RHCPP 01–02 (CONF)

Statistical Issues in Heavy Flavour Physics†

Glen Cowan
Physics Dept., Royal Holloway, University of London, Egham, Surrey, TW20 0EX

Abstract. Two statistical questions relevant to heavy flavour physics are discussed. First, the problem of how best to combine measurements of branching ratios of rare decays is treated using the method of maximum likelihood. Second, the general question of how to quantify theoretical uncertainties is examined.

1. Introduction

In this paper two statistical issues are addressed that come up in many physics analyses but which are particularly relevant for heavy flavour physics. The first, presented in Section 2, is the problem of combining measurements of the branching ratios of rare decays where some of the measurements are based on only a few or perhaps no candidate events. The solution recommended is a straightforward application of the method of maximum likelihood. The next question, discussed in Section 3, is much more open-ended, namely, how to quantify theoretical uncertainties. Both issues are of course not unique to heavy flavour physics, and the approaches recommended can be applied in many other contexts.

2. Combining branching ratios for rare decays

Consider first a single experiment in which one searches for rare decay and finds n candidate events out of N total decays. Suppose the efficiency for the signal is ε and there are b expected background events. In general, n should be treated as a binomially distributed variable where N represents the total number of trials and the binomial probability $\mathcal{B}$ is the branching ratio. If $\mathcal{B}$ is very small and N very large, however, we can treat n as following a Poisson distribution with expectation value ν, i.e.,

$$P(n;\nu) = \frac{\nu^n}{n!} e^{-\nu} \ . \tag{1}$$

Here the number of events observed n is the sum of n_s signal and n_b background events and the expectation value of n, here called ν, is also the sum of the signal and background expectation values s and b:

† Contribution to the Proceedings of the UK Phenomenology Workshop on Heavy Flavour and CP Violation, 17 – 22 September 2000, St John's College, Durham; to appear in Journal of Physics G.

$$n = n_s + n_b \ , \tag{2}$$

$$\nu = s + b \ . \tag{3}$$

The branching ratio $\mathcal{B}$, which is the parameter we want to estimate, is related to the expected number of signal events s, the efficiency ε and the total number of events N by

$$\mathcal{B} = \frac{s}{\varepsilon N} \ . \tag{4}$$

For a single experiment, the likelihood function $L(\mathcal{B})$ is simply the Poisson probability (1) evaluated with the observed number of events n and expressed as a function of the branching ratio $\mathcal{B}$. Maximizing L gives the maximum likelihood (ML) estimator $\hat{\mathcal{B}}$,

$$\hat{\mathcal{B}} = \frac{n - b}{\varepsilon N} \ , \tag{5}$$

where here estimators will be denoted by hats. The statistical uncertainty in $\hat{\mathcal{B}}$ can be quantified by giving an estimate of its standard deviation,

$$\hat{\sigma}_{\hat{\mathcal{B}}} = \frac{\sqrt{n}}{\varepsilon N} \ , \tag{6}$$

or by reporting a confidence interval (see e.g. [1]).

Now consider the case where we have m independent measurements of the branching ratio $\mathcal{B}$. The m experiments report $n_1, \ldots, n_m$ candidate events out of $N_1, \ldots, N_m$ total decays; they have signal efficiencies $\varepsilon_1, \ldots, \varepsilon_m$ and expected numbers of signal and background events $s_1, \ldots, s_m$ and $b_1, \ldots, b_m$. Assuming the measurements are statistically independent, the likelihood function is the product of the Poisson probabilities for each,

$$\begin{aligned} L(\mathcal{B}) &= \prod_{i=1}^{m} \frac{\nu_i^{-n_i}}{n_i!} e^{-\nu_i} \\ &= \prod_{i=1}^{m} \frac{(\varepsilon_i N_i \mathcal{B} + b_i)^{n_i}}{n_i!} e^{-(\varepsilon_i N_i \mathcal{B} + b_i)} \ . \end{aligned} \tag{7}$$

The value of $\mathcal{B}$ that maximizes (7) is the ML estimator $\hat{\mathcal{B}}$, which in practice will need to be determined numerically, e.g., by solving

$$\frac{\partial \ln L}{\partial \mathcal{B}} = \sum_{i=1}^{m} \left[\frac{n_i \varepsilon_i N_i}{\varepsilon_i N_i \mathcal{B} + b_i} - \varepsilon_i N_i \right] = 0 \ . \tag{8}$$

As in the case of a single measurement, the statistical uncertainty in $\hat{\mathcal{B}}$ can be quantified by estimating its standard deviation using, for example, the curvature of the log-likelihood function at its maximum,

$$\hat{\sigma}_{\hat{\mathcal{B}}} = \left(- \left. \frac{\partial^2 \ln L}{\partial \mathcal{B}^2} \right|_{\mathcal{B} = \hat{\mathcal{B}}} \right)^{-1} \ , \tag{9}$$

or by constructing a confidence interval.

As the final result must be determined numerically it is difficult to see what the properties of the ML estimator are. If, however, we consider the case where there is no background, i.e., $b_i = 0$ for all i, then the estimator $\hat{\mathcal{B}}$ takes on the simple form

$$\hat{\mathcal{B}} = \frac{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} \varepsilon_i N_i} \, , \tag{10}$$

for which the standard deviation can be estimated by

$$\hat{\sigma}_{\hat{\mathcal{B}}} = \frac{\sqrt{\sum_{i=1}^{m} n_i}}{\sum_{i=1}^{m} \varepsilon_i N_i} \, . \tag{11}$$

For this case, it is easy to see that there is no problem if some of the n_i are small or even zero. Such problems would arise, for example, in least squares averaging of individual estimates of $\mathcal{B}$. Furthermore, in the real case with nonzero backgrounds, the ML estimate is not spoiled by including experiments that carry less information than others because, for example, they have larger backgrounds or are based on smaller data samples.

It should not be controversial that the ML estimator has many advantages and should be used unless there is a good reason not to. The important point to note here is that it is not sufficient for experimenters to report only an estimate of the branching ratio and a standard deviation. In order to combine the measurements using the method of maximum likelihood, one requires the number of observed events, the efficiency, the expected number of background events and the total number of events for all experiments.

3. Quantifying theoretical uncertainties

In many experimental analyses, the dominant uncertainty in the determination of a parameter does not stem from measurement errors but from the approximations used in deriving the theoretical prediction for the quantities observed. An important example from heavy flavour physics is the theoretical prediction for the differential decay rates of B mesons into various final states. Although the Standard Model makes a well-defined prediction for the functional form of the probability distribution for measured decay times, the exact form of this prediction cannot be derived because of calculational difficulties, especially for multi-body final states. Instead, approximations are used, based, for example, on a perturbation series calculated only to some finite order. As a result, the estimated parameter values are in general uncertain even in the absence of measurement errors, statistical or otherwise. This is what is meant here by the 'theoretical uncertainty' in an estimated parameter.

Here we will argue that the Bayesian approach to statistics provides a potentially useful framework for quantifying theoretical uncertainties. To keep the discussion as general as possible, suppose we measure a vector of data $\mathbf{x}$, whose probability distribution function (pdf) is predicted as a function of a vector of parameters $\boldsymbol{\theta}$. For example, the vector $\mathbf{x}$ could represent a set of measured decay times for B meson decays into $\mathrm{J}/\psi\mathrm{K}^0_\mathrm{S}$ and the vector $\boldsymbol{\theta}$ would then be related to CKM matrix elements as well as other Standard Model parameters that may enter.

Now suppose that we do not know the exact form of the pdf $f(\mathbf{x};\boldsymbol{\theta})$ but the function that we have to work with is only an approximation. This approximation can be viewed as belonging to a family of functions, one of which is the true prediction of the theory, or at least close to it. Suppose we can parametrize this family by some vector of *nuisance parameters* $\boldsymbol{\mu}$. The elements of $\boldsymbol{\mu}$ could represent, for example, higher order coefficients in a perturbation series that are not included in the prediction actually available, but which could be calculated in principle. It is necessary in the procedure presented here for one of the members of the family of functions to be sufficiently close (in principle, equal) to the true prediction. The family can always be made arbitrarily large by including more nuisance parameters, but this will have the end result of inflating the uncertainties in the parameters of interest.

In Bayesian statistics, all knowledge about the parameters of a theory is summarized by giving the *posterior* pdf, $p(\boldsymbol{\theta},\boldsymbol{\mu}|\mathbf{x})$. This is the conditional pdf for the parameters given the data $\mathbf{x}$. In our case, the parameters consist of those in which we are interested, $\boldsymbol{\theta}$, as well as the nuisance parameters $\boldsymbol{\mu}$. According to Bayes' theorem, they are related to the likelihood function $L(\mathbf{x}|\boldsymbol{\theta},\boldsymbol{\mu})$ and the *prior* pdf $\pi(\boldsymbol{\theta},\boldsymbol{\mu})$ by

$$p(\boldsymbol{\theta},\boldsymbol{\mu}|\mathbf{x}) \propto L(\mathbf{x}|\boldsymbol{\theta},\boldsymbol{\mu})\pi(\boldsymbol{\theta},\boldsymbol{\mu}) \ . \tag{12}$$

The likelihood function $L(\mathbf{x}|\boldsymbol{\theta},\boldsymbol{\mu})$ is the conditional probability for the data given the parameter values, and the prior pdf $\pi(\boldsymbol{\theta},\boldsymbol{\mu})$ summarizes our knowledge of the parameters before carrying out the experiment.

It should be emphasized here that the probability associated with the value of a parameter cannot be interpreted meaningfully as a frequency of an outcome of a repeatable experiment. Instead it is understood to reflect the degree of belief that the parameters have particular values, an interpretation called *subjective* probability. In order to determine this degree of belief given the data, we need to say what it was before carrying out the experiment, i.e., we need the prior pdf $\pi(\boldsymbol{\theta},\boldsymbol{\mu})$. Bayesian statistics provides no fundamental prescription for determining the prior pdf. Given $\pi(\boldsymbol{\theta},\boldsymbol{\mu})$, however, Bayes' theorem specifies how our knowledge about the parameters should change in the light of the data.

The usefulness of the Bayesian approach in the present problem will depend on our being able to parametrize the pdf $f(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\mu})$ in terms of the nuisance parameters $\boldsymbol{\mu}$, and to make a reasonable statement about their probabilities. Often this means interrogating the theorist to the point where he or she tells you how large, say, the next order coefficient in a perturbation series could be. At first, the theorist may say something like 'Until I've calculated it, all values are equally likely'. Under pressure, however, it will be admitted that if it were to differ by many orders of magnitude from the previous term, or from one's naive expectation of the value based on past experience with perturbation theory or Padé approximants or some other procedure, then this would be surprising.

The key now is to quantify this level of surprise, by *calibrating* it against something for which probability can be defined in terms of a frequency. The theorist should state the range for the nuisance parameter such that the degree of surprise at finding out that it is outside this range is equal to, say, the level of surprise that one would experience at drawing a white ball out of an urn of balls, 32% of which are white. In most cases this level of surprise will be a smoothly varying function of the parameter, which could be represented by some sort of a bell-shaped curve like a Gaussian. It would be rare, for example, that a uniform distribution between fixed limits would provide a reasonable model for quantifying the uncertainty in such a parameter. Assuming a Gaussian model, the central interval containing 68% of the probability corresponds to the one-sigma error. In Bayesian parlance these are often called 68% *credible* intervals. More information on subjective probability and Bayesian statistics can be found in [2]; its application to problems of data analysis in physics is treated at various levels of detail in [3, 4, 5, 6].

It is clear that this calibration of a degree of belief will never be perfect, and we should be happy if it is reasonable to within a factor of two or even a factor of ten. In principle the level of calibration could be measured by forcing theorists to guess the 68% credible intervals, waiting long enough for the calculations to be made and then determining the frequency with which they are right. An alternative proposed at this workshop would be to force them to bet (large) sums of money on the values of nuisance parameters that at some point will be calculated. The Bayesian framework will be useful if reasonable prior probabilities can be assigned; there are no doubt cases in which this is not possible.

The importance of making reasonable statements about prior probabilities holds as well for the model parameters $\boldsymbol{\theta}$, not only the nuisance parameters $\boldsymbol{\mu}$. Here the posterior will be insensitive to the prior pdf for $\boldsymbol{\theta}$ as long as it is a more slowly varying function of $\boldsymbol{\theta}$ than is the likelihood function. If this is not the case, then the Bayesian approach is less likely to yield useful results, since the posterior pdf will be determined more by the prior beliefs than by the data.

Providing one can quantify the necessary prior probabilities, it is a straightforward exercise to apply Bayes' theorem to obtain the posterior pdf $p(\boldsymbol{\theta}, \boldsymbol{\mu}|\mathbf{x})$. We are then in a position to apply one of the most powerful tools of Bayesian statistics: we eliminate the nuisance parameters $\boldsymbol{\mu}$ from the problem simply by computing the *marginal* pdf $p(\boldsymbol{\theta}|\mathbf{x})$, i.e.,

$$p(\boldsymbol{\theta}|\mathbf{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{\mu}|\mathbf{x})\, d\boldsymbol{\mu}\ . \tag{13}$$

One can then summarize the posterior pdf by, for example, presenting contours of constant probability density in two-dimensional subspaces or by characterizing it by various sorts of location parameters such as the mode. Regions in the parameter space can be determined which contain specified probabilities, e.g., 68%, 90%, etc.

The sizes of the contours of constant probability density reflect both the statistical and systematic errors. These two components of the uncertainty can be separated by constructing estimators $\hat{\boldsymbol{\theta}}$, e.g., by taking the mode of the posterior pdf. These will be specific functions of the data, and as such they are random variables in the sense of frequentist probability. One can then ask, for example, how $\hat{\boldsymbol{\theta}}$ would be distributed upon repetition of the experiment, and in this way the statistical and systematic components can be separated, if desired.

Although up to now we have only taken $\boldsymbol{\mu}$ to quantify errors due to approximations made in deriving theoretical predictions, it can just as easily represent other types of systematic uncertainties. In either case, the assumptions made in quantifying the uncertainties should clearly be reported along with the result. It should be emphasized that the posterior pdf by itself does not in general provide a useful summary of the result of the experiment, as it combines both the data and the subjective beliefs needed to construct the prior pdf. If only the posterior pdf is reported, it is impossible for a consumer of the result to separate the influence of the experimenter's prior beliefs from the influence of the data.

The experimenter should therefore separate as well as possible a summary of the observations from their interpretation. In a paper, for example, this may be done in sections titled 'results' and 'discussion' or 'interpretation'. In the section on results one can report the values of observables that to the greatest reasonable degree are free of subjective judgements, theoretical or otherwise. For example, one should report the various numbers of events observed that satisfy specific criteria, and not only the estimates of parameters that are derived from these data. In the interpretation of the results, however, there is no reason for experimenters to shy away from making reasonable (stated) assumptions based on subjective probability and showing what they imply for the values of parameters of interest.

4. Conclusions

A straightforward technique based on the method of maximum likelihood has been presented for combining measurements of branching ratios. In order to employ this method, experimenters must report the number of candidate events seen, the efficiency of the selection procedure, the expected background and total number of decays observed. It cannot be used in general if only an estimate of the branching ratio and its standard deviation are reported.

The use of Bayesian statistics in quantifying systematic errors, in particular those due to uncertainties in the functional form of theoretical predictions, has been discussed. In this framework one must quantify the uncertainty of unknown constants using subjective probability, i.e., probability based on a degree of belief. The method is useful to the extent that the uncertainty can be parametrized and quantified, e.g., by interrogation of the theorist responsible for the prediction. Owing to difficulties in calibrating one's degree of belief relative to frequentist probabilities, this is necessarily an approximate procedure and all assumptions made in quantifying the uncertainty should be reported along with the result.

Acknowledgements

It is a pleasure to thank the organizers of the Durham workshop, especially Nicolo de Groot, Mike Whalley and Adrian Signer, for an enjoyable and fruitful meeting. I also thank David Summers for stimulating discussions about theoretical uncertainties.

References

[1] Roger Barlow, *Confidence Levels for Pedestrians*, in these proceedings.

[2] A. O'hagan, *Kendall's Advanced Theory of Statistics*, Vol. 2B, *Bayesian Inference*, Edward Arnold, London (1994).

[3] W.T. Eadie, D. Drijard, F.E. James, M. Roos and B. Sadoulet, *Statistical Methods in Experimental Physics*, North-Holland, Amsterdam (1971).

[4] G. D'Agostini, *Bayesian reasoning in high-energy physics: principles and applications*, CERN 99-03 (1999).

[5] D.S. Sivia, *Data Analysis: a Bayesian Tutorial*, Oxford University Press (1996).

[6] G. Cowan, *Statistical Data Analysis*, Oxford University Press (1998).