INTERVAL ESTIMATION AS VIEWED FROM THE WORLD OF MATHEMATICAL STATISTICS

Peter Clifford
Address for correspondence:
Statistics Department, 1 South Parks Road Oxford, OX1 2TG,
clifford@stats.ox.ac.uk

1. HYPOTHESIS TESTING

Modern statistics dates back to the beginning of the 20th century. It developed in response to questions raised in two important new areas:

- Biometrics — the quantitative measurement of living things, as pioneered by Darwin and Galton.
- Production Control — process monitoring in industrial mass production.

In both these areas, a starting point for scientific enquiry is usually the formulation and testing of a *null hypothesis*, the hypothesis of no change. In agricultural trials, for example, the null hypothesis would assert that a new fertiliser had no effect on wheat yield. For an industrial production line it would assert that the process is under control.

Faced with growing numbers of enthusiastic data gatherers, early statisticians saw benefits in devising simple rules for testing the null hypothesis. Pressure of work dictated expediency: "time is precious – analyse the data and move on to the next client". The statistician's perspective was made explicit by Neyman. He argued that:

The ensemble = a life-time of statistical advice.

Neyman's advice was to control the frequency of Type I error within this ensemble. In other words, in your career as a statistician, arrange that the frequency of rejecting null hypotheses incorrectly is no more than, say, 5%. Naturally, you should also try to maximise the *power* within this constraint, i.e., you should try to make sure that you reject null hypotheses as often as possible when they are false.

2. CONFIDENCE INTERVALS

In many applications, the statistical model is determined by a real-valued parameter θ. To obtain an interval estimate for θ, Neyman suggested testing each value of θ individually as a null hypothesis; the confidence region is then the set of θ that are not rejected. For a suitable class of tests, the region will be an interval. If all of the tests have a 5% Type I error then a 95% confidence interval is obtained.

By constructing intervals in this way, you can ensure that in your lifetime as a statistician you will successfully cover the true value of the parameter 95% of the time (no matter what the true value of the parameter is). In other words the coverage probability is 95% on average. From the statistician's perspective, this is highly satisfactory!

So how does this work in practice? A client collects data, $\mathbf{x}$, and wants to test the null hypothesis that the mean of the sampled population is some specified number θ. The client goes to a statistician and asks for a ruling. Here are the strategies of two statisticians who specialise in controlling Type I error.

Statistician A No matter what $\mathbf{x}$ or θ is, reject θ when $U < 1/20$, where U is a newly simulated random variable from a uniform distribution on $(0, 1)$.

Using this procedure, Statistician A will reject the null hypothesis 5% of the time. The Type I error probability is 5%. The power is also 5%.

Statistician B When $54.0 < \theta < 54.0001$ don't reject it, otherwise reject θ when $U < 1/20$, where U is a newly simulated random variable from a uniform distribution on $(0, 1)$. Here the probability of Type I error is bounded above by 5%.

What will the confidence intervals look like? For Statistician A the confidence interval will be empty 5% of the time and it will be the whole real line 95% of the time. The statistician is happy because the coverage probability is 95%. For Statistician B, 5% of the time the confidence interval will be $(54.0, 54.0001)$ i.e., some arbitrary small interval, and the rest of the time the confidence interval will again be the whole real line. The coverage probability is now slightly larger than 95%. Again the statistician is happy.

Now look at things from the client's perspective. From Statistician A they get either the whole line or the empty set. This is clearly unacceptable to the client. So they go to Statistician B, and luckily get the interval $(54.0, 54.0001)$. Now the client is happy too, because the interval is small. Does this make sense?

A similar situation arises when constructing confidence intervals for a parameter constrained to be positive. In the simplest case, the model is that the observation x is sampled from a Gaussian distribution with mean μ and known variance σ^2, where $\mu > 0$. The two-sided test of the hypothetical value μ rejects when $|x - \mu|/\sigma$ is larger than 1.96. The 95% confidence interval $C(x)$ associated with this family of tests is given by

$$C(x) = \begin{array}{ll} (x - 1.96\sigma, x + 1.96\sigma) & \text{if } x > 1.96\sigma\ , \\ (0, x + 1.96\sigma) & \text{if } x > -1.96\sigma, \\ \text{empty} & \text{if } x < -1.96\sigma. \end{array} \quad (1)$$

From the point of view of coverage probability there is nothing particularly wrong with this family of intervals. They do cover the unknown value of μ with the right frequency. However, they are not necessarily a satisfactory summary of our beliefs about μ. For example, if $\sigma = 1$ and $x+1.96\sigma = 0.0001$, the confidence interval for μ is $(0, 0.0001)$, an unconvincingly precise confidence interval.

Neyman would say: "a bad test has led to a bad confidence interval". In Neyman's view a good system for constructing confidence intervals is one which minimises the chance of the intervals containing false values of the parameter. This relates directly to the notion of uniformly most powerful (UMP) tests. Unfortunately, UMP tests don't often exist. Neyman's suggested compromise is to use tests and hence confidence intervals based on the maximised likelihood ratio (i.e., the recently rediscovered "unified approach").

3. PROBLEMS WITH CONFIDENCE INTERVALS

3.1 Discreteness

In discrete problems, i.e., problem involving counts, coverage probabilities for confidence intervals cannot be fixed precisely at 95%. This is because the associated tests of null hypotheses have discrete probability distributions. The usual practice is to construct conservative intervals, i.e., intervals whose coverage probability is no smaller than 95%. Various methods have been proposed to obtain coverage probabilities closer to the nominal value.

3.1.1 Randomisation

Suppose that the test statistic $T(x, \theta)$ for the hypothetical value θ rejects when $T > k$. The critical value k has to be chosen so that the probability of rejection is 5% under the null hypothesis θ. If T has a discrete distribution, then it may turn out, for example, that $k = 5$ is too large and $k = 4$ is too small, i.e., $P(T \geq 5) < 0.05$ and $P(T \geq 4) > 0.05$. One suggestion is to reject when $T \geq 5$ and when $T = 4$, reject when

$$U < \frac{0.05 - P(T \geq 5)}{P(T = 4)},$$

where $U \sim \text{Unif}(0, 1)$.

The rejection probability is now exactly 5% and so the confidence interval constructed from this type of test will have exact coverage probability

Another possibility is to convert the discrete variable into a continuous one, e.g.,

$$T + U \quad \text{where } U \sim \text{Unif}(0, 1).$$

These ideas are mathematically interesting, but they are rarely used in practice. It should be noted however that randomised intervals are always shorter than conservative intervals constructed without randomisation.

Yet another technique is to use *mid-p* values. In this approach, tail probailities are calculated with the convention that

$$P(T \geq 4) \approx \frac{1}{2}p_4 + p_5 + \ldots.$$

Intervals obtained in this manner may have good average coverage probabilities.

3.2 Post-data conditioning

Mathematical statisticians have devoted a great deal of energy to the study of Neyman's approach to hypothesis testing and confidence intervals in the past 70 years. Many disturbing aspects of the method have been exposed, despite its widespread acceptance in applications. Important questions are raised by the possibility of post-data conditioning and various illustrative examples have been devised and discussed.

3.2.1 Two measuring devices

Suppose that you need to measure a physical quantity, and two portable measuring devices are available, both measure subject to experimental error. The first has standard deviation 1 and the second has standard deviation 10. On any particular day, only one of the devices will be in the laboratory and there's a 50% chance it is the accurate one. Nevertheless you plan to use whichever device is there. From a frequentist viewpoint, your average standard error will then be 7.01. So when you make a measurement x, you can report a 95% confidence interval $(x - 13.9, x + 13.9)$.

However, when you arrive in the laboratory, you see that the the accurate device is there. Now it seems sensible to report a confidence interval $(x - 1.96, x + 1.96)$. In other words, when you are given information about which device is available, you construct a different confidence interval. This is an example of post-data conditioning.

3.2.2 Estimating the middle of an interval

Another example of this type is as follows. Suppose that a sample $(x_1, \ldots, x_n)$ is taken from a uniform distribution on the interval $(\theta - 1/2, \theta + 1/2)$, where θ is an unknown parameter.

The *sufficient statistics* are $x_{\min}$ and $x_{\max}$. A simple estimate of θ is

$$\tilde{x} = \frac{x_{\min} + x_{\max}}{2}$$

so that

$$(\tilde{x} - b_n, \tilde{x} + b_n) \quad \text{where } 2b_n = 1 - (0.05)^{1/n}$$

is a 95% confidence interval for θ. Notice that the width of the interval is fixed at $2b_n$, regardless of the values in the sample $(x_1, \ldots, x_n)$. When $n = 10$, for example, the width is 0.26.

Now consider the range $r = (x_{\max} - x_{\min})$. If $r = 0.99$ say, we know for sure that θ is within 0.01 of $(x_{\min} + x_{\max})/2$. However, when $n = 10$, for example, the confidence interval is certain to cover the true value of θ. In fact the confidence interval is 26 times wider than it need be.

In the statistical literature $x_{\max} - x_{\min}$ is said to be an *ancillary* statistic. It is a function of the sufficient statistics that has a distribution that does not involve the unknown parameter θ. It is reasonable to condition on the value of the range r and construct a conditional confidence interval based on $\tilde{x}$. Intervals obtained in this way will have a width which depends on r. The coverage probability will be right for each value of the conditioning variable and since the distribution of the conditioning variable does not depend on θ, the correct coverage probability is guaranteed universally.

3.2.3 Poisson count data with background

For Poisson count data with a known background b, the probability that n events are observed is

$$P(N = n) = \frac{e^{-(b+\theta)}(b+\theta)^n}{n!}, \quad \text{where} \quad b, \theta > 0.$$

The sufficient statistic for θ is n.

Conceptually, the random variable N can be written as $N = X + Y$ where $X \sim$ Poisson(b) and $Y \sim$ Poisson(θ), although neither of these component variables are observable.

Since

$$N \leq n \quad \text{implies} \quad X \leq n,$$

it is tempting to condition on the event $X \leq n$. However, the sufficient statistic n is one dimensional and there is no non-trivial function of n with a distribution not involving θ. In particular, the event $X \leq n$ is not ancillary. Because of this there is no guarantee that the coverage probability of intervals obtained by conditioning on $X \leq n$ will be correct. It is also worth noting that

$$N \leq n \quad \text{also implies} \quad X \leq n + 1,$$

and it is not clear whether there are advantages in conditioning on $X \leq n + 1$ rather than $X \leq n$.

3.2.4 The standard t-interval

Finally, there are still surprises in even the most standard problems. Suppose that $(x_1, \ldots, x_n)$ are sampled from a Gaussian with mean θ and variance σ^2, both unknown. The usual $100(1-\alpha)\%$ C.I. for θ is

$$C(\overline{x}, s) = \left(\overline{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

where $\overline{x}$ and s^2 are the sample mean and sample variance and t_α is the $100(1-\alpha)$ percentile of the t distribution with $n-1$ degrees of freedom.

For $n = 2$, $\alpha = 0.5$, we thus have for all θ, σ^2,

$$P_{\theta,\sigma^2}(\theta \in C(\overline{x}, s)) = 0.5$$

However, Brown (Ann. Math. Stat. 38, 1967, 1068-1071) showed that

$$P_{\theta,\sigma^2}(\theta \in C(\overline{x}, s) \mid |\overline{x}|/s \leq 1 + \sqrt{2}) \geq \frac{2}{3}$$

In general, a set A is said to be a *negatively biased relevant subset* for a $100(1-\alpha)\%$ confidence interval C_x if there exists $\epsilon > 0$ with

$$P_\theta(\theta \in C_x \mid x \in A) \leq 1 - \alpha - \epsilon$$

for every θ, and said to be *positively biased relevant subset* if

$$P_\theta(\theta \in C_x \mid x \in A) \geq 1 - \alpha + \epsilon$$

for some $\epsilon > 0$. In this example,$\{(x_1, \ldots, x_n) : |\overline{x}|/s \leq 1 + \sqrt{2}\}$ is a positively biased relevant subset.

3.3 Inadmissibility

Even the most obvious confidence regions may turn out to have unacceptable properties. This is particularly so for higher dimensional regions. Suppose that you have p quantities $\mu_1, \ldots, \mu_p$ of interest and measurements $x_1, \ldots, x_p$ on each. For simplicity assume the measurements all have a Gaussian distributions with standard deviation 1. The obvious confidence region for $(\mu_1, \ldots, \mu_p)$ is

$$C(\mathbf{x}) = \{\mu : \sum_{i=1}^{p} (x_i - \mu_i)^2 \leq \chi^2_p(95\%)\}$$

where $\chi^2_p(95\%)$ is the 95-percentile of the χ^2 distribution with p degrees of freedom.

The coverage of the confidence region is exactly 95%. However, when $p > 2$ the region

$$C_{\text{JS}}(\mathbf{x}) = \{\mu : \sum_{i=1}^{p} (t_i - \mu_i)^2 \leq \chi^2_p(95\%)\}$$

where $t_i = 0$ if $\sum x_i^2 < p - 2$ and $t_i = x_i(1 - (p-2)/\sum x_i^2)$ otherwise, has the same volume but a higher coverage probability.

4. BAYESIAN METHODS

Bayesian methods predate the frequentist approach of Neyman and his co-authors. It can be argued that they are the 'right' way to do statistical inference, i.e., the right way to modify one's beliefs in the face of uncertain information. A stumbling block is the question of the choice of prior, since posterior beliefs are a reflection of prior beliefs and the likelihood function. A number of suggestions have been made for an 'objective' choice of prior. Foremost among these is Jeffreys prior.

4.1 Jeffreys priors

Consider $x \sim \text{Binomial}(n, p)$. A plausible 'non-informative' prior for p is the uniform prior on $(0, 1)$, expressing 'ignorance' about the value of p. Note, however, that if $p \sim \text{Unif}(0, 1)$, the square root of p has a non-uniform distribution with higher density near 1 than 0. Thus, ignorance about p translates to knowledge about the parameter $\sqrt{p}$.

In some settings, it might be argued that there is a single 'Natural' or 'important' parameterisation, so that a specification of ignorance for that parameterization is natural. In others, priors which are non-informative for some parameterisations but not others may be undesirable.

For a model with parameter space $\Theta \subseteq \mathbf{R}$, the Fisher information is

$$I(\theta) = E_\theta \left(\frac{\partial \log(f(x \mid \theta))}{\partial \theta} \right)^2$$

where $f(x \mid \theta)$ is the sampling distribution and the expectation is taken over $f(x \mid \theta)$. Under regularity conditions,

$$I(\theta) = -E_\theta \left(\frac{\partial^2 \log(f(x \mid \theta))}{\partial \theta^2} \right).$$

In such a setting, the *Jeffreys Prior* for θ is defined by $\pi(\theta) \propto I(\theta)^{1/2}$, to be proportional to the square root of the Fisher Information at θ. Note that in general the Jeffreys prior may be improper (i.e., it may not have a finite integral).

Note that by the chain rule,

$$I(\theta) = I(h(\theta)) \Big(\frac{dh}{d\theta} \Big)^2.$$

If θ has the Jeffreys prior and h is a monotone differentiable function of θ, the prior induced on $h(\theta)$ by the Jeffreys prior on θ is

$$\pi(h(\theta)) = \pi(\theta)|\frac{dh}{d\theta}|^{-1} \propto I(\theta)^{1/2}|\frac{dh}{d\theta}|^{-1} = I(h(\theta))^{1/2}.$$

Thus the Jeffreys priors are invariant under reparameterisation.

Recall the interpretation of $I(\theta)$ as the ability of the data to distinguish between θ and $\theta + d\theta$. If the prior favours values of θ for which $I(\theta)$ is large, the effect is to minimize the effect of the prior distribution relative to the information in the data and hence to be uniformative about θ.

4.1.1 Jeffreys prior for count data

The Jeffreys prior for a signal θ with Poisson data n and background b is inversely proportional to $\sqrt{\theta + b}$. The posterior density of θ is then proportional to

$$\frac{1}{\sqrt{\theta + b}} \frac{e^{-(\theta+b)}(\theta + b)^n}{n!}, \quad \text{for } \theta > 0.$$

The 95% highest posterior density (HPD) credible interval, is the interval of θ values that contains 95% of the posterior density, with the property that any value of θ outside the interval has a lower density than any value inside.

5. SUMMARY

Increasingly, Bayesian methods are being used in the analysis of complex data sets, where typically there is

- a high dimensional parameter space
- a reservoir of wisdom from which prior beliefs can be distilled (at least approximately)
- willingness to use computer intensive methods for simulation and model-sensitivity analysis.

Modern statistical practice distinguishes between routine problems, where standard frequentist methods are used (small consultancy fee!), and elaborate problems, where computer-intensive Bayesian methods are increasingly popular. Examples are: image processing, large-scale clinical trials in medicine, mixture modelling, non-parametric regression, etc. The techniques involve sampling the (high-dimensional) parameter θ from a posterior density proportional the product of the likelihood and the prior density. The Metropolis algorithm (Markov chain Monte Carlo) is used to provide the samples.

It should be noted that Bayesian methods are used routinely in engineering applications. Signal processing and control engineering depend heavily on the Kalman filter, a Bayesian updating formula applicable to linear Gaussian systems. There has been recent interest in extending these techniques to non-Gaussian signal processing problems. The new computer-intensive techniques are known generically as *particle filters*.

5.1 The individual and the collective

Neyman devised confidence intervals as a method for analysing mass-produced statistical problems:

- no need to elicit prior information (or build an expert system)
- simple to construct (for naive practioners)
- good for the ensemble (not necessarily good for the individual)

There is an analogy with the popularity of certain computer algorithms. For example, QUICKSORT is the most commonly used method for sorting N numbers. It is a randomised algorithm, with an expected running time of $A_1 N \log(N)$. The worst case running time is of order N^2; this can happen by chance on any particular occasion.

The algorithm is good for the ensemble, but you might be the unlucky one! What to do about it? It turns out that there is a different (non-randomised) algorithm which runs in $A_2 N \log(N)$ time, but with $A_2 > A_1$. If you only had one very large set of numbers to sort in your life, it would be a 'safer' strategy to use this algorithm.

The lesson for data analysis is that if you are going to spend a lot of time and money on collecting and analysing a particular set of data, you may not be interested in how a particular statistical technique performs for the ensemble. It makes more sense to adopt a selfish approach and build personal confidence in your knowledge. In such circumstances, Bayesian methods are appropriate.

5.1.1 Reading

The italicised terms in the text are defined and placed in a historical context in: *The Encyclopaedia of Statistical Science*.

There is still a great deal of interest in comparing various methods of constructing frequentist confidence intervals.

Newcombe(1998) Two-sided confidence intervals for a single proportion: comparison of seven methods. *Statistics in Medicine*

Newcombe(1998) Two-sided confidence intervals for differences between two proportions: comparison of eleven methods. *Statistics in Medicine*

Bayesian methodology is covered in

- *The Bayesian Choice*. Christian Robert
- *Bayesian Methods: Kendall's Theory of Statistics*, Tony O'Hagan
- *Bayesian Statistics*, many volumes edited by Bernardo and Smith.
- Bayesian computation via Gibbs and related Markov chain Monte Carlo methods (with discussion), *Journal of the Royal Statistical Society (B)*, Vol **55**, pp 3 – , 1988.

Discussion after talk of Peter Clifford. Chairman: David Cassel.

Giulio D'Agostini

Can you please comment about the physicist's point of view, which has been essentially oriented to induction and inference? You have shown the statistician's point of view, and I am happy, it was nice, but the physicist's point of view was always induction. We try to understand something about the nature of making statements about true values, about theories and so on.

Peter Clifford

Well, maybe I didn't spell it out, but before I came here I assumed that all physicists were Bayesians. Physicists are interested in induction, they want to modify their beliefs about the true state of nature on the basis of the data that they've observed. When you are busy integrating out parameters, in a sense, you keep slipping into a Bayesian mode of operating. What I've been seeing at this meeting is a sort of flip-flop phenomenon between Bayesian and classical ways of thinking. People in the workshop who I assumed were avowedly Bayesians, are now saying: 'Well classical methods are maybe OK'.

My training is as a frequentist, I was a student of Neyman's, but I would say that nowadays since we have the computing resources available, in specific problems where we've got the time and the manpower to really analyze these problems, Bayesian methods are the best way forward.

Michael Woodroofe

I was interested in your statement that the prior doesn't matter too much in high-dimensional problems. That's not universally true. We know examples where it's false, for example order parameters. Could you give us a little idea of when and where it will be true that the prior doesn't matter in high-dimensional problems?

P. Clifford

I may have appeared to say it. What I meant to say was that Bayesian methods are being used in high-dimensional problems, and that it appears that they're getting away with it, not because the problems are high-dimensional, but because in the examples which work, there's sufficient data to swamp the priors. So what I wanted to say was that the methods which are being used successfully in practice, are methods where the data are really telling you what's going on. The prior is really there as a support for your inference, but it's a support which gradually you're able to remove, as the data starts to dominate.

F. James

I think we should be careful not to confuse two different situations. One is where there is a prior probability, a phase space or something, and that's the case for Kalman filtering, the maximum entropy method and so forth, and there everyone uses those methods. You don't have to be a Bayesian, that's because the idea of using a prior probability makes sense because there *is* a prior probability. The problem is, do you want to put in a prior probability for something like the mass of the Higgs where there is no prior probability? I define a Bayesian as somebody who uses a prior probability that he pulls out of a hat, for example. The physicist who tries to be objective does not want to put in his prior feelings in those cases.

P. Clifford

Let me just say that in signal processing there is no prior when you're looking at a tracking problem and you don't know where some object of interest is. You're attempting, by radar scanning or some other means, to work out where the object is. The object actually starts off at a fixed position. This position is not random but when you use the Kalman filter for the problem, you expediently put in some representation of your beliefs about where the object might have been initially. What happens is that as the data flows in, the initial belief is modified by real data, using Bayes formula. It doesn't really matter that you might have got things wrong for the first few observational steps because the data starts to swamp the prior. It's not true that there's a natural prior in Kalman filtering. The prior is just something convenient and vague.

F. James

That's just the case where it doesn't matter. If it doesn't matter what prior you put in, then that's fine as well. The problem is when it does matter, and you don't want to put a prior in, and that's the case that we're worried about here.

P. Clifford

Right, but the Bayesian response to that would be that you do a sensitivity analysis. Let's see how sensitive the conclusions are to the prior that you put in.

H. Prosper

My comment was somewhat similar to Fred's. Of course I'm quite happy to use these methods, but the difficulty that I always find is that my colleagues will say 'well, but our result changes if one changes the prior because we have seen no events', and the question is what should be the response. I agree with you, the response should be that if in fact your answer depends very much on the prior, then the conclusion should be that you have insufficient data to say anything sensible.

P. Clifford

I think that's absolutely right, but if you're in a situation with high prior sensitivity, where the data is really not telling you a whole lot, then that's an important piece of information too.