The Last Fifty Years of Statistical Research and their Implications for Particle Physics

D.R. Cox
Nuffield College, Oxford, England

1 Introduction
The title may be interpreted in at least two ways. The last fifty years, although drawing heavily on earlier work, have seen both the development of a large number of particular statistical techniques and also their general availability achieved through relatively painless software packages. These procedures are very widely used in a great range of scientific and technological fields. In so far as these methods are based on probabilistic models of the data, the models tend to be broad descriptions of commonly occurring patterns of haphazard variability; the models relatively rarely contain an important specific subject-matter basis. One possible interpretation of the title is that within this vast mass of material lies the answer to at least some of the demanding statistical issues facing particle physics. Such a source is not the route taken in this paper. While the provision of such an armoury of methods can be claimed to be a massive contribution to science, it seems more fruitful to regard the problems of particle physics as ones to be tackled largely from first principles. See, however, the comments in this volume of Cox and Reid (2007) on some of the so-called statistical wishlist.

The emphasis in this paper is therefore largely on broad principles. I have recently (Cox, 2007) reviewed the nature of statistical considerations in scientific research, strongly emphasizing the need for unity of statistical and subject-matter thinking. In many fields a statistician involved in discussions of a research study will be the most quantitatively-minded member of the group. Clearly this is not the case in discussions with physicists with their very strong and highly successful tradition of independent mathematical thought, so that many of the points in my review hardly apply. The objective of the present paper is therefore partly to outline some general principles and partly, in order to be more specific, to describe in outline issues connected with discovery against a large essentially random background, including an account of false discovery rates.

2 Some methodological themes
A very broad classification of statistical concepts is as follows:

- ideas not specifically based on probability models, for example
  - clustering algorithms
  - visualization of multivariate data
- design of experiments, including simulation studies, and of sampling (data capture) procedures
- construction of probability models
  - families of empirical models of haphazard variation
  - substantive stochastic models
- statistical methods for analysis and interpretation of data in the light of a probability model
  - model checking
- formal theory of inference

Reid (2007) below discusses design of experiments, noting that the experiments discussed in the statistical literature are comparative studies, for example of a modified condition with a control, and are
thus unlike experiments in the present context, so that the ideas are of most relevance in particle physics
in the context of computer experiments.

The construction of special stochastic models, especially for dynamic problems, is now virtually
a special area of its own. In some countries, for example UK but not in US, it used to be firmly part of
the statistical field. It is unclear what the role of such work may be in the context of particle physics.

In the next section a few aspects of more formal theory are discussed before proceeding to more
specific matters.

3 An aspect of formal theory

The notion of probability used in formulating a statistical model for data is based on long-run frequency
under real or hypothetical repetition. It aims to capture the essence of the data-generating procedure.
Probability is used also in defining and interpreting the most appropriate method of analysis. There
are various ways in which this can be done, for example by significance tests, confidence limits and so
on or by posterior probabilities. This last usually involves a different notion of probability, namely as
in some sense a degree of belief. These two approaches are broadly called frequentist and Bayesian
respectively. Both ideas occur often in statistical discussions in particle physics and the following brief
note is intended to clarify the distinctions between them and, very importantly, the distinctions between
different interpretations of Bayesian analyses.

Key ideas of the frequentist approach were set out in very major papers by R.A. Fisher (1922,
1925). Neyman and Pearson, in a series of papers, reformulated Fishers ideas aiming, as they said,
for greater clarity. Some recent authors consider that while Neyman and Pearson certainly did achieve
mathematical clarity and introduced important new ideas, there was some loss of scientific relevance and
there has been some move towards Fisher’s original formulations. For many purposes these differences
are relatively minor; the key point is that statistical procedures are calibrated by their performance under
hypothetical repetition.

By contrast there are at least five quite different interpretations of Bayesian theory which makes
the use of the word Bayesian decidedly ambiguous if not actually confusing. The common link is the
mathematical one of using the basic laws of probability to pass from the probability of data given expla-
nation to the probability of the explanation given the data, accounting for the older and, in some ways
preferable, term inverse probability. This calculation requires the existence of and knowledge of proba-
bilities for the various explanations in the absence of the specific data under analysis, the so-called prior
distribution.

The five interpretations are in outline as follows:

– the prior distribution is a frequency distribution known or estimable from appropriate data. The
use of inverse probability is then uncontroversial, and all probabilities are frequentist. This may
be called empirical Bayes.

– the prior is intended to be neutral in the sense of introducing no or very little information about
the issue under study, leaving the data to supply that information. The idea goes back at least
to Laplace and has been developed in detail by Jeffreys and later Jaynes. Demortier (2007) has
described the latest thinking on this in a particle physics context, the contributions of Bernardo
(2005) to the notion of a reference prior being central; another major contributor is Berger as in
Berger (2007). Important contributions to a counting problem with noise have shown that some-
what casual choice of a flat prior can have very bad consequences. This approach may be called
objective Bayes.

– the prior distribution may be indirectly data-based, or at least evidence-based in some sense, and
provides a way of introducing into an analysis important additional information.

– the prior may encapsulate the opinions of a particular individual, usually called You, and as such
be the basis of Your decision making. There is no necessary implication for any other individual. This may be called the personalistic view.

- Bayesian calculations with a suitably standardized prior may be regarded purely as a convenient algorithm for finding procedures with good frequentist properties, no special notion of probability being needed.

It is important that these, while united by the mathematical techniques used, represent very different views. The first is entirely uncontroversial. The second pursues the same objectives as those of frequentist inference. The third offers the important possibility of incorporating additional information; note, however, that if this is directly based on empirical data, techniques for the combination of information may be used and these should include looking at the mutual consistency of the two sets of data. The personalistic approach is strongly focussed on personal decision making and, while it may be helpful to clarify personal thinking on an issue, it is not of clear relevance to the public discussions of scientific research and to the presentation of evidence for public discussion. The position over Bayesian calculations as algorithmic procedures with a frequency justification is not entirely clear. With a single unknown parameter very close matching between Bayesian and frequentist solutions is achieved with the Jeffreys prior. Such close matching is achievable with more than one parameter only in exceptional cases, although with modest numbers of parameters reasonable results may often be achieved.

Major issues arise when the number of parameters fitted is large relative to the amount of information available. Naive use of objective priors leads to very bad answers. Direct use of maximum likelihood may be misleading. Frequentist techniques for overcoming such difficulties stem from Bartlett (1937) and have been the subject of much recent work; for an account with many examples, see Brazzale, Davison and Reid (2007). Issues associated with models with many parameters are a challenge for all approaches to statistical inference.

4 Many hypotheses
4.1 General formulation
The remainder of the paper is concerned with much more specific issues. Suppose that data are available to test a large number $n$ of null hypotheses, each hypothesis may or may not be true. If only the smallest $p$-value is reported we are likely to be misled if in fact all null hypotheses are true, for example if the data are totally noise. This is a well-understood selection effect.

There are two distinct problems:

- some small but almost certainly nonzero number of the hypotheses are false and it is required to assess which those are
- it is quite possible that all the null hypotheses are true: how strong is the evidence against this on the basis of $m = \min(p_j)$

These two questions require quite different answers.

4.2 Selection of real effects
There are two broad approaches to this issue, one involving notional error rates and the other, probably the preferable one, empirical Bayesian in formulation. The former approach was first studied systematically by Schweder and Spjotvoll (1982). Suppose that $R$ of the null hypotheses are rejected of which $F$ are in fact rejected in error. Then the false discovery rate is defined as

$$E(F/R \mid R > 0).$$

Procedures are required to ensure that the false discovery rate does not exceed some specified limit (Benjamini and Hochberg, 1995; Storey, 2002). A very minor modification is to define the false rejection
rate \( F/R \) to be zero if \( R = 0 \). In a more elaborate version, \( F/R \) is regarded as a random variable which is required to be less than some specified limit with suitably high probability (Genovese and Wasserman, 2006).

For the second approach in its simplest formulation suppose that the \( n \) test statistics \( T_1, \ldots, T_n \) have under the respective null hypotheses the densities \( f_0(t) \), whereas under the alternative hypothesis in each case the density is \( f_1(t) \). Suppose further that a proportion \( \theta \) of null hypotheses are false.

In an important special case the two distributions are Gaussian distributions with unit variance and means zero and \( \mu_1 \) for the null and for the alternative hypotheses respectively. The unit variance under the null hypothesis is achieved virtually without loss of generality by definition of the test statistic. The unit variance under the alternative is a simplifying assumption which could be tested given sufficient data.

Then for any given \( t \) the posterior odds that the value comes from the alternative rather than from the null distribution are

\[
\log \frac{P(f_1 \mid t)}{P(f_0 \mid t)} = \log \frac{\theta}{1-\theta} + \mu_1(t - \mu_1/2).
\]

Thus, provided \( \theta \) and \( \mu_1 \) can be reasonably estimated, the posterior odds corresponding to any given \( t \) can be found. Numerical work suggests that at least for a preliminary analysis estimation of the two parameters from the first two moments of the test statistics gives good results (Cox and Wong, 2004). A threshold in \( t \) could be set to achieve a preassigned false recovery rate; a main advantage of the method, however, is that it attaches a measure of uncertainty to each value of \( t \) rather than merely giving a dichotomy.

A more elaborate approach (Efron et al, 2001) assumes that both densities \( f_0(t) \) and \( f_1(t) \) are unknown and need to be estimated nonparametrically. The emphasis of the discussion is rather different depending on whether the immediate interpretation of the analysis is important or whether a multi-step selection procedure is involved. In the latter those hypotheses chosen in step one are then tested more searchingly in a second or further stages.

### 4.3 Global null hypothesis

Suppose now that we consider \( m = \min(p_j) \) as the test statistic for the global null hypothesis that all individual null hypotheses are simultaneously satisfied. If the individual tests are independent the \( p \)-value allowing for selection is

\[
1 - (1 - m)^n
\]

and without the independence assumption

\[
mn
\]

is an upper bound, often sharp.

If \( n \) is large, for example of the order of \( 10^3 \), achievement of an interesting level of significance requires \( m \) to be extremely small. In particular this involves sensitivity to the assumptions involved in calculating the individual \( p_j \) to a degree that will often be quite unreasonable.

A possible solution, which has been used independently by Professor David Clayton in a genetical context, is as follows.

The first step is to produce a graphical summary of the \( \{p_j\} \) that will emphasize the \( p_j \) of most interest, namely the small values. For this write \( y_j = -\log p_j \) so that under the null hypothesis these have an exponential distribution with unit mean. Equivalently the \( 2y_j \) have a chi-squared distribution with two degrees of freedom.

Order the values in the form

\[
y(1) \geq y(2) \geq \ldots y(n)
\]
Under the global null hypothesis these have expected values

\[ 1 + 1/2 + \ldots + 1/n, 1/2 + \ldots + 1/n, \ldots, 1/n. \]

These are close to the quantiles of the unit exponential distribution.

Under the global null hypothesis and, assuming that the formal distribution theory is totally appropriate, a plot of the ordered \( y \) against their expected values should show a straight line of unit slope. Failure of one or a small number of the null hypotheses is shown by the final points being well above the line. On the other hand, provided that it is unlikely that many of the null hypotheses are false, failure in the tails of the underlying distribution theory is shown by the plot producing a smooth curve; outlying points can then be assessed as departures from this smooth curve.

Detailed discussion of the assessment of significance requires some further work. Under simple assumptions the ordered \( y_{(j)} \) considered as functions of \( j \) for \( j = n, n - 1, \ldots, 1 \) form a simple Markov process, in fact a nonstationary random walk.

References