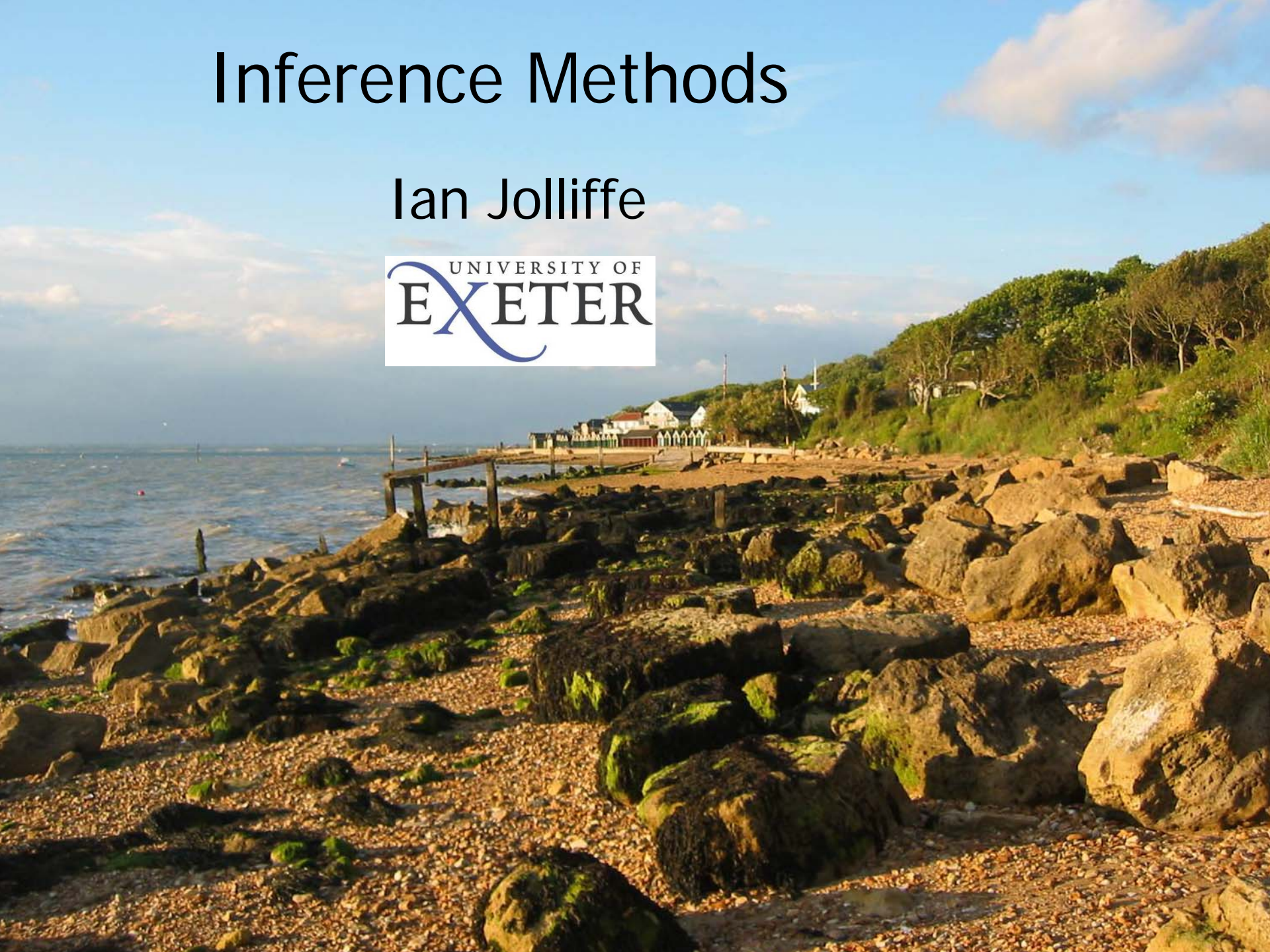


Inference Methods

Ian Jolliffe



Outline

- Introduction
- Interval Estimation
 - Confidence Intervals
 - Bayes Intervals
 - Bootstrap Intervals
 - Prediction Intervals
- Hypothesis Testing
 - Links between intervals and tests
 - P-values
 - Permutation tests
- Complications
 - Non-Gaussian distributions
 - Errors in observations
 - Spatial and temporal correlations
 - Simultaneous inference and multiple testing

Introduction

- Statistical inference is needed in many circumstances. Here the context is that we have a value of a verification measure and wish to account for the uncertainty associated with that measure.
- The emphasis here is on interval estimation.
- The presentation draws heavily on Jolliffe (2007) – some of the results are slightly different.

Types of inference

- Point estimation – e.g. simply give the value of a verification measure, with no indication of the uncertainty associated with it.
- Interval estimation - a standard error could be attached to a point estimate, but it is better to go one step further and construct a confidence interval, especially if the distribution of the measure is not close to Gaussian.
- Hypothesis testing - in comparing values of a measure at different times, hypothesis testing may be a good way of addressing the question of whether any improvement could have arisen by chance.

Approaches to inference (Garthwaite et al., 2002)

1. Classical (frequentist) parametric inference.
2. Bayesian inference.
3. Non-parametric inference.
4. Decision theory.
5. ...

Note that

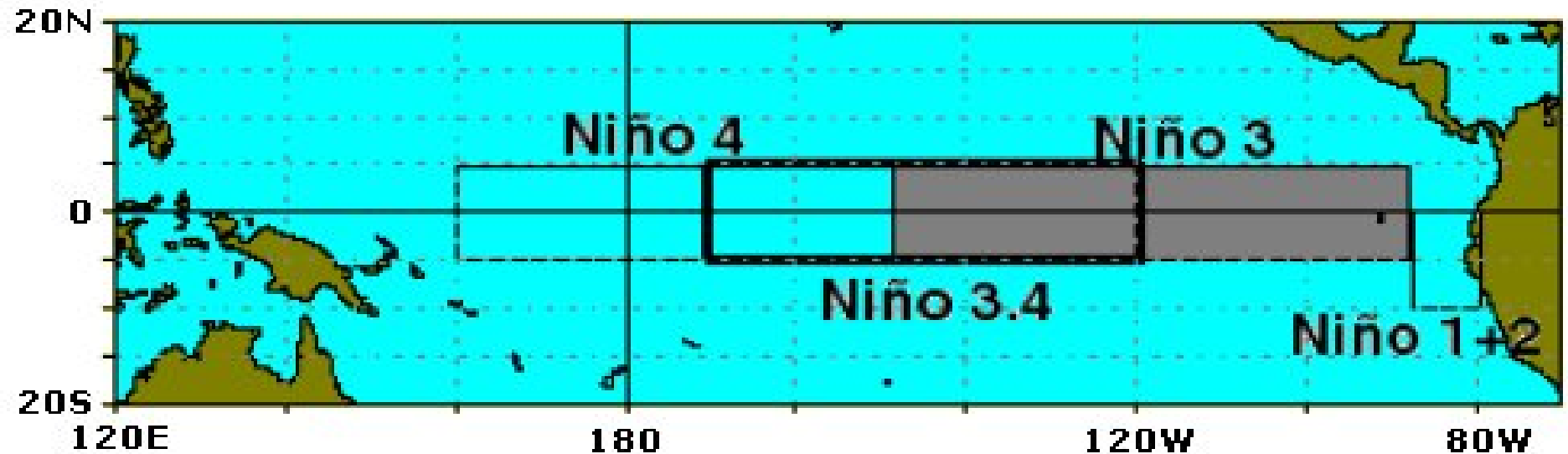
- The likelihood function is central to both 1 and 2.
- Computationally expensive techniques are of increasing importance in both 2 and 3.

Confidence intervals

- Given a sample value of a measure (statistic) 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

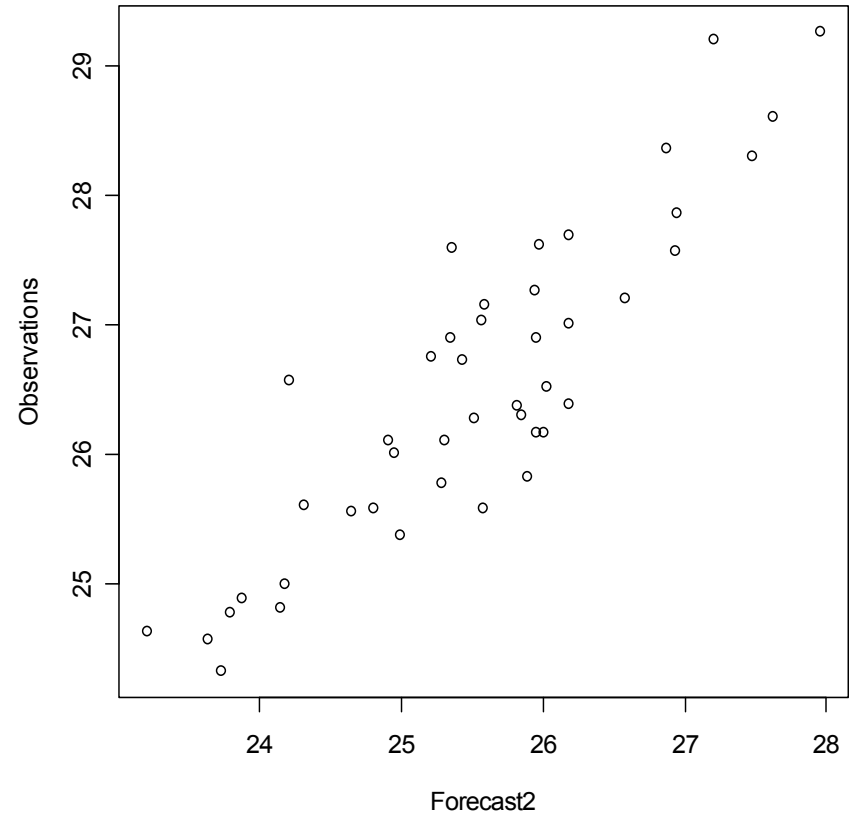
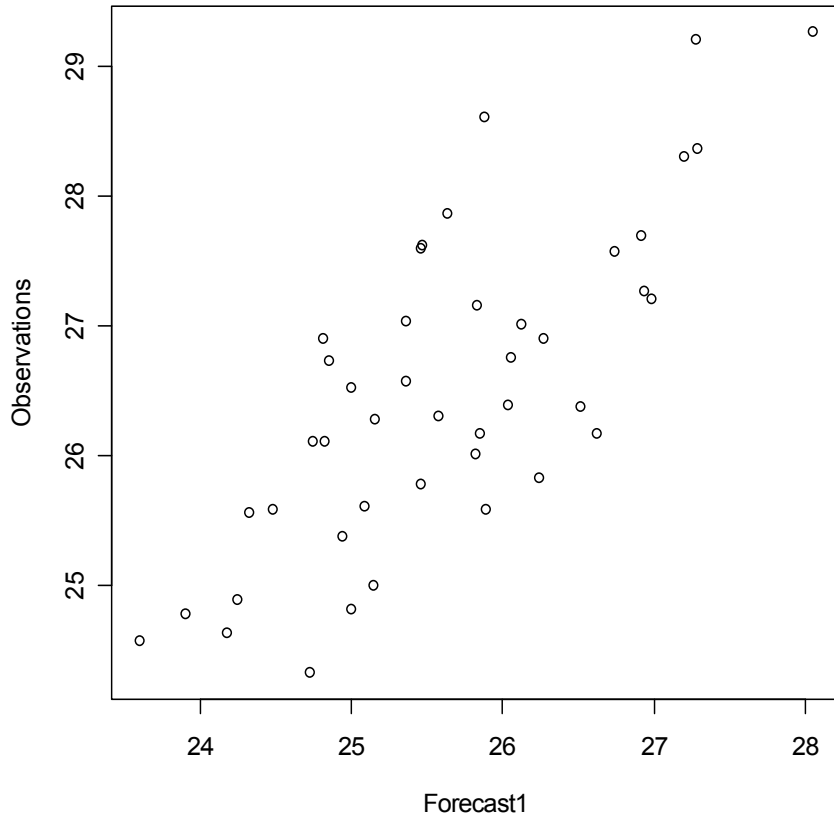
- Niño 3-4 SST for 44 years 1958-2001. Data + 9 hindcasts produced by a ECMWF coupled ocean-atmosphere climate model with slightly different initial conditions for each of the 9 members of this ensemble (data from Caio Coelho).
- 9 time series, which we refer to as ‘forecasts’, are constructed from the ensemble members and compared with observed data.



Verification measures

- 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 (probability of detection) as a verification measure.
 - Continuous: use correlation as a verification measure.
- The next two slides show
 - Scatterplots of the data against two of the forecasts (labelled Forecast 1, Forecast 2) with correlations $r = 0.767, 0.891$.
 - Data tabulated according to whether they are above or below average, for two forecasts labelled Forecast 1, Forecast 3 with hit rates 0.619, 0.905.

Two scatterplots: $r = 0.767, 0.891$



Binary data for two forecasts (Hit rates 0.619, 0.905)

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

Confidence intervals 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 forecast, 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

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

$$\frac{p + z_{\alpha/2}^2 / 2n \pm z_{\alpha/2} \sqrt{p(1-p)/n + z_{\alpha/2}^2 / 4n^2}}{1 + z_{\alpha/2}^2 / n}$$

Binomial confidence intervals II

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 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 and there is a function in R for doing so.

Bayesian intervals

- In the Bayesian approach to inference, a prior distribution for the parameter of interest (here π) is combined with the likelihood function for the data to give a posterior distribution for π (Epstein, 1985).
- Bayesian intervals are a different sort of interval from confidence intervals – they assume that π is random, not fixed, and use percentiles from its posterior probability distribution.

Bayesian intervals for a binomial parameter

- The obvious type of prior distribution for π is a Beta distribution. Such distributions are:
 - Defined on the range $[0,1]$, like π ;
 - Reasonably flexible in their shape;
 - Conjugate – a Beta prior implies a Beta posterior.
- The pdf for a Beta distribution with parameters α and β is

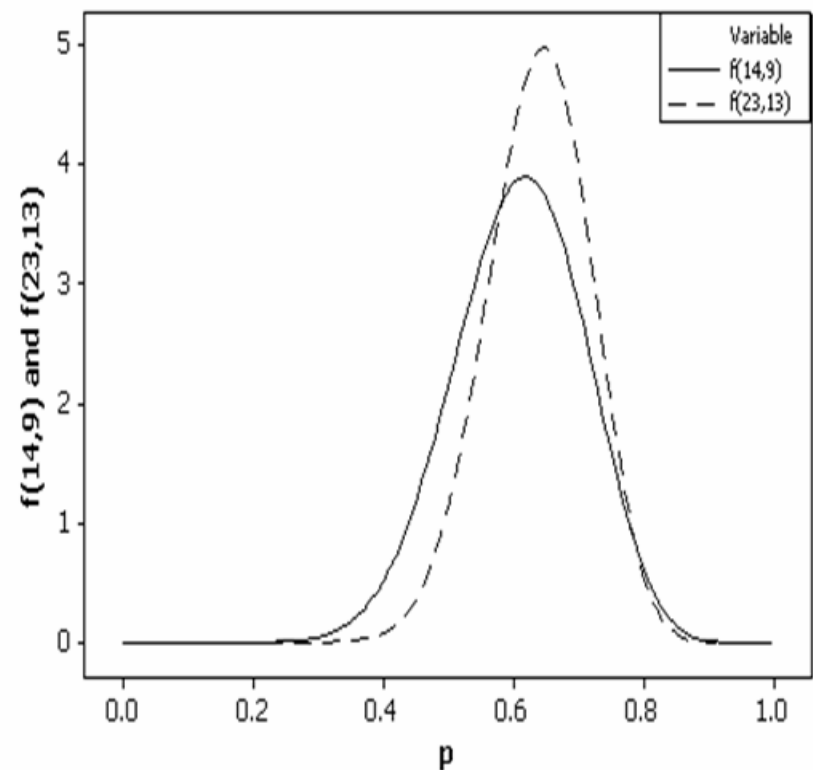
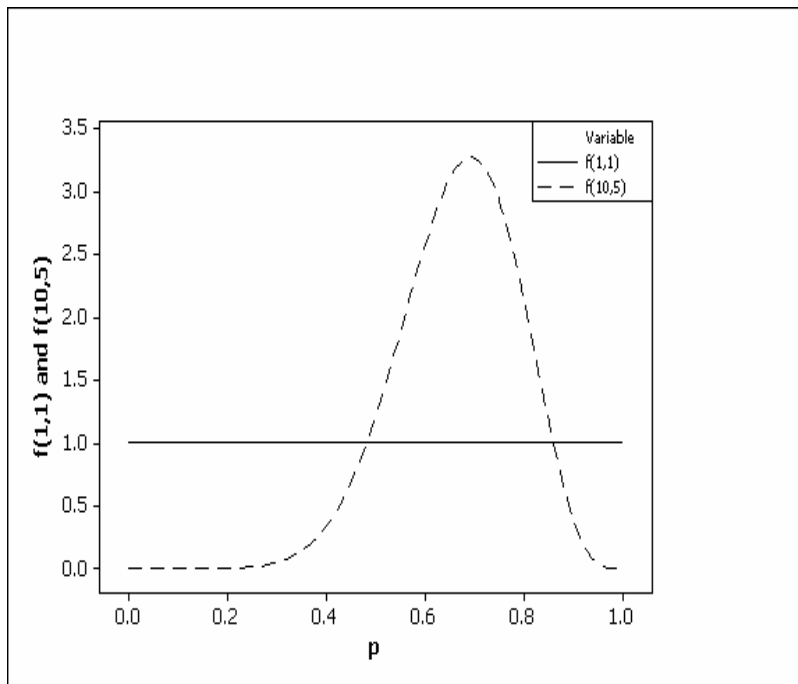
$$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1} (1 - \pi)^{\beta-1}$$

The likelihood function (simply the binomial probability function for x successes in n trials) is

$$\frac{n!}{x!(n-x)!} \pi^x (1 - \pi)^{(n-x)}$$

Multiplying these leads a Beta posterior with parameters $(\alpha+x)$, $(\beta+n-x)$.

Two Beta prior (left) and corresponding posterior (right) distributions for Forecast 1



Bootstrap intervals (Efron & Tibshirani, 1993)

- The data set for Forecast 1 consists of 13 successes (1's) and 8 failures (0's).
- Take B random samples of size 21 **with replacement** from these 21 values and calculate p for each sample.
- Rank the B values of p . For a confidence coefficient $(1-2\alpha)$ find the $B\alpha^{\text{th}}$ smallest and $B\alpha^{\text{th}}$ 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.
- Results are given for $B = 1000$.

Binomial example – some comments

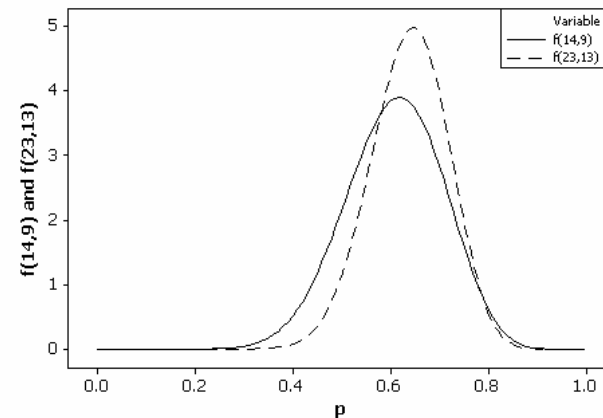
	Forecast 1	Forecast 3
Crude approx.	(0.41,0.83)	(0.78,1.03)
Better Approx.	(0.41,0.79)	(0.71,0.97)
'Exact'	(0.38,0.82)	(0.70,0.99)
Bayes – uniform	(0.41,0.79)	(0.71,0.97)
Bayes – informative	(0.48,0.79)	(0.66,0.92)
Percentile bootstrap	(0.43,0.81)	(0.76,1.00)
Basic bootstrap	(0.43,0.81)	(0.81,1.05)

- There is very little difference between the intervals for Forecast 1 ($p = 13/21$). This demonstrates that $n=21$ is large enough, and p far enough from 0 or 1, for the Gaussian approximations to work reasonably well. There are larger discrepancies for Forecast 3, where $p = 19/21$ is closer to 1.
- For Forecast 3 the upper limit exceeds 1 for the crude approximation and the basic bootstrap, which is unsatisfactory.
- The 'exact' interval is wider than any of the others, but this may be because its confidence coefficient is greater than 95%.

Binomial example – more comments

	Forecast 1	Forecast 3
Crude approx.	(0.41,0.83)	(0.78,1.03)
Better Approx.	(0.41,0.79)	(0.71,0.97)
'Exact'	(0.38,0.82)	(0.70,0.99)
Bayes – uniform	(0.41,0.79)	(0.71,0.97)
Bayes – informative	(0.48,0.79)	(0.66,0.92)
Percentile bootstrap	(0.43,0.81)	(0.76,1.00)
Basic bootstrap	(0.43,0.81)	(0.81,1.05)

The informative prior has mean $2/3$. The corresponding Bayes interval is narrower and shifted upwards compared to that for the uniform prior for Forecast 1, and shifted downwards for Forecast 3.



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 $\pi_1 - \pi_2$.
- In the present example it is pretty clear that, because of the small sample sizes, any interval will be very wide.
- However, as an illustration we find an approximate 95% confidence interval for $\pi_1 - \pi_2$ for our current data, with $p_1 = 13/21$, $p_2 = 19/21$.

Confidence intervals for differences - example

An approximate 95% interval has endpoints

$$(p_1 - p_2) \pm 1.96 \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$$

- Substituting gives -0.29 ± 0.24 , so interval is $(-0.53, -0.05)$. This does not include zero, implying that π_1, π_2 are different.
- This interval is based on the crude approximation. However the percentile bootstrap gives a very similar interval $(-0.52, -0.05)$.
- Note that all the pairs of individual 95% intervals for π_1, π_2 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. Looking at overlap of intervals is often misleading.
- Note that the interval above assumes independence of p_1, p_2 . If they were positively correlated, the interval would be narrower. Bootstrapping can incorporate pairing between forecasts and gives a percentile interval $(-0.48, -0.10)$.

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 $r \pm z_{\alpha/2}(1-r^2)/\sqrt{n}$.
 - 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 $1/(n-3)$.
- Bayesian and bootstrap approaches could also be used.

Confidence intervals for correlation coefficients - example

	Forecast 1	Forecast 2
Normal approximation	(0.65,0.89)	(0.83,0.95)
Fisher's transformation	(0.61,0.87)	(0.81,0.94)
Percentile bootstrap	(0.61,0.87)	(0.80,0.95)
Basic bootstrap	(0.67,0.92)	(0.83,0.98)

- There is very little difference between these intervals.
- In general, the second should give a better approximation than the first.
- Bootstrap will be preferred if there is doubt about distributional assumptions.

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 interval's endpoints are fixed points in its distribution, whereas for a confidence interval the endpoints are random and the quantity being covered is fixed but unknown.
- 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 $\rho = 0$. Using the crudest approximation, r has a Gaussian distribution with mean zero, variance $1/n$ and a 95% prediction interval for r , given $\rho = 0$, has endpoints $0 \pm 1.96\sqrt{1/n}$.
- Our example has $n=44$, so a 95% prediction interval is $(-0.295, 0.295)$.
- **Prediction interval:** given $\rho = 0$ we are 95% confident that r lies in the interval $(-0.295, 0.295)$.
- **Confidence interval:** given $r = 0.767$, we are 95% confident that the interval $(0.61, 0.87)$ contains ρ .

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 in values have arisen by chance if there was no difference in underlying skill for the two systems/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 $\rho=0$

- Continue our example with $r = 0.767$, $n=44$ and null hypothesis $H_0: \rho=0$.
- 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 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 $(1-\alpha)$ prediction interval $(-0.295, 0.295)$ for r found earlier.
- Clearly H_0 is rejected at the 5% level or, indeed, much more stringent levels.

* atmospheric scientists, but hardly anyone else, sometimes refer to this as 95%

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.

P-values

- Hypothesis tests can be treated as a clear-cut decision process – decide on a significance level (5%, 1%) and derive a critical region (a subset of the possible data) for which some null hypothesis (H_0) will be rejected.
- Alternatively a p-value can be quoted. This is the probability that the data, or something less compatible with H_0 , could have arisen by chance if H_0 was true.
- IT IS NOT the probability that H_0 is true.
- The latter can be found via a Bayesian approach.
- For more see Jolliffe (2004).

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) , $i = 1, \dots, n$.
 - Fix the f_i s, and consider all possible permutations of the o_i s.
 - Calculate the correlation between the f_i s and permuted o_i s 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, though the terms permutation test and randomisation test are often used synonymously.

Complications – non-Gaussian measures

- The examples above illustrate that measures may have approximately Gaussian distributions, even for quite small samples, but this cannot be assumed.
- If the measures are non-Gaussian there are several options, some of which have been illustrated ...

Complications – non-Gaussian measures

Can be tackled by the following

- Transform to approximate Gaussianity.
- Use an exact alternative distribution based on known theory.
- Use a computationally intensive non-parametric procedure such as bootstrap or permutation test.
- Use a more traditional non-parametric method.

If there are insufficient past data on values of the measure to roughly judge its distribution, use a non-parametric procedure.

Errors in observations

- To be addressed tomorrow ...
- In general observation error seems to make verification measures appear worse than they really are.
- Errors can be adjusted for, but knowledge of the error structure is needed.
- Their effect should be less important when measures are compared from different times/systems ... but is it better not to try to reduce observation errors, so as not to complicate such comparisons?

Spatial and temporal dependence

- Much inference on verification measures assumes independence of the data points from which the measure is calculated – this is often invalid.
- Solutions
 - Take one or more subsets of the data so that data within subgroups are independent (data reduction)
 - Model the dependency (parametric approach)
 - Allow for dependency in a non-parametric approach (e.g block bootstrap)

Simultaneous inference

- Ideas for inference about one measure or parameter can be extended to two or more parameters.
- For differences or ratios of parameters we essentially reduce things back to a single derived parameter.
- A simple example of two parameters is hit rate and false alarm rate, corresponding to points on a ROC curve. Pepe (2003) gives confidence rectangles for the two; also confidence intervals for hit rate, given false alarm rate.
- If intervals are given for several parameters, confidence coefficients may need adjustment (corresponding to the multiple testing problem in hypothesis testing).

Multiple testing

- Field significance is one aspect of this (Livezey and Chen, 1983). It also has the feature of spatial dependence.
- Traditionally try to control the family-wise error rate (the probability of wrongly rejecting at least one of the null hypotheses) by reducing the significance level of each individual hypothesis test e.g. Bonferonni adjustments.
- More recently controlling the false discovery rate (FDR) = 'proportion of rejected null hypotheses that are actually true' has attracted a lot of attention (Ventura et al, 2004).
- Multiple testing, especially approaches based on the FDR, is a 'hot' topic in statistics, driven largely by microarray data from genetics.

Final remarks

- It is important to quantify the uncertainty associated with computed values of verification measures.
- Standard errors, confidence intervals, Bayes intervals, bootstrap 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.
- There are a number of complications that are, at least in part, specific to verification measures.

Questions, comments, discussion?

i.t.jolliffe@ex.ac.uk



References

- Efron B and Tibshirani RJ (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- Epstein ES (1985). *Statistical Inference and Prediction in Climatology: A Bayesian Approach*. Meteorological Monograph. American Meteorological Society.
- Garthwaite PH, Jolliffe IT & Jones B (2002). *Statistical Inference, 2nd edition*. Oxford University Press.
- Jolliffe IT (2004) P stands for ... *Weather*, **59**,77-79.
- Jolliffe IT (2007). Uncertainty and inference for verification measures. *Wea. Forecasting*, **22**, 637-650.
- Livezey RE & Chen WY (1983). Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111**, 46-59.
- Ventura V, Paciorek CJ & Risbey JS (2004). Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *J. Climate*, **17**, 4343-4356.