcast, given that the event occurred. Call this probability π.
- The situation is the standard one of finding a confidence interval for the 'probability of success' in a binomial distribution, and there are various ways of tackling this.

Binomial confidence intervals

- Common approximation, based on the fact that the distribution of p can be approximated by a Gaussian distribution with mean π and variance p(1-p)/n where n is the 'number of trials'. The interval has endpoints $p \pm z_{\alpha/2}\sqrt{}p(1\text{-}p)/n$, where $z_{\alpha/2} = 1.96$ for a 95% interval.
- A slightly better approximation is based on the fact that the distribution of p can be approximated by a Gaussian distribution with mean π and variance π(1- π)/n. Its endpoints are given by the roots of a quadratic equation.

Binomial confidence intervals II

- For small n we can find an interval based on the binomial distribution itself rather than a normal approximation. Such intervals are sometimes called ‘exact’, though their coverage probability is generally not exactly that specified, because of the discreteness of the distribution. Details are not given, but charts are available for finding such intervals.
- Bayesian intervals assume some prior knowledge about π. Such intervals are a different sort of animal – they assume that π is random, not fixed, and use percentiles from its posterior probability distribution.
- Bootstrap intervals – illustrated later for the correlation coefficient.

Binomial confidence intervals - example

- Consider hindcast Sim3 for which $p = 16/24 = 2/3 = 0.67$. Find 95% confidence interval for π
- Usual approximation (0.48,0.86)
- Improved approximation (0.47,0.82)
- ‘Exact’ interval (0.47,0.84)
- Bayesian interval based on uniform prior distribution (0.47,0.83)
- These demonstrate that n=24 is large enough for the approximations to work well

Confidence intervals for differences

- Suppose we have two forecasts and we wish to compare their hit rates by finding a confidence interval for the difference between the two underlying parameters π_1-π_2.
- In the present example it is pretty clear that, because of the small sample sizes, any interval will be very wide.
- As an illustration suppose we have the observed hit rates, p_1=0.762, p_2=0.667 but based on much larger samples n_1=n_2=200. Find a 95% confidence interval for π_1-π_2.

Confidence intervals for differences II

- Large samples so approximation is fine.
- Interval has endpoints p_1-p_2±1.96√$p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$ (everything to the right of √ is square-rooted).
- Substituting gives 0.095±0.090, so interval is (0.005,0.185). This does not include zero, implying that π_1,π_2 are different.
- Note that 95% intervals for π_1, π_2 are (0.703,0.821), (0.602,0.732) respectively. These overlap, suggesting that π_1, π_2 may not be different.
- In comparing parameters it is usually more appropriate to find a confidence interval for the difference than to compare individual intervals.
- Note that the interval above assumes independence of p_1, p_2. If they were positively correlated, the interval would be narrower.

Simultaneous intervals for more than one measure – ROC curves

- If two measures have a known joint distribution, then the distribution could be inverted to find a confidence region for the two corresponding population measures
- The idea seems not to have been much developed in verification
- One situation where it might be useful is for ROC curves where hit rate (one measure) is plotted against false alarm rate (another measure) for several thresholds (so further multiplicity of measures)

ROC curves II

- Pepe (2002), 'The statistical evaluation of medical tests for classification and prediction' has
  - confidence rectangles
  - confidence intervals for hit rate, given false alarm rate
- More could be done
  - rectangles ignore dependence (though hit rate and false alarm rate are statistically independent for a given threshold, so rectangles are OK in this context)
  - if intervals are given for several thresholds, confidence coefficients need adjustment (corresponding to the multiple testing problem in hypothesis testing)

Confidence intervals for Pearson's correlation coefficient

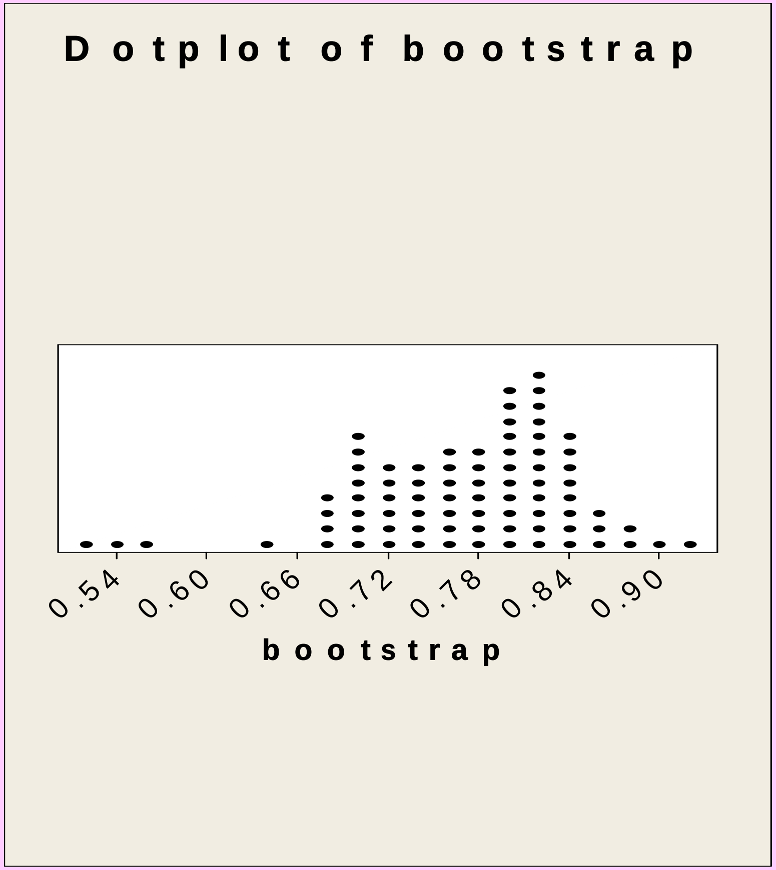
- We have r, a sample value. We want a confidence interval for ρ, the corresponding population quantity.
- There are various approximations
  - Interval with endpoints $r \pm z_{\alpha/2}(1-r^2)/\sqrt{n}$. A 95% interval is (0.645,0.889) for r=0.767 and n=44.
  - Based on Fisher's z-transformation, $\frac{1}{2}\ln[(1+r)/(1-r)]$ is approximately normally distributed with mean $\frac{1}{2}\ln[(1+\rho)/(1-\rho)]$ and variance 1/(n-3). A 95% interval for r is (0.609,0.867) for r=0.767 and n=44.

Bootstrap confidence intervals for correlation coefficients

- Take B random samples of size n with replacement from the paired data and calculate r for each sample. Rank the r values. For a confidence coefficient (1-2α) find the Bαth smallest and Bαth largest of the r values. Call these l and u.
  - The percentile method uses the interval (l, u).
  - The ‘basic bootstrap’ uses (r–(u-r), r+(r-l)).
  - There are various other ‘improved’ bootstrap confidence intervals.

Bootstrap example

- r = 0.767, n = 44, B = 80; 95% interval required. l=0.540 (2nd smallest); u=0.904 (2nd largest).
- Percentile interval (0.540, 0.904).
- 'Basic' interval (0.630, 0.994).
- 'Basic' interval using Fisher's transformation (0.487, 0.890).
- Note the outliers at left of bootstrap distribution.



Confidence intervals for correlation coefficients - discussion

- We have 5 intervals. Unlike the hit rate intervals they are quite different –which is to be preferred?
- Normal (0.65,0.89); Fisher (0.61,0.88); Percentile B (0.54,0.90); Basic B (0.63,0.99); Fisher Basic B (0.49,0.89).
- No best buy – you can't place too much faith in the exact coverage of an interval without strong belief in the assumptions underlying it.
- But … Fisher should usually be a better bet than normal; 'Basic' needs assumptions which are better satisfied after Fisher transformation.
- Whether the bootstrap or usual Fisher interval is preferred depends on reaction to the outlying bootstrap samples – do these reflect the distribution of r in the underlying population or are they simply reflecting odd behaviour in our particular data set?

Prediction intervals

- A prediction interval (or probability interval) is an interval with a given probability of containing the value of a random variable, rather than a parameter
- The random variable is random and the endpoints are fixed points in its distribution, whereas the interval is random for a confidence interval
- Prediction intervals can also be useful in quantifying uncertainty for verification measures

Prediction intervals for correlation coefficients

- We need the distribution of r, usually calculated under some null hypothesis, the obvious one being that ρ =0. Using the crudest approximation, r has a Gaussian distribution with mean zero, variance 1/n.
- If a confidence interval for ρ doesn't include zero we conclude that there is a (linear) relationship between forecast and data.
- Prediction intervals provide a dual way of tackling the same question, by ascertaining whether the prediction interval includes the observed value of r.

Prediction intervals for correlation coefficients - example

- Using the crude approximation, a 95% prediction interval for r, given ρ=0, has endpoints ± $1.96\sqrt{1/n}$.
- With n=44, these are ± 0.295. All the observed values of r are well outside this interval, indicating a relationship between hindcasts and data – the hindcasts have skill.

Hypothesis testing

- The interest in uncertainty associated with a verification measure is often of the form
  - Is the observed value compatible with what might have been observed if the forecast system had no skill?
  - Given two values of a measure for two different forecasting systems (or the same system at different times), could the difference have arisen by chance if there was no difference in underlying skill for the two systems (the two times)?

Hypothesis testing II

- Such questions can clearly be answered with a formal test of the null hypothesis of ‘no skill’ in the first case, or ‘equal skill’ in the second case
- A test of hypothesis is often equivalent to a confidence interval and/or prediction interval

Correlation coefficient - test of ρ=0

- Continuing our example with 0.767, n=44.
- Null hypothesis H_0: ρ=0
  - Use the crude approximation that, under $H_{0,}$ r has a Gaussian distribution with mean zero, variance 1/n. Then reject H_0 at the 5% significance level (atmospheric scientists but hardly anyone else may refer to this as 95%) if and only if r is larger than $1.96\sqrt{}1/n$ or less than $-1.96\sqrt{}1/n$; in other words, if and only if r is outside the (1-α) prediction interval for r found earlier. Clearly H_0 is rejected at the 5% level.

Correlation coefficient - test of $\rho=0$ via confidence intervals

- We could also use any of our earlier confidence intervals to test H_0. We gave 95% intervals, and would reject H_0 at the 5% level if and only if the interval fails to include zero, which it does in all cases.
- If the intervals were 99%, the test would be at the 1% level, and so on. Similarly for prediction intervals.

Permutation and randomisation tests of $\rho=0$

- If we are not prepared to make assumptions about the distribution of r, we can use a permutation approach:
  - Denote the forecasts and observed data by (f_i, o_i), i =1, …n.
  - Fix the f_is, and consider all possible permutations of the o_is.
  - Calculate the correlation between the f_is and permuted o_is in each case.
  - Under H_0 all permutations are equally likely, and the p-value for a permutation test is the proportion of all calculated correlations greater than or equal to (in absolute value for a two-sided test) the observed value.
- The number of permutations may be too large to evaluate them all. Using a random subset of them instead gives a randomisation test.

Conclusions

- It is important to quantify the uncertainty associated with computed values of verification measures
- Standard errors, confidence intervals, prediction intervals, tests of hypotheses, can all be used to do so
- Which to use, and which variety, depends on the context and on the assumptions that can be safely made