Introduction to Frequentist and Bayesian Fitting

F. James
CERN, Geneva, Switzerland

A brief introduction to Bayesian and frequentist statistical analyses, emphasising the differences between the two.

Frequentist Probability

Frequentist *statistics* is based on frequentist *probability*, which is defined as a *limiting frequency*, over a set of (hypothetical) repetitions of the same experiment.

$$P(A) = lim_{N \to \infty} \frac{N(A)}{N}$$

where A occurs $N(A)$ times in N trials.

Frequentist probability is used in most scientific work, because it is *objective*. It can (in principle) be determined to any desired accuracy and is the same for all observers.

It is the probability of Quantum Mechanics.

It can only be applied to repeatable phenomena.

Frequentist Fitting

An important principle in *frequentist statistics* is that you should always be able to calculate *how often you are wrong*:

- How often the quoted error bar does not include the true value.
- How many unwanted events will pass the cuts and contaminate the sample.
- How many good events will be rejected by the cuts and reduce the signal.
- How often the Chisquare Test will fail even if the hypothesis is correct.

The hard part of frequentist statistics is to find methods such that the probability of being wrong is *independent of the value(s) of any unknown parameter(s)*.

For confidence intervals, J. Neyman (with the help of Gary Feldman and Bob Cousins) showed us how to do this.

Bayesian Probability

Bayesian Statistics is based on the *Bayesian definition of Probability* which is a *degree of belief.*

The operational definition most often used is de Finetti's *coherent bet.* Your belief in A is proportional to the amount you would bet on A happening.

This allows you to define the probability of non-repeatable events.

In particular, the *prior probability* of A is your belief in A *before you do any experiments* to measure A.

Bayesian Fitting

All Bayesian methods require as input a *prior probability.* This is a kind of *phase space*: All experimental results are multiplied by the *prior* to obtain the *posterior probability.*

Concerning the *prior*, there is some good news and some bad news:

1. The *bad news* is:

   - Unlike physical phase space, there is no principle that tells you what *prior probability* to use. It is arbitrary.
   - No matter what your experiment tells you, it is always possible to find a prior that gives *any result you want.*

2. The *good news* is: The prior becomes less important as the amount of data increases. However, even this has to be taken cautiously because:

   - The limit in which the data dominate the prior is anyway the limit in which also frequentist methods give the same results.
   - When the hypothesis is multidimensional (several parameters estimated simultaneously), it cannot be shown that the data must dominate the prior.

So Why would any Physicist Use Bayesian Methods?

Bob Cousins [Am. J. Physics **63**, 398 (1995)] gives a good analysis why physicists mostly use frequentist statistics.

But some physicists are Bayesian. Why?

- Statisticians consider Bayesian methods valid, and in some sense even more coherent than frequentist methods.
- Bayesian has become "trendy", "modern".
- Bayesian is easier, more natural.
- The *Karmen Problem.*

The Karmen Problem before Feldman-Cousins

The Karmen Problem is the observation of more events (signal plus background) than are expected from background alone, making the apparent observed signal negative. Before the work of Feldman and Cousins, the standard frequentist upper limit in this case was given by the table below, which shows that it could be negative. This method would still have overall exact coverage, since the probability of such a result is very small.

Frequentist 90% Upper Limits for Poisson with Background

observed =	0	1	2	3
background = 0.0	*2.30*	3.89	5.32	6.68
0.5	*1.80*	3.39	4.82	6.18
1.0	*1.30*	2.89	4.32	5.58
2.0	*0.30*	1.89	3.32	4.68
3.0	*-0.70*	0.89	2.32	3.68

The naive Bayesian upper limit (with uniform prior) for the same case gave just what the physicist wanted:

Bayesian 90% Upper Limits (Uniform Prior)

observed =	0	1	2	3
background = 0.0	*2.30*	3.89	5.32	6.68
0.5	*2.30*	3.50	4.83	6.17
1.0	*2.30*	3.26	4.44	5.71
2.0	*2.30*	3.00	3.87	4.92
3.0	*2.30*	2.83	3.52	4.37

The Karmen Problem after Feldman-Cousins

Feldman and Cousins discovered that we had not been applying frequentist statistics in the correct way, and it should be done using the maximum likelihood ratio ordering principle, which would assure a *unified approach* to upper limits and two-sided limits, while at the same time eliminating negative upper limits.

Feldman-Cousins 90% Upper Limits for Poisson with Background

observed =	0	1	2	3
background = 0.0	*2.44*			
0.5	*1.94*	3.86		
1.0	*1.61*	3.36	4.91	
2.0	*1.26*	2.53	3.91	5.42
3.0	*1.08*	1.88	3.04	4.42

At the same time, people questioned the uniform Bayesian prior, and found the the Jeffreys prior had a better theoretical base and corresponded better to actual prior belief. Unfortunately, the Jeffreys prior gives:

Bayesian 90% Upper Limits (Jeffreys Prior)

observed =	0	1	2	3
background = 0.0	*0.00*	2.30	3.89	5.32
0.5	*0.00*	0.00	0.00	0.00
1.0	*0.00*	0.00	0.00	0.00
2.0	*0.00*	0.00	0.00	0.00
3.0	*0.00*	0.00	0.00	0.00

Therefore, the Karmen problem, or more generally the problem of getting reasonable upper and lower limits (with correct frequentist coverage) in delicate cases with very small numbers of events, is no longer a reason to adopt Bayesian statistics. On the contrary, recent advances show the exact frequentist methods to be closer to what the physicist wants.

What can go wrong with Bayesian analyses?

One of the most striking examples of misfunctioning of data analysis due to the use of Bayesian principles can be found in the paper "Evidence for neutrinoless double beta decay", by Klapdor-Kleingrothaus et al, submitted to World Scientific on May 20, 2006. This research article finds strong statistical evidence for a peak where many physicists don't see any peak. One of the criticisms of the analysis was that the authors did not do a standard goodness-of-fit test for the no-peak hypothesis to see whether the data required a peak. But the standard chi-square test is not allowed in the Bayesian framework because it violates the Likelihood Principle. The explanation given by the authors was as follows:

Criticism: There is no null hypothesis analysis demonstrating that the data require a peak.

Reply: The statement that there is a peak with probability K implies that there may be none. Since the results are probabilistic, it is not possible to demonstrate that the data do require a peak.

Note that the model used to analyze the data includes the possibility that the intensity of the line is zero. The error intervals given in KDHK have been chosen such as to exclude the value of zero. Therefore every result in KDHK can especially be read as: "With probability K the intensity zero is outside the error interval". In this sense the null hypothesis is rejected with the probability on which the error interval is based.

The difficulty we have in understanding this reply makes us question whether Bayesian methods are really so natural and intuitively appealing.

Conclusions: Bayesian or frequentist?

1. The main problem in Bayesian methodology is the *prior*. Use Bayesian methods when you *know the prior* and have a good reason to use it. The only case I know where that is true is *maximum entropy image processing.*

2. Use Bayesian decision theory to make it clear what are the subjective criteria for your decision. [Example: where to look for new physics.]

3. For everything else, in particular objective data analysis, I don't see any reason to use Bayesian methods. We now know how to handle all the situations (nuisance parameters, systematic errors) that used to cause problems in the frequentist methodology.

4. Very few people would believe a result that can only be obtained by a Bayesian analysis with an arbitrary prior.