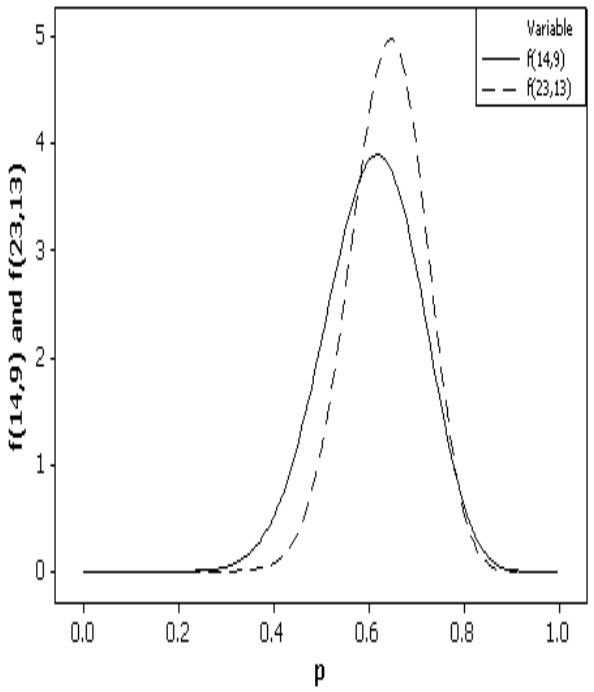
Statistical Inference

Confidence Intervals (and Hypothesis Tests)

Ian Jolliffe



Introduction
- Types of Inference
- Illustrative Examples

Interval Estimation
- Confidence Intervals
- Bayes Intervals
- Bootstrap Intervals
- [Prediction Intervals]

Hypothesis Testing

Links between intervals and tests

Introduction

- Statistical inference is needed in many circumstances, not least in forecast verification.
- We explain the basic ideas of statistical inference (some old, some newer), some of which are often misunderstood.
- A simple example is used to illustrate the ideas

Inference – the framework

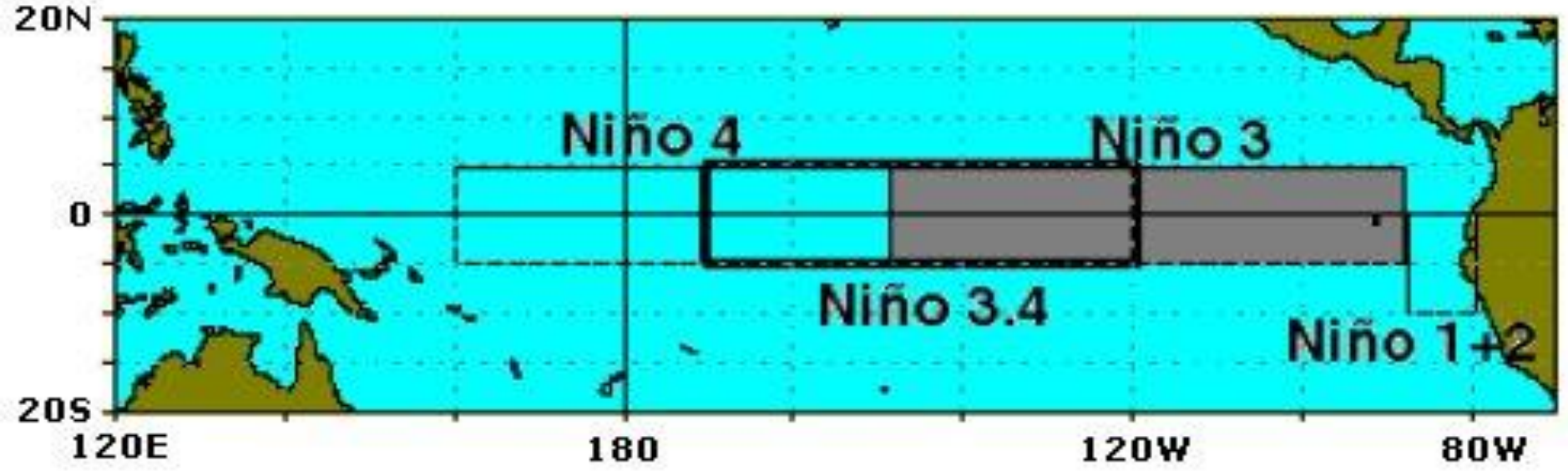
- We have data that are considered to be a sample from some larger population, for example:
  - A real population of people
  - A theoretical population of all possible forecasts and observations of next-day maximum temperature
- We wish to use the data to make inferences about some population quantities (parameters),
  - Proportion of population with antibodies
  - Correlation between forecast and observed temperature

Types of inference

- Point estimation – e.g. simply give a single number to estimate the parameter, with no indication of the uncertainty associated with it – the observed proportion with antibodies, the sample correlation or some other verification measure
- 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 e.g. an interval for which we have a specified level of confidence that it includes the underlying 'population' value of antibody proportion, correlation
- Hypothesis testing - in comparing estimates of a parameter (e.g. correlation for two different forecasting techniques), hypothesis testing can sometimes be a good way of addressing the question of whether the difference (in correlation) for the two forecasting techniques could have arisen by chance.
- Our emphasis is on interval estimation.

Example

- Niño 3-4 SST1958-2001. 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, for the 44 years 1958-2001, which we refer to as ‘forecasts’, are constructed from the ensemble members and compared with observed data.



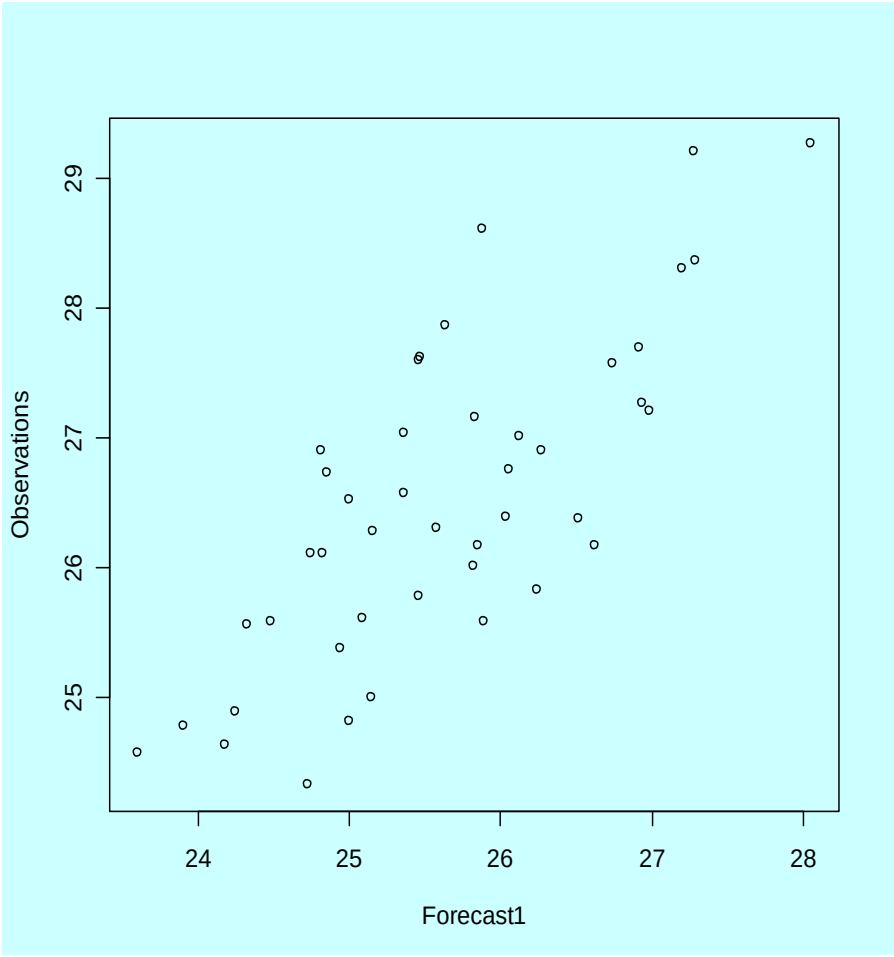
Verification measures and uncertainty

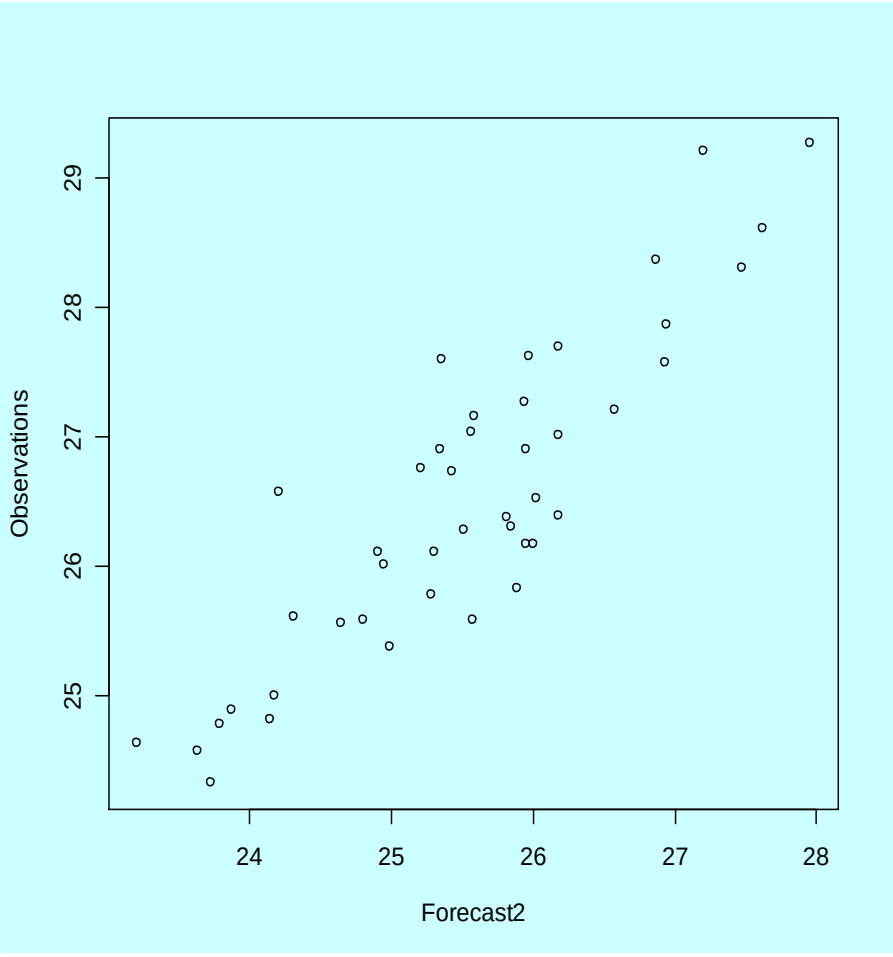
- We could compare the ‘forecasts’ with the observations in a number of ways – for illustration consider
  - For a given forecast series, compare the 44 forecast SSTs with their corresponding observed values using the correlation coefficient
  - Convert to binary data (is the SST above or below the mean?): use probability of detection (POD) (aka hit rate aka sensitivity in disease detection) as a verification measure.
- The values of these measures that we calculate have uncertainty associated with them – if we had a different set of forecasts and observations for Niño 3-4 SST, we would get different values.
- Assume that the data we have are a sample from some (hypothetical?) population and we wish to make inferences about the correlation and POD in that population.

Example - summary

- The next two slides show
  - Scatterplots of the observations against two of the forecasts (for convenience labelled Forecast 1, Forecast 2) with the lowest and highest correlations of the nine ‘forecasts’: r = 0.767, 0.891.
  - Data tabulated according to whether they are above or below average, for two forecasts labelled Forecast 1, Forecast 2 with lowest and highest POD 0.619, 0.905.
  - The variation in values between these forecasts illustrates the need for quantifying uncertainty.
- We will look at various ways of making inferences based on these correlations and POD.

Two scatterplots: r = 0.767,0.891





Binary data for two forecasts (POD 0.619, 0.905)

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

Approaches to inference

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

Note that

- The likelihood function (the joint probability distribution of the data viewed as a function of unknown parameters of interest) is central to both 1 and 2.
- Computationally expensive techniques are of increasing importance and accessibility in both 2 and 3.

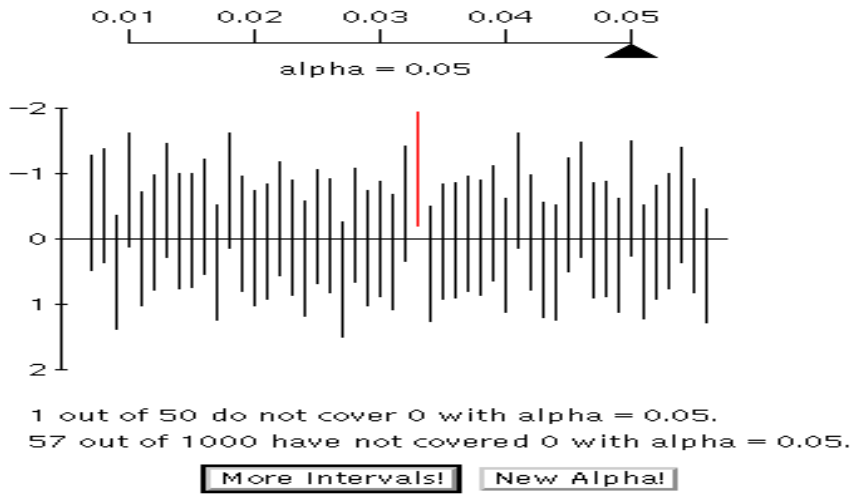
Interval estimation

What is

- A confidence interval?
- A Bayes or credible interval?
- An interval obtained by bootstrapping? (an important type of non-parametric interval)
- [A prediction or probability interval?]

What is a confidence interval?

- 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 (parameter).



Note:

- The interval is random; the population value is fixed – see diagram
- The confidence level is the long-run probability that intervals include the parameter, NOT the probability that the parameter is in the interval

Confidence intervals for POD

- Like several other verification measures, POD is the proportion of times that something occurs – in our sample we have the proportion of occurrences of above average SST that were forecast. Denote such a proportion by p.
- A confidence interval can be found for the underlying probability of a correct forecast (if you were to use the same forecasting procedure many many times), given that the event occurred. Call this probability θ.
- 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 θ 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 (2.58 for a 99% interval, etc.).
- A slightly better approximation is based on the fact that the distribution of p is better approximated by a Gaussian distribution with mean θ and variance θ(1-θ)/n. A little algebra shows that the endpoints of an interval based on this 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 example, for doing so.

Binomial (POD) example - 95% confidence intervals

- There is very little difference between the intervals for Forecast 1 (p = 0.62). This demonstrates that n=21 is large enough, and p far enough from 0 or 1, for the approximations to work reasonably well. There are larger discrepancies for Forecast 2, where p = 0.91 is closer to 1.
- For Forecast 2 the upper limit exceeds 1 for the crude approximation.
- The 'exact' interval is wider than any of the others, but this may be because its confidence coefficient is greater than 95%.

	Forecast 1	Forecast 2
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)

What is a Bayes interval? (also called a credible interval)

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

Bayes intervals for a binomial parameter

- The obvious type of prior distribution for θ is a Beta distribution. Such distributions are:
  - Defined on the range [0,1], like θ;
  - 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

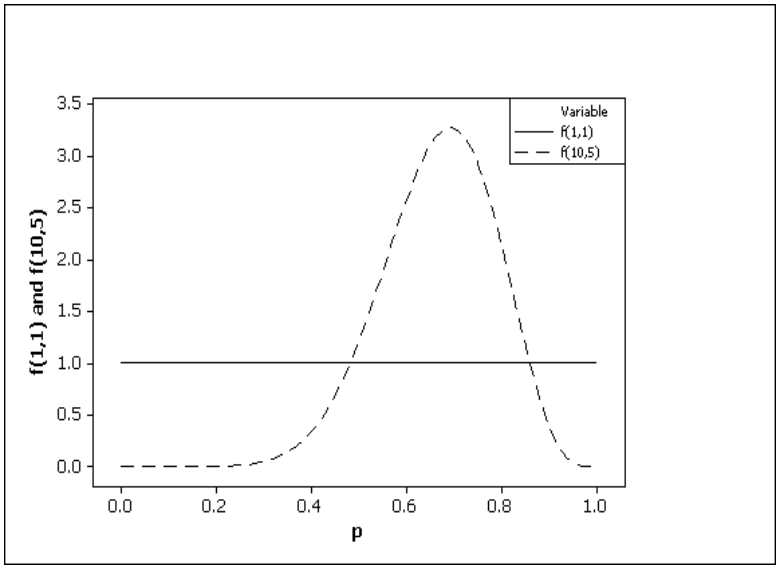
$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

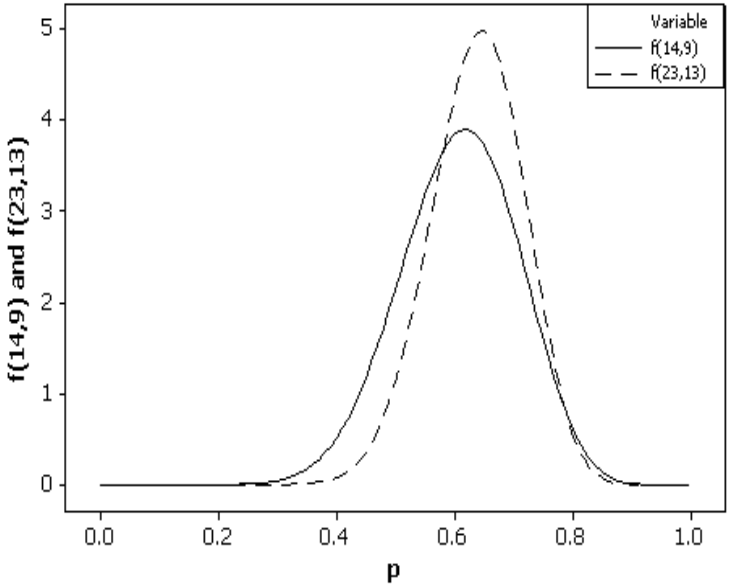
The likelihood function (simply the binomial probability function for x successes in n trials) is

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

Multiplying these leads a Beta posterior with parameters (α+x), (β+n-x).

Two Beta prior (left) and corresponding posterior (right) distributions - one informative and one uninformative - for Forecast 1





For a 95% Bayes interval, find values in the posterior distribution that cut off 2.5% probability in each tail. These then form the end-points of the interval. Similarly for other confidence levels.

Binomial (POD) example - 95% Bayes intervals

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

	Forecast 1	Forecast 2
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)

What is a bootstrap interval?

- 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 level $(1-2\alpha)$ find the $B\alpha^{th}$ smallest and $B\alpha^{th}$ largest of the r values. Call these l and u.
- There are various bootstrap confidence intervals of varying complexity. The easiest to understand and implement is the percentile method, which uses the interval (l, u).
- Results in the Table are for B = 1000.

Other bootstrap intervals

- There are other bootstrap intervals with greater complexity but improved properties, for example
  - Basic bootstrap
  - Parametric bootstrap
  - Bootstrap-t intervals
  - BC_{α}
- All share the idea of sampling with replacement.
- It is often the case that data are not independent (temporal and/or spatial correlation) Adaptations are then necessary e.g. block bootstrap for temporally correlated data in which you sample, with replacement, blocks of consecutive observations.

Binomial (POD) example - 95% percentile bootstrap intervals

- The 95% percentile bootstrap is not dissimilar to the other intervals, but is not reliant on the distributional assumptions of those intervals

	Forecast 1	Forecast 2
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)

Confidence intervals for differences

- Suppose we have two forecasts and we wish to compare their PODs by finding a confidence interval for the difference between the two underlying parameters θ_1-θ_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 θ_1-θ_2 for our current data, with $p_1 = 0.619$, $p_2 = 0.905$.

Confidence intervals for differences - example

An approximate 95% interval has endpoints

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

- This interval is based on our earlier crude approximation of the binomial distribution by a Gaussian.
- Substituting gives -0.29 $\pm$ 0.24, so the interval is (-0.53,-0.05). This does not include zero, implying, with 95% confidence, that θ_1, θ_2 are likely to be different

Percentile bootstrap interval for differences

- Take a random sample of size 21 with replacement from Forecast 1's 13 '1's and 8 '0's and calculate p_1 for the sample.
- Take a random sample of size 21 with replacement from Forecast 2's 19 '1's and 2 '0's and calculate p_2 for the sample.
- Calculate p_1-$p_{2,}$ and repeat these steps B times.
- Rank the B values of p_1-p_2. For a confidence level (1-2α) find the Bα^{th} smallest and Bα^{th} largest of these values. These are the endpoints of our interval.
- In the example, with B=1000 the interval is (-0.52, -0.05) very similar to the confidence interval

More on intervals for differences

- Returning to our earlier results, intervals for Forecasts 1 and 2 overlap, suggesting that θ_1, θ_2 may not be different, contradicting our conclusions based on differences
- 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.

	Forecast 1	Forecast 2
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)

More on intervals for differences II

- Note that the intervals above assume independence of p_1, p_2. If they are positively correlated, as seems likely, the interval would be narrower.
- To bootstrap incorporating pairing between forecasts, take a random sample with replacement of size 44 from the triples (Forecast 1, Forecast 2, Observed) for the 44 years. Calculate p_1-p_2, and repeat B times.
- Rank the B values of p_1-p_2. For a confidence level (1-2α) find the Bα^{th} smallest and Bα^{th} largest of the B values. These are the endpoints of our interval.
- In the example, with B=1000 the 95% percentile interval is (-0.48,-0.10), narrower than the unpaired intervals, as expected

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, $\frac{1}{2}\log_e[(1+r)/(1-r)]$ is approximately normally distributed with mean $\frac{1}{2}\log_e[(1+\rho)/(1-\rho)]$ 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)

- 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.

What is a prediction interval?

- 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 the interval is random for a confidence interval.
- Prediction intervals, as well as confidence intervals, can be useful in quantifying uncertainty when estimating parameters.

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 ρ = 0 we are 95% confident that r will lie 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 (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 $\rho=0$

- 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 $1.96\sqrt{1/n}$ (0.295) or less than $-1.96\sqrt{1/n}$ (-0.295); in other words, if and only if r is outside the 95% 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.

Decision theory and p-values

- Hypothesis tests can be treated as a clear-cut decision process – choose a significance level (5%, 1%) and derive a critical region (a subset of the possible data) for which H_0 will be rejected.
- For a full decision theory approach, we also need a loss function and prior probabilities.
- 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.
- Because of the misuse of p-values, a Psychology journal banned them in 2015. This was followed by much discussion including the American Statistical Association issuing a *Statement on p-Values and Statistical Significance* and publishing a special issue of The American Statistician containing 43 articles on the subject.

What have we learned?

- When calculating a verification measure, there is (almost?) always uncertainty associated with the value of that measure.
- Statistical inference can help to quantify that uncertainty.
- Sometimes we may wish to test a specific hypothesis such as ‘are the forecasts better than chance?’ or ‘does a new forecasting system give better forecasts than an old one?’.
- More often, a confidence interval, or some other type of interval, is a more informative way of quantifying uncertainty.

What have we learned II

- We have seen several different types of ‘uncertainty interval’: confidence intervals, Bayes intervals, bootstrap intervals, [prediction intervals].
- For a given dataset, there may be different ways of calculating these intervals.
- The choice between intervals and method of calculation, depends on the assumptions that can be made about the distribution of the data. Bootstrap (and other non-parametric) intervals typically make fewer assumptions than other intervals. Try to understand your choice – in an ideal world ask a statistician.
- There are also links between interval estimation and hypothesis testing

Concluding (cautionary) remarks

- We have covered some of the main ideas, but only a tiny part, of statistical inference. For example, there was nothing on traditional non-parametric inference.
- Inference has many subtleties. *The American Statistician* often has examples of this in relatively simple contexts. For example, see Tuyl et al. (2008) for a discussion of what is an 'uninformative' prior distribution for a binomial parameter – a situation we considered; recall also the controversy surrounding p-vales.
- For some standard verification measures, software and/or formulae exist for quantifying uncertainty, but in many cases this is not yet the case. This is no excuse for ignoring uncertainty.

References

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.

Gilleland E. (2020) Bootstrap methods for statistical inference. Part 1: Comparative forecast verification for continuous variables. 37, 2117-2134.

Gilleland E. (2020) Bootstrap methods for statistical inference. Part 2: Extreme value analysis. 37, 2135-2144.

The American Statistician (2019) Special issue (401 pages) on p-values and alternatives.

Tuyl F, Gerlach R & Mengersen K (2008). A comparison of Bayes-Laplace, Jeffreys, and other priors: the case of zero events. *Amer. Statist*., **62**, 40-44.