

## PANEL DISCUSSION

Panel Members: Bernard Silverman, David Cox, Jerry Friedman and Bob Cousins  
Chairman: Louis Lyons

**Louis Lyons** I think members of the panel need no introduction whatsoever: Bernard Silverman, David Cox, Jerry Friedman and Bob Cousins.

What I was going to do was to ask the members of the panel if there are any of the questions they would particularly like to talk about. I told David I would give him the first go, so David let's start with you.

**David Cox** My impression is that most advances in the use of statistical methods come not from looking through a library of techniques which are available but by those with a primary interest in statistics and with some knowledge of the subject matter sitting down over a period with a group in the subject field who have some knowledge of statistics. Issues of formulation are crucial. Eventually, one hopes, new ideas will emerge that both address the subject-matter questions at issue and which maybe will be more widely useful.

Most of the really interesting and important developments in statistical methods have come that way. Direct transplanted ideas from one field into another are less commonly totally successful. Now collaboration with physicists and statisticians is going to be particularly difficult for statisticians because you know so much already and also of course you have an enormously strong tradition of independent mathematical thought.

To address Question 1 I feel it is helpful to think of statistical methods in four chapters and two of the chapters have been quite strongly represented in this meeting, which incidentally I have highly enjoyed; it has been much more interesting than most statistics meetings.

The four chapters, not in any particular order, are first of all likelihood, Bayes, Neyman, Fisher, confidence interval calculations and so forth; clearly there is a lot of interest in that. It may be there are things in the statistical literature, and matters that could be absorbed into the thinking, perhaps particularly the notion of taking profile likelihoods and modifying them to make them perform better, particularly when there are a large number of nuisance parameters which is the critical issue in many contexts.

In a sense at the other extreme, there's the enormous collection of particular statistical methods that had been found useful somewhere or other, some highly exploratory, some partially exploratory, some graphical, some numerical. You might say that that's statistics as covered or potentially covered by R.

But I mentioned there are two chapters that have been hardly represented here. Now one would be what we call roughly applied stochastic processes. I know in some contexts this might not be regarded as a part of statistics at all. I mean constructing particular probabilistic models often dynamic, but not necessarily so, and studying their properties as issues in applied mathematics rather than in the pure mathematics of probability theory. It may be you regard all that as part of theoretical physics anyway and not part of statistics. So there is the issue of constructing new models that incorporate the physics and the probability, not quantum probability, but the physics of observational probability into new models.

Now the fourth chapter that is scarcely represented is the issue of the design of investigations and there are two major sections of statistical theory here, one to do with sampling and I point out that includes stereology. The other aspect is the design of experiments and I think this particularly relevant here to the design of computer

experiments, including systematic sensitivity studies of models with many adjustable input features. Notions of Latin hypercube sampling and fractional replication which come originally largely out of technological industrial statistics could be useful.

**Louis Lyons** What exactly is stereology?

**David Cox** The study of the properties of objects in three dimensions by two or one dimensional probes or in general  $k$  dimensions, by appropriately sampling in lower numbers of dimensions.

**Louis Lyons** Well I'm sure that your first remark, about getting statisticians involved in analyses, will be welcome news to members of the physics department here and we'll be knocking at your door with problems that we would like to have you help us solve. Do any other members of the panel want to add anything to this?

**Jerry Friedman** In terms of the machine learning component, you seem to be coming up to speed rather fast. If you want to see if there are some things you have overlooked that might be useful, I'll shamelessly recommend our book. Get it from the library and go over the table of contents and see if you find some things in there that look a little strange that you haven't seen before and go to the relevant chapter. But as I said, I think that in the machine learning area you have got up to speed rather fast.

**Bernard Silverman** A very interesting aspect of this Conference is that the main impetus for the development of statistics in the twentieth century was its relevance to agriculture and biological and later medical applications. To see it in a physics context is fascinating because one of the things you seem to be doing, quite reasonably, is to recapitulate a lot of the discussions that went on in statistics in other contexts previously and to appropriate those discussions to the physics context. I suspect you need to think more about Bayesian methods, and to be more comfortable about using Bayesian techniques. While not everyone on the panel would necessarily approve, that has been a major shift in statistics in recent decades. There are many issues in these questions where essentially frequentist methods are more problematic, and the Bayesian approach might be more natural.

But I would really like to stress what David Cox said, which is that you should not think that statistics is about techniques. You would never think physics was about techniques, would you? Physics is about ideas and understanding and intuition and so on, and the techniques are just on top of that. Astronomy is not about techniques, is it? It's not, because it is about ideas, and the techniques are just a way through to the ideas. It's the same with statistics. Statistics is a developed scientific field and it is not a collection of techniques. It's a way of thinking which then gives rise to techniques and it is important to get into that way of thinking, rather than to say "Wouldn't such and such a technique be more appropriate for this problem?"

**Louis Lyons** Let's move on to Bernard on Question 4.

**Bernard Silverman** Now you'll probably discover when we do this that when you have two statisticians you have three opinions. I was fascinated by Question 4 in a way because the real issue to Question 4 is to try and formulate the problem in such a way that you can actually see what's going on. This problem seems to have arisen from someone saying : "I've got an experiment and I have observed the data point  $x$  which has a Poisson distribution with parameter  $\lambda$  and I want to know if  $\lambda$  is bigger than  $\lambda_0$ ". Then they said "By the way I know something about  $\lambda$ , I know that  $\lambda$  is  $b_0 \pm \sigma_b$ ." Then you realise that that's not what they know at all; in the question we were posed we were told it could be thought of as being determined in a subsidiary experiment. So what I actually had to do on reading this was to re-work out the original experiment. There seems to be another experiment going on where there is another data point  $y$  which has Poisson distribution

with parameter  $r\lambda_0$ . So it isn't only the main experiment that is conducted, but this other one as well, and really you just need to make sure you have written both of them down. Once you do that, you can see that to do any inference at all you want to know what the distribution of  $x$  is, conditional on  $y$ .

One way to approach this is to take a Bayesian viewpoint, putting a prior on  $\lambda$  and then calculating  $p(x|y) = p(x|\lambda) * p(\lambda|y)$ , using Bayes theorem to find the latter conditional probability.

The key point, however, is not the use of a Bayesian vs frequentist approach, but the need to go right back to the experiment that was conducted, to look at all the experiments at once and write everything down about it. That's better than saying that  $b$  has been measured in a subsidiary experiment as something or other.

**Louis Lyons** We've got two more statisticians here so I expect four more opinions! Jerry, David do either of you want to say something about that?

**David Cox** And the audience could have opinions too, maybe an infinite number!

**Jerry Friedman** I'm going to disappoint you, I don't have another opinion.

**David Cox** You have to write down the likelihood. It's the key and then the issue is what do you do with the likelihood. If you have a prior that is evidence-based, I don't think anyone would dispute that you should use it. If you've not got such a prior, you're into this issue of reference priors and flat priors, and there's been a lot of work about that, in a way for two hundred years, and certainly for the last two days. It's a minefield and done properly the reference prior would give a beautiful answer that we are all satisfied with, but what does a beautiful answer mean? Well if I wanted to be argumentative, and I'm totally not argumentative, I would say if it gives something that has at least tolerable frequentist properties. If it gives an answer that has very bad frequentist properties, I can't think anyone would defend it. So in the end I would be very happy to use the Bayesian formalism as a way of getting an answer. I have done so and have no qualms - if that's the way to get an answer, I'll do it. But in the last analysis, I have to say, "I doing something that is going to produce answers close to the right answer most of the time, in other words is it calibrated properly?" which is much about what that means. In some sense I would see it as a hypothetical frequentist interpretation.

**Steffen Lauritzen** Just to keep things down to earth, in this example when you formulated it the way that Bernard did, if  $r$  is known, then  $x + y$  is sufficient for  $\lambda$  and it becomes a straightforward simple hypothesis in the conditional binomial, just with the scale factor  $1 + r$  entering, whether you are Bayesian or not.

**Louis Lyons** Bob do you want to say anything?

**Bob Cousins** Well I had my 40 minutes on exactly this problem so I won't repeat too much. I'll just say what Sir David said, that at the end of the day you want to say that some probability  $P$  equals some number. If a student asks "How do you define probability  $P$  in that statement?" you would like to be able to give them an answer. I think that in our field the frequentist answer is the easiest one to explain, but I'm also perfectly happy with the subjective answer. What I'm unhappy with are the ones in between, those priors which are used in a Bayesian machinery but unless the probability  $P$  comes out with the right frequency I don't know how to interpret it. Since this particular problem is so interesting in our field, it's been studied a lot and for the variety of methods that I talked about, their frequentist coverage has been studied. I refer, for example, to the papers by Conrad and collaborators on integrating out the nuisance parameter formally and seeing how it works. What we call the MINUIT MINOS method is profile likelihood, and is discussed in a recent paper by Rolke, Conrad and Lopez. Those papers contain a lot of information, and they also tell you what sort of study you can do with whatever technique you use. Then if you're really up to it you can do a full-blown

frequentist construction the way Kyle Cranmer and Giovanni Punzi talked about. Then you don't have to check if it covers because it does so by construction, but it likely overcovers. So that is interesting to study as well.

**Luc Demortier** I just thought it would be useful to clarify that when statisticians say flat priors they don't mean uniform priors, they just mean objective priors. Am I correct in that because there might be some confusion? We try to discourage the use of flat uniform priors but not necessarily the use of objective priors or reference priors. But the statisticians sometimes use the terminology flat prior to mean objective prior in general.

**David Cox** I certainly had in mind either the Jeffreys prior in simple cases or something like Bernardo's reference priors in more complicated ones.

**Bob Cousins** Well it's OK because you did not say in what metric your prior was flat! It will be flat in some metric!

**Tomi Zivko** I would like to give a short comment about priors. Yesterday I gave a talk in which I presented work of my colleague and myself. In the talk I claimed that we started from Jaynes, Polya and Cox's desiderata. That means that we started from a purely Bayesian point of view, and using only those assumptions which are found in the desiderata, we obtained calibrated solutions, which is a purely frequentist result. But there were no comments after the talk, no objections.

**Louis Lyons** OK so I guess people need to have a chance to read it and absorb the ideas there.

Maybe we should move on to another question. I think Bob hasn't had a chance to choose anything yet.

**Bob Cousins** So I've thrown out all the ones about nuisance parameters and Bayesian analysis, and I've instead chosen the one about blind analysis (Question 13). I have been involved in three experiments that performed blind analyses, including the BNL E791 rare kaon decay experiment, where the blind analysis led by Bill Molzen seems to have led to widespread use of blind analyses in HEP, and also one including Josh Klein, who has recently written a review with Aaron Roodman [Ann. Rev. Nucl. Sci. 55 (2005) 141]. So the question is "What do you do if you open the box and despite of all your due diligence, you see that you've been stupid and there is an obvious background that you did not anticipate?" This actually happened to me. We opened up the box and looked at the events in the signal region, and found that two events had all their ADC (analogue-to-digital converter) readouts zero – not even at the pedestal value, but really zero. So we came up with the criterion that it is OK to throw away an event after you open the box, if you would look foolish by not throwing it away. You should feel free to throw away an event, rather than go to a conference and stand up and say "I'm going to stick to my principle of blind analysis and keep this event, even though my read-out was not working."

In our case, the effect on acceptance was completely negligible when we added the 'ADC read-out was working' cut to all events, and it was such a clear-cut case that we did not take any further action. If this was not the case, however, as the question points out, this can introduce a bias in a subtle way. Suppose there are twelve possible backgrounds that you haven't thought about, and if you cut on them each would introduce a 5% inefficiency. Then you open up your box you find there is only one of them that actually appears that you haven't thought about. You add this one cut and take a 5% hit on efficiency. But if you had really thought about all your backgrounds in advance and decided to eliminate them, you would have had twelve hits on efficiency, each of 5%. So this is a source of bias one is left with, but probably it is even worse for an unblind analysis. In practice I have found that people doing blind analyses are very good about thinking hard about

potential background sources precisely because they want no surprises when they open the box.

There is another principle I'm convinced of from seeing all these blind analyses. You should freely look at 10% of the data inside the blind box, especially if it's a new experiment and it's not the third data set or what not. If you are going to tune your cuts on 10% of data, even if you do a bad job, the bias is not very strong. So I would first of all try to prevent this problem of unexpected backgrounds by looking at 10% of the data; and second of all I think if you really can improve your analysis after you open the box (for example by better calibration), then go ahead and improve the analysis.

**Louis Lyons** You are assuming you keep that 10% in the final analysis?

**Bob Cousins** Yes, keep the 10%. There are plenty of people in the world who think you can tune on 100% of the data and still have valid answers, so I'm perfectly happy to tune on 10% of the data and keeping that 10% in the final sample.

**Gary Feldman** I just wanted to add to Bob's comment that Josh Klein and Aaron Roodman have recently written a very nice paper on blind analysis and in it they have a great line which says "Doing a blind analysis is not an excuse for publishing a wrong result." Their recommendation is if you find a problem once you've opened the box, fix it and then just explain in your paper what you did.

**Louis Lyons** There's one blind analysis that I recently heard about and that's from the TWIST experiment at TRIUMF. They are doing a precision measurement to determine the value of a parameter by comparing their data with Monte Carlo simulations with different values of the parameter. They blind the value of the parameter used in the Monte Carlo experiment. So they can look at their data as much as they like and see if there any problems there because it's the Monte Carlo parameter that's blind. That seems quite a nice technique.

**Bangalore Sathyaprakash** I'm part of the LIGO scientific collaboration looking for rare events in our gravitational wave detector. To cut down the background, we look for coincidences within a certain time window - when there is an event in one of the instruments, we look for an event in the other instrument. After opening the box, we later discovered that one particular event was associated with an aeroplane flying over the instrument. Now, who could have thought an aeroplane-veto beforehand? So it was very hard. So I'm very heartened to hear that the advice that we should not just do blind analysis blindly. When you open the box, if you find something funny, just go ahead and allow for it in your further analysis.

**Rajendran Raja** The dangers of a blind analysis far outweigh its benefits since while you are blinded you are not monitoring the data in the blind box. Any loss in objectivity in an unblind analysis can be overcome by having simulators that model the data well and using the simulators to set the cuts. If D0 had been looking for the top quark blindly, we'd still be looking for a 65 GeV top quark.

**Byron Roe** I agree with Raja that one should be cautious about using blind analyses. I know that we have a minority opinion on this, but to me blind analyses are very useful when you don't know what to do with the data to give you a positive result for the parameters of interest. That is, an unconscious bias will not be able to bias the data to give you a positive result. The blind experiment I'm in is MiniBooNe, and in that experiment, you do know what to do. I think that diminishes greatly the value of blindness.

Now at breakfast Gary Feldman was pointing out there is a second problem called the stopping problem where you keep going until you've got an answer you like and then you stop looking for corrections. That certainly is a point to be worried about but you have to balance against that the problems that you introduce in your

analysis by having things blind. Surely you can always correct it afterwards, but it certainly is not such a great idea if you can do the analyses correctly in the first place.

**Louis Lyons** OK, so maybe we move on to another question. Jerry would you like to choose a topic?

**Jerry Friedman** I guess the message I get from this discussion of blind analysis that it's good to think outside the box!

There were a number of questions about machine learning, mainly about variable selection that can be dealt with rather quickly so I'll just try to give short answers to many of them, and some of them were the same. There were a couple of questions about variable selection and whether there are well understood techniques for selecting subsets of variables. The answer to that question is 'yes' but the question should be whether there are good techniques for variable selection. Variable selection techniques fall into two categories of filters and integral. With filters, before you apply whatever machine learning procedure you are going to use, you apply a different procedure to filter out bad variables before you run the machine learning procedure. This is often fast, but the problem is that the criterion you are using to select the variables is not the machine learning procedure that's going to use them, and so you may filter out useful variables. An example I have seen here is where you look at the power of each variable one at a time and filter out those appearing to be weakly related to the outcome variable. The problem is that a variable may not be strongly related by itself, but in combination with others it may have quite an effect. So the best way to do variable selection is in the context of the procedure that is going to be using the variables. I discussed this a bit in the techniques that I talked about during my talk where you actually get the relevance of the variables as used by the procedure that was trying to do the learning. Then you could filter out the ones that the algorithm said it didn't need.

There was another question about using many variables when you are suspicious that your Monte Carlo may in fact not describe the experiment; so maybe you should use less variables rather than try to use more variables that attempt to capture more features to do the discrimination between signal and background. I'm certainly sympathetic to the fact that the Monte Carlo may not be exactly correct. This is a common problem even when you have actual data. We call it non-stationarity in statistics and concept drift in the machine-learning literature. You take data at some time, you build a model that can describe that data rather well and you can cross-validate it. But then you apply it in the future and it doesn't work very well because the relationship between the variables has simply changed – you don't have the same system any more. There's been some work on that, but it's very hard and nothing is really satisfactory. In all these machine learning techniques, the presumption is that your training data is a random sample from the population of the future predictions. If that's not the case there's really not a lot you can do. You can try and over-regularise, because regularisation implies not fitting your data as well as you can. There are two reasons for doing this. One is because the data is random and randomness can lead you astray. That's the kind of thing statisticians deal with. Then there is the situation where the data has simply changed and again a solution to that is not to trust it too much, don't fit it quite so strongly. But systematic ways of doing that in the presence of concept drift are really not well developed in the machinery described in the literature.

There was one question on the Kolmogorov-Smirnov method for the goodness of fit test for multi-dimensional data. I was amused by that, well actually nostalgic, because as a young physicist it was that problem that got me interested in statistics. I realised back then that it was an important problem so I started to try and solve it. That led me into the statistics literature and I said "Hey this is pretty interesting". So that's how I got into statistics and I never left. At the last meeting I talked about general multi-variable goodness of fit testing and a procedure for doing that based on machine learning procedures. That's in the PHYSTAT2003 proceedings. Actually, the first statistics paper I ever wrote came out of this problem. It is published in the Annals of Statistics, and had in its title "Kolmogorov-Smirnov test in high dimensional data". It's around

1989-1990 Annals and if you look plus or minus a year in those you'll find a multi-variable generalisation of the K-S test and multi-variable generalisations for other goodness of fit tests like the Wald-Wolfowitz. You cannot straightforwardly extend K-S tests in high dimensions in the obvious way using the multi dimensional CDF. That doesn't work at all because of the curse of dimensionality. The CDF of the joint variables would tend to realise only two values (either 0 or 1) because it's the number of observations that are dominated by whatever point you are considering in the high dimensional space. You tend to dominate very few points simultaneously in all of the dimensions.

Are there any other machine learning questions? I thought Question 11 was kind of interesting. You have a system where you know that the target function can only be a function of a certain set of variables, like the matrix element technique talked about in the question. Since you know all of the variables that the target function could possibly depend upon, is there any value in constructing new variables that are functions of those variables and extending the variable set? The answer is 'yes'. The reason for this is that all machine learning procedures have some functions that they are good at learning and some functions that they are bad at learning. When you add the new variables, you change the function. I tend to try and understand things sometimes by considering extreme cases. As an example, suppose that the background was on a two-dimensional ball and the signal is on a larger ball surrounding it. It's only a function of those two variables. Given a perfect procedure those two variables would separate perfectly. But given finite data and an imperfect procedure (and all procedures are imperfect), if you added the variable which is just the sum of the squares of the two and put that into your machine learning algorithm you'd do a whole lot better. It would ignore  $x_1$  and  $x_2$  and just use  $x_1^2 + x_2^2$ . So the answer to that is 'yes' and the best way to do it is to use knowledge about the problem. If you use a technique based on trees that are not sensitive to lots of irrelevant variables, then you can feel free to add many derived variables if you have any suspicion at all that they might be useful. Then the natural variable selection technique of tree-based procedures will weed them out if they are not good, but include them if they are.

Finally, concerning the question on the James-Stein estimator. I guess I would answer that with a question: "What on earth is wrong with biased estimators?" Accuracy is an important thing, and if the lack of accuracy comes not so much from bias but from the variance, it is still lack of accuracy. If you can get a much more accurate answer by allowing a little bit of bias, I just don't see the downside.

**Louis Lyons** I was just going to say, maybe not every member of the audience knows about James-Stein estimators. So could somebody provide us with a two sentence explanation of James-Stein estimators?

**Bernard Silverman** James-Stein simply says this: If you are estimating a parameter of a large number of dimensions, then shrinking it back towards zero will give you a more accurate estimate than simply making it equal to the data. James-Stein in fact works with about four or more dimensions, but if you had thousands of observations - suppose you had a vector of a thousand parameters and you had one observation on each parameter - it's pretty obvious that it's much better to shrink the observations than simply to let them be equal to the parameters.

I want to add something to what Jerry said about Question 18 which is very interesting. It is not about James-Stein estimators as such, but to note that the discussion around this question is an example of recapitulating discussions that have gone on in statistics before. The objections raised in Question 18 are what people said when the James-Stein estimator was originally suggested, and so don't be surprised or worried if you're alarmed by issues which statisticians have been alarmed at before. The thing to bear in mind is that we may have thought through some of them already.

**Jerry Friedman** There are lots of other shrinkage estimators that you can probably use, such as ridge

regression, and lasso regression that I mentioned in my talk. Those are all shrinkage estimators and in general for prediction as opposed to estimation, shrinking is almost always a lot better than selecting variables. Variable selection is also a shrinkage estimator by the way; you just shrink some of the coefficients of the variables to zero.

**David Cox** The mathematically peculiar thing about the James-Stein estimator is that there is an apparent gain if you shrink towards anywhere. You've talked about shrinking towards the origin, which is natural, or generally to some linear space represented by a regression. If you shrink towards that, you can view that as a kind of empirical Bayes, even though it is not explicitly formulated in that way. The mathematical paradox or semi-paradox about it is that you may shrink towards anywhere you like, but you won't gain very much.

Going to the unbiased estimates, I can see only one situation in which unbiased estimates are particularly compelling. That would be if you had a lot of data in sections and you had a parameter that you were interested in for each section. You analyse each section of data separately and you get an estimate of that parameter and then you put those estimates into some linear representation. Then biasing the estimates would be a systematic error that persisted through the whole analysis.

**Jerry Friedman** But could you combine the data and do a general analysis with the shrinkage?

**David Cox** Yes, but in some contexts it is both easier but also I think more insightful to proceed stage by stage and you can see, as it were, what's happening in each bit of the data first before you put it into some big system.

**Bob Cousins** I'll just make one point that the way bias is usually expressed is in terms of a mean, and that is a metric-dependent statement. The most common bias we learn about in freshman's physics class is to take a sample variance and correct by  $n/(n-1)$ , which is to correct for the known bias in the maximum likelihood estimate. This gives a non-biased estimate for the variance but the RMS is biased. I think that for historic reasons it was probably defined that way. Fred James has suggested that if you are going to worry about bias, you should consider trying the median instead of the mean, because the median is independent of metric.

**Jerry Friedman** One last thing is that Bayesian techniques seem to be popular in Particle Physics, and all Bayesian estimates are shrinkage estimates.

**Steffen Lauritzen** Just to add to the number of opinions, what would worry me about the James-Stein estimate is certainly not the bias (and in that sense I agree completely with the rest of the statisticians) but rather the lack of invariance on changes of units and scale. I think that if I was a physicist, this would send a chill down my spine.

**Kyle Cranmer** I'd like to address this question and the physical context in which I looked at it, and I ran into exactly these changes of scale issues and funny things like that. We are looking at super-symmetry and what we would do in that case is that we'd measure masses of 10 or 12 particles, something like that, and the idea that we'd then shift all the masses from what we actually measured to stick into some other calculation seems really bizarre at first.

I'd like to briefly extend the question as if what we have really is some fundamental theory that might have a lot of parameters like a super-symmetry with 105 parameters and they predict the masses of these particles. Then we'd measure the masses of those particles. We'd be using the James-Stein part to try and improve our estimator of the masses but then back propagate that to try and get the parameters of the more fundamental theory. I'm wondering in that more extended context – if that made any sense – are there any more things to

worry about? It's a sort of two step procedure.

**Jerry Friedman** There are always plenty of things to worry about.

**David Cox** Could I make a couple of comments about the bias issue? One is the issue of  $n - 1$ . Of course it doesn't matter in one set of data but if you have variance being built up from various different sources which you think have about the same variability, then it would matter for the same reason that I indicated before.

The other point is that empirical Bayes estimates are typically shrinkage estimates, but not necessarily. If you do empirical Bayes and allow priors with very long tails, highly non-Gaussian, then under some circumstances empirical Bayes can be anti-shrinkage. It can take relatively extreme observations and push them further out rather than pull them closer in. One doesn't often see that, but certainly mathematically that's the situation.

**Harrison Prosper** I think our obsession with bias is really historical. In the old days it was much simpler to have each experiment do their analysis internally and then provide some summaries of what they have done. In that circumstance of course you'd like to have the summaries be such that you can combine them linearly and have an unbiased answer. Today we have 2, 3 Gigahertz machines and thousands and thousands of CPUs all over the planet. If ever we got to the point where we'd be willing to publish our data in some form that could be usable by other people, we could then do what Jerry suggests, which is that you do a large analysis of all these data and the whole issue of whether the thing should be unbiased becomes moot.

**David Cox** There's also of course the connection indirectly with Question 6. It seems to me a particularly important question and that's headed "Bayesian treatment of systematic uncertainties" because the terminology bias tends to suggest that a biased estimate is the same as one with a systematic error and that in some sense is misleading. Systematic errors are surely very important and there is a lot of concern in many fields that conventional statistical analyses, largely whether they are Bayesian or frequentist, deal with the errors that arise out of the random variations in the data, not out of any systematic errors in measurement. Bias from them is assumed eliminated by design. It is a strength of the Bayesian treatment that if you can put a reasonable prior on these systematic errors, then of course they can be incorporated into a fuller assessment of error. One danger there concerns independence assumptions.

**Louis Lyons** In particle physics some of the systematic errors come from trying to correct for biases and then the contribution to the systematic error will be how uncertain we were that we have allowed for this bias correctly.

**Bernard Silverman** The question posed is quite interesting. It uses the words 'systematical errors' but it's trying to get at some other kind of error which is an error which is in some sense unknown. But we know roughly what it might be and so if I were the prophet Dennis Lindley I would say the problem with the way the question is posed is that the last line isn't particularly correct. It says "I'd like to know whether there are methods that recognise the different nature of the statistical and systematic errors". Within a pure Bayesian way of thinking there is no different nature between the systematic and statistical errors. That's the whole point. For a Bayesian there is only one kind of randomness. Errors do not have different natures, they may have different origins physically, but in terms of how you model them they do not have different natures. That's the strength and weakness if you like of the Bayesian approach.

**Geoff Nicholls** I agree that in the Bayesian inference there is just one kind of error – the modelled error, errors that you've accommodated correctly, the statistical error from the fluctuations, that is to say uncertainties due to randomness in the realisation of the data. Our Bayesian error bars measure this error very well. The focus in Bayesian inference is on fitting a parametric model – so model misspecification errors, biases

caused by fitting the wrong model, are not expressed in the error bars we report. One of the things I've found interesting about the physicists' contribution to systematic errors is that they attempt to report them in a rather explicit way. I like the way you often see in physics papers  $\pm x$  followed by  $\pm y$ , the second being an attempt to quantify uncertainty due to variation in the model. You can formalise this model-error by fitting a larger class of models, but that often isn't computationally feasible. So the physicists' approach of simply having a go – considering at least the obvious and physically important modes of model variation – is a lot better than simply ignoring the problem.

**Rajendran Raja** I would like to speak about the distinction between the systematic errors and the statistical errors. In an experiment there are quantities which have a certain frequency of occurrence. Some of them are longer lived than others. So depending on how long the experiment lasts, some things will be systematic. For example, if CDF lasted a hundred years the luminosity error will be statistical because the luminosity errors would change many times during that time, but if it lasted a few years the error on the luminosity will be systematic because it will have one value during that interval of time. That's something that I haven't seen discussed, as to when errors become systematic as opposed to statistical, depending on the timescales involved.

**Louis Lyons** Can I encourage members of the panel to express a view on Question 2, about parameter intervals? When we estimate some parameter and get some range for the parameter, what properties would we like these intervals to have? Has anybody got an opinion on that?

**Bernard Silverman** They should have the properties they are claimed to have! If they are confidence intervals, they should have frequentist coverage.

**Louis Lyons** OK but we could widen the question a little bit so it didn't necessary say confidence intervals but rather any intervals. Is coverage an overriding feature? How unhappy would we be about the method that gave empty intervals or sometimes very very short intervals? What should we aim to do when we are investigating methods for producing confidence intervals?

**David Cox** Assuming you are using a reasonably high level of confidence or posterior probability or whatever, any true value should lie within the interval most of the time – that would be my answer. Secondly there is the issue that in most cases, I think one wants not intervals but upper limits and lower limits separately. For instance the Poisson problem is very clearly a situation where you can formally do a good job for the upper limit, but perhaps all you could possibly say about a lower limit is that it could perfectly well be zero. So there's that aspect.

No empty intervals? Well think one must accept empty intervals, in certain situations, because if the confidence interval, or a posterior interval, is a list of those parameter points that are reasonably consistent with the data and the model, the answer may be that no value is consistent with the data.

Then there is a complementary problem where the confidence set is the whole space, and any parameter value is consistent with the data. Again you are making a statement that's trivially true. I don't see the difficulty with that as a formalization of what the data imply.

**Bernard Silverman** Would you be worried by a very short interval? There is a danger here; you could have a very short interval because the model didn't really work and there was only a very small parameter set that fitted, so it's a sort of limiting case of the empty interval. That's scary because people would interpret that to mean we have estimated this parameter with very great accuracy, where what we've actually said is the model doesn't really fit but there is a very small range where it just about OK.

**David Cox** Yes but there's qualitative prior knowledge involved in judging what is small and that prior knowledge has to be used, if only informally.

**Nancy Reid** I think there's been a little bit more emphasis than necessary in the physics literature on exact coverage, that seemed to be bordering on an obsession! Coverage refers to this property in the long run over a whole lot of experiments and something that's been argued about in statistics over many many years is how many of those conceptual experiments are relevant to the one you have. That was laid out most clearly by David in 1958 with the two measuring instruments problem. I think that, in some cases, you are almost duplicating the two measuring instruments problem, by using the Neyman construction to get intervals that are guaranteed to cover in such a wide variety of situations that you're losing for the particular situation that you are going to use it. There's lots and lots of literature on this and it's not an easy literature to study but I think you'd be well advised to consider a little bit more the more pragmatic view that's been expressed by David and Bernard. It came up in the question on blinding: "If I look at the data then I'm going to ruin my coverage." But that coverage refers to a whole lot of perfectly carried out experiments where nothing weird happened, and you have a different experiment where something weird has happened, so that coverage is not really relevant in that situation. There's obviously a tension because every experiment is in some sense unique but we are talking about statistics so we have to average over something. So there is a tension between the two and it's not easy to resolve, but I don't think the right resolution is to average over everything.

**Sergei Bityukov** I want to express my opinion about confidence intervals. If we have a procedure which allows us to construct intervals, we can also construct the confidence density, and that contains more information.

**Bruce Yabsley** Just on the question of empty intervals and why we might be concerned about them: It's my impression part of our problem with statistical methods is that we use them for non-statistical purposes as well as for statistical purposes. You put a statement in a paper, and rather than just saying "Our confidence interval is such and such" or "Our upper limit is such and such", it tends to be overloaded with what I'm going to call (in a non-technical sense) sociological claims like "We have observed something" or "It's not there". This creates a serious problem with upper limits [in the case where there's a small excess over the expectation for background only] because if we are not absolutely confident that we've seen a signal, people want to quote an upper limit, even if they're quoting that limit at 90% confidence, which is the convention in the field. So you might have a weak signal (where a 90% interval in a unified approach excludes zero, but (say) a 99.7% interval includes it) but still want to quote a 90% upper limit: we return to this business of flip-flopping that Gary Feldman and Bob Cousins fixed. People throw away the solution, i.e. a unified approach to interval-setting, because they're nervous about what a two-sided interval would imply. And so you get a situation where something has a perfectly clear statistical meaning — 90% of the time the real value will be in the interval and 10% of the time it won't — but people aren't willing to stop there. So returning to the case of an empty interval, it will be taken as saying "We're confused, maybe our model was wrong or maybe the data is discrepant or we just don't know." I think it's very hard to imagine a physics collaboration actually writing that in a paper, even though it might be perfectly valid as it stands.

**Bob Cousins** I think my opinions are pretty well advertised on most of these so I'll just comment on the shortness issue. For two-sided intervals, 'shortness' of course is a metric-dependent statement. I think the way you want to look at it is that the coverage of confidence intervals is just a statement about one type of error and you do need to worry about the other type of error. So you want the most powerful intervals against alternative hypotheses. People know that this is not generally possible for all alternative hypotheses.

For a discrete observable you can make the acceptance region shortest in the construction direction, but that

is not necessarily the same as in the parameter direction. Crow and Gardner did this many years ago, but it seems not to be popular.

**Louis Lyons** I think that actually points to a difficulty if you are trying to decide between which of two methods you are going to use, before you look at the data of course, and you want a method that gives intervals which are not too long but not too short. It's not quite obvious what criterion you would use to optimise the choice of interval.

**David Cox** In the case of empty intervals the data are sending the message that something is wrong and of course if at all possible the source of the confusion must be identified.

**Bob Cousins** I'll just repeat something I said at the first Confidence Limits Workshop at CERN in 2000. I am not sure that statisticians are aware of it, but in Particle Physics we have a sort of thing that goes under the name of robustness. The Particle Data Group does lots of averaging of several measurements of the same parameter. When these are inconsistent with each other and have a large  $\chi^2$ , then the error on the weighted average is not determined simply by error propagation, but is scaled until the  $\chi^2$  per degree of freedom becomes something reasonable like unity. Of course it's like all robustness things in that it's kind of a black box that may or may not fix the actual problem, because you don't know what the actual problem is, or otherwise you'd fix it. In a 1999 paper Mike Chanowitz wrote on Higgs mass constraints, the input data from LEP and SLAC had some discrepancies, so he suggested blowing up the errors in an analogous way before producing a combined constraint on the Higgs mass.

One thing I thought was at least worth exploring five years ago, and I don't know it has ever been done, is to take that error on the background we were talking about and blow it up until there is a reasonable probability that you got the data you got. Then you quote an upper limit, including that larger error on the background. As with the PDG method, you don't really know if this is solving the problem, but I think it is worth exploring.

**Louis Lyons** Maybe at this stage we could ask members of the audience if any of them wanted to comment on any of the issues here.

**Bangalore Sathyaprakash** This is really a question rather than a comment. We are looking for signals that are weak and rare. When you're really not expecting very strong signals there is probably not much point in debating about which methods we follow; frequentist or Bayesian, does it really matter? We would certainly learn a lot from particle physicists who have had this experience for generations so I would like to see some discussion of that.

What I'm asking is this: let's suppose we are looking for a specific type of signal in a time series. We could follow either a frequentist approach and do an analysis wherein we try to evaluate the likelihood and either claim a detection or set an upper limit. Alternatively, you could follow a Bayesian approach and assume a prior – we know nothing about the prior – and then follow that procedure and get an upper limit. These two upper limits are different but does it really matter? Should we really quarrel about it? Should we not worry about detecting rather than setting upper limits?

**Bob Cousins** I've been talking about this with my colleagues at the Large Hadron Collider. There has been an enormous amount of interest over the last fifteen years or so, including the Confidence Limits Workshops, on how we measure upper limits. I do look forward to having signals to worry about. We can take as one example the statistics issues at the TeVatron with the discovery of the top quark, and how CDF and D0 measure its mass. I'll also just mention that Gary Hill gave a talk here on the difference between optimising

for upper limits and for discovery.

**Louis Lyons** I wanted to add that at the Fermilab conference in 2000, Ilya Narsky investigated what you get for upper limits by analysing the same data, using the number of events seen and the expected background, with a whole series of different methods. He's got an interesting plot that compares all these different methods. What becomes clear is that you can get very different upper limits by varying your technique and in some cases you can get variations by a factor of 10 or more, especially when the number of observed events is less than the estimated background. So it's a very good idea to choose your method before you look at the data, rather than tuning up the method that gives you the tightest or the weakest upper limit, depending on your feelings about things.

**Harrison Prosper** I'm one physicist who is not quite so obsessed with coverage. Certainly I'm very happy, and I also insist, that if one invents a method one should at least see whether it works well on average in some ensemble. But the crucial thing to realise is that this is a hypothetical ensemble. Our real experiment presumably is embedded in some ensemble, but we don't know what it is. We make some assumptions, for example we assume that things are perfectly Poisson, but presumably that's not exactly true; things are Poisson to some degree. So I'm happy to have approximate coverage in an ensemble, which is, after all, hypothetical. In fact such a situation arose in a collaboration of which I'm a member. We had two analyses measuring the top quark mass. The question arose as to whether it was sensible to choose the better of the two answers. We looked at the error that was computed for each analysis and asked whether or not we should use the analysis that gave the smaller error. As I noted then, that means inventing some ensemble in which we decide to toss a coin and choose which answer to report. The point is that you can invent any number of ensembles that are plausible and for each of these you'll get different coverage. This is why I'm not quite so obsessed with exact coverage.

**Jeremy Lys** We're missing out a lot here on the sociology as Bruce referred to before, when Harrison asks whether it is better to use one method rather than the other. You can ask "Better in what sense?" Many of us are experimental physicists, we know we are in competition with other people with other groups, we are in competition for our livelihoods in a sense, we are in competition for grants and so on and it's essential that we appear to be doing good physics and hence it's better to get what we call an accurate answer rather than an inaccurate one, so it's clearly tempting for us to change the way we define coverage for example. If you release the conditions of coverage and get a better answer then you might say that you should publish that result, maybe not. There are sociology points that arise here, that's all.

**Rajendran Raja** I would like to harp back to the Durham conference which was what got me into this whole thing. It's not just coverage that's important, but the stability of the limit in an ensemble of similar experiments is also important. The degree of fluctuation of a one-sided limit is important to compute and quote. We had this problem when we were looking for the top quark, with CDF publishing one limit and D0 was publishing another limit. The question was which was better. They were within 20 GeV of each other, but if you change the accepted sample by one event by changing the cuts slightly, the limit fluctuates by about 20 GeV. So it's important not only what the actual value of the limit is but also what the band of fluctuation is. Until we see an analysis along those directions, we'll be preoccupied with the coverage, and we won't get the full picture.

**Bob Cousins** Back to these criteria for parameter intervals, maybe I've said it 20 times but I think it's important that, when an interval is associated with some probability  $P$ , there exists a definition of what  $P$  is. For confidence intervals  $P$  is defined as frequentist coverage. Harrison's point is that you need a well-defined ensemble in order to define frequentist coverage, and that brings us back to points both of us made at the first Confidence Limits Workshop. Professional statisticians seem to go in the direction of conditioning, and that

is something we should look at.

The other way of defining P which I think is quite useful is subjective degree of belief. Real Bayesians, like Michael Goldstein at the Durham Conference, are comfortable with that P. What I have trouble with are intervals in between which come out of so-called objective Bayes, where you don't know what the P is: it's not subjective degree of belief and it's not necessarily frequentist coverage. Therefore I think that when we use the machinery of objective Bayes, the P we want to teach our students to apply is the frequentist P, and then check that it works. Joel Heinrich gave an outstanding talk on this in one of the parallel sessions yesterday, using the Bayesian machinery to obtain intervals which undercovered by the criterion of frequentist P, and he figured out how to change priors so that the result was more consistent with frequentist expectations. This is quite useful, but we should make sure that our students understand that the result of all this Bayesian machinery is not a P which is subjective degree of belief.

**Byron Roe** Improbable events do happen and sometimes your empty interval or almost empty interval tells you have an improbable event, even if you are confident in your parameterization. There are certainly famous examples, as Bob well knows. My question is really for the statisticians. Suppose you have an improbable distribution and by various means you really know it's an improbable distribution. Do you have any general ideas for what should be done to set limits on parameters?

For example, suppose you are measuring signal plus background and you know the mean value of your background well, and you get an observation which is much lower than the background. You know something strange has happened to you. My question is: "In this kind of situation can you make suggestions as to what one might want to do?"

**Unknown** Get a better estimate of your background!

**Bernard Silverman** This is a bit like the story about Mendel and the beans, and counting number of plants of different kinds in one particular experiment. If the theory is true, then you expect three of one sort to one of the other. Mendel's published data is at the wrong end of the  $\chi^2$  distribution. In other words it fits far better than you would expect at random. So you could then, I suppose, reject random selection. The interesting thing is that you get results which appear to challenge the very randomness assumption sometimes. There was an explanation: it was suggested that Mendel had a gardener who was adjusting the results to what his boss wanted to get! Fortunately, the experiment was correct anyway but there was a little fiddling going on. But I think that you might look for some reason for that low value. In other words, interact with your data intelligently and say "Maybe we're spotting something that we weren't expecting to see, perhaps it's some other phenomenon". I don't think you can get a statistical answer to the question; you just need to know it will happen sometimes.

**Louis Lyons** That brings us to coffee time, so let's close the session. Thanks to all of you who contributed to this session, and especially to our Panel members: David, Bob, Jerry and Bernard.