TOPICAL REVIEW • OPEN ACCESS

Statistical issues in searches for new phenomena in High Energy Physics

To cite this article: Louis Lyons and Nicholas Wardle 2018 J. Phys. G: Nucl. Part. Phys. 45 033001

View the article online for updates and enhancements.
Topical Review

Statistical issues in searches for new phenomena in High Energy Physics

Louis Lyons$^{1,2}$ and Nicholas Wardle$^{1,3}$

1 Department of Physics, Imperial College London, London SW7 2AZ, United Kingdom
2 Particle Physics, University of Oxford, Oxford OX1 3RH, United Kingdom
E-mail: l.lyons@physics.ox.ac.uk and n.wardle09@imperial.ac.uk

Received 27 September 2016, revised 19 September 2017
Accepted for publication 17 October 2017
Published 23 January 2018

Abstract
Many analyses of data in High Energy Physics are concerned with searches for New Physics. We review the statistical issues that arise in such searches, and then illustrate these using the specific example of the recent successful search for the Higgs boson, produced in collisions between high energy protons at CERN’s Large Hadron Collider.

Keywords: High Energy Physics, precision measurements, parameter confidence interval

(Some figures may appear in colour only in the online journal)

1. Introduction

In 2012, the ATLAS and CMS experiments at the CERN Large Hadron Collider (LHC) announced that they had discovered the long-sought Higgs boson $H^0$. At about the same time, the existence of Higgs-like bosons of somewhat heavier masses was excluded, as were many other hypothesised particles in specified mass ranges (e.g. SUperSYmmetric (SUSY) particles with particular decay modes, 4th generation neutrinos, heavier vector bosons, leptoquarks, etc). This article will review the statistical issues involved in such claims in searches for New Physics in High Energy Physics (HEP) experiments.

The structure of this article is as follows. In this section, we discuss the various forms of statistical activities in HEP. The different statistical procedures relevant to searches for New Physics are discussed in section 2; these can result in the discovery of New Physics, or its exclusion, or the inability to choose between them. While the emphasis in section 2 is on discovery, section 3 deals with upper limits, sometimes resulting in mass exclusion ranges for...
New Physics. As an example, the exclusion of certain types of excited quarks is described. The recent discovery of the Higgs boson was very exciting and resulted in Nobel Prizes for Englert and Higgs. Details of the search, which illustrate the ideas of section 2, are given in section 4.

The emphasis of this article is to provide an introduction to the topic for young scientists who, if they are going to be serious data analysers looking for new phenomena, will need to read the relevant original papers. Thus we aim to provide an understanding of the issues involved, rather than a cook-book step-by-step account of the necessary procedures.

Although this article is designed primarily for Particle Physicists, we hope that some parts may be useful also for scientists in neighbouring fields.

1.1. Why statistics?

Experiments in HEP are expensive, both financially, and also in terms of the human effort in designing, building and running the detectors and accelerators. It is thus important to use effective statistical procedures for extracting the maximum amount of information (but not more than this) from HEP data. Not only is statistical analysis inexpensive, but usually it does not require an excessive amount of time.

There are several types of statistical procedures that are applicable to HEP:

• Parameter determination: Here we determine the value of the parameter in a theory, and also its uncertainty. If there is more than one parameter, (e.g. the gradient and intercept of a straight line fit to some data), as well as the values and uncertainties of each parameter, we also need to specify the correlations between them. Parameter determination is very common in Particle Physics, e.g. the mass of the Higgs boson.

• Goodness of fit: This could involve checking the consistency of our data with the predictions of the SM, with no new particles.

• Hypothesis testing: An alternative to goodness of fit could be to see which of two models was more in agreement with our data e.g. just the SM or the SM plus a SUSY particle. This article concentrates mainly on hypothesis testing.

• Decision making: This does not have much formal use in Particle Physics, but was applied informally in, for example, deciding in 2000 whether to continue running CERN’s LEP Collider (whose latest data at that time had hints of possible production of a Higgs boson of mass 115 GeV), or to stop running and to dismantle LEP in order to start building the LHC in the same tunnel. The formal procedure would have required numerical values for the cost function, which involved the money for the further running of LEP and for delaying the construction work of contractors for the modification of the collider’s tunnel; the delay in starting to build the LHC; and on the other hand the possibility that the Higgs boson might be convincingly discovered at Fermilab’s Tevatron collider rather than at CERN; etc. The decision to shut down LEP, taken after much detailed consideration, turned out to be correct as the Higgs boson’s mass is now known to be 125 GeV, and hence would have been beyond the accessible range at LEP.

An every-day example of parameter determination would be estimating the probability $P_h$ of a flipped coin coming down ‘heads’, from the result of 100 flips. Goodness of fit could be whether the results are consistent with $P_h = 0.5$, while deciding whether $P_h = 1/2$ or $P_h = 2/3$ is a better description of the data would be hypothesis testing. If we suspected that the person we were betting with was using a biased coin, decision theory could help us make up our mind whether to accuse him of cheating. This would require us to assign numerical
values to the cost function, including the ‘costs’ of what we would lose by not complaining, as compared with the possibility of being shot if we did.

2. Statistical issues for searches

In this section, we consider various statistical issues that are relevant in searches for new phenomena. A more detailed description of a particular search is given in section 4.

2.1. Bayes or frequentism?

Here we briefly describe two fundamental but very different approaches to statistics. Because we will be discussing the probability of data, given a specific hypothesis $P(\text{data}|\text{hypothesis})$, and also the probability of the hypothesis being true, given the data $P(\text{hypothesis}|\text{data})$, we first need to consider the meaning of probability in the Bayesian and frequentist approaches.

2.1.1. What is probability? There is a mathematical approach, developed largely by Russian mathematicians, which provides a series of axioms which probability must obey. They include concepts such as probability being represented by a number in the range zero to unity inclusive; zero probability means the relevant event never happens; the probability of something happening is 1 minus the probability of it not happening; etc. These axioms provide the necessary theoretical underpinning to the Bayesian and frequentist approaches to probability, but gives no real understanding of what it is.

The frequentist approach is to define the probability of a process as the limit as $N$ tends to infinity of the fraction of times it occurs in $N$ essentially identical independent trials. Since the definition involves a large number of repetitive trials, it does not apply to a unique event or to the value of a physical quantity. The following are examples of cases where a frequentist would not assign a probability:

- Rain in Manchester:
Frequentists would not ascribe a probability about whether it was raining in Manchester on the day of the Brexit vote. They would say either it was or it was not, and although we personally do not know, we could find out by looking at meteorological records. There is no way we can have a large number of repetitions of that day, to produce an ensemble for us to see the fraction of days there was rain. Furthermore frequentists would similarly refuse to provide a probability estimate, even if ‘Manchester’ was replaced by ‘The Sahara Desert’.

- Astronauts to Mars:
Again, frequentists would not assign a probability to whether the first astronauts to Mars will return to Earth alive, because there is only one first set of astronauts going there. This does not mean it is an uninteresting question, especially if you are one of those first astronauts, but in that case do not ask a frequentist for an assessment of the probability.

- Physical constants:
Similarly, frequentists will not speak about probabilities for numerical values of physical constants, or whether a particular theory is true or false. Examples would be whether the amount of dark matter in the Universe is above 25% of the critical density; or whether

4 There are also more ad hoc approaches e.g $\chi^2$, which is used in parameter determination and goodness of fit; and likelihood, which is part of both Bayesian and frequentist approaches.
SUSY particles of mass below 1 TeV exist. At present, we do not know, but these are not issues where we can have a series of repeated Universes with differing answers, for us to see in what fraction the statement turns out to be true.

In contrast Bayesians regard probability as a personal issue, expressing the degree to which one believes a statement. Thus people with different information available or with different beliefs can assign different numerical values to a probability. For example, if I toss a coin, you would be likely to decide that the probability of heads is 0.5. However, I might have cheated and seen that it is tails, and so for me the probability is zero. Alternatively, perhaps I had a very quick glance so that you would not see that I was cheating, and I think that it is tails but am not completely sure, so maybe I assign a probability of 0.2 to heads.

Given that Bayesian probability is a measure of personal belief, it does not require an extended series of trials, and so can be applied to a single event, or to the numerical value of a physical constant. The decision of what numerical value to assign a probability is determined by the concept of a fair bet: if you think that there is a 20% probability of something specific happening, you should be prepared to offer 4 to 1 odds and allow a person to bet against you in whichever direction they want. If you assign the probability incorrectly, you are liable to lose money.

2.1.2. Bayes’ theorem. Bayes’ theorem plays a central role in Bayesian analysis. It starts with the probability \( P(A \, \text{and} \, B) \) of \( A \) and \( B \) both happening. As can be readily verified, this satisfies

\[
P(A \, \text{and} \, B) = P(A|B)P(B) = P(B|A)P(A),
\]

where \( P(A|B) \) is the conditional probability of \( A \) happening, given the fact that \( B \) has occurred; and the last part of the equation follows because \( P(A \, \text{and} \, B) \) must be the same as \( P(B \, \text{and} \, A) \). If we divide the second part of equation (2.1) by \( P(B) \) we obtain

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}.
\]

This is Bayes’ theorem, and relates the conditional probabilities \( P(A|B) \) and \( P(B|A) \).

Not everyone appreciates that these probabilities are in general not equal\(^5\). This sometimes manifests itself in statements about the probability of data, given the ‘No New Physics’ hypothesis, being incorrectly interpreted as the probability of ‘No New Physics’, given the data.

Bayes’ theorem is acceptable not only to Bayesians, but also to frequentists, provided that all the probabilities are in accord with frequentist views of probability. The controversy begins when Bayesians replace \( A \) by ‘hypothesis’ or ‘parameter value’ and \( B \) by ‘data’. Then, for example, Bayes theorem becomes

\[
P(\text{param}|\text{data}) \propto P(\text{data}|\text{param})P(\text{param}).
\]

Here \( P(\text{param}) \) is the Bayesian prior which is the probability density that we assign to different parameter values before we perform our experiment; \( P(\text{data}|\text{param}) \) is the likelihood function, encapsulating what we can say about the parameter as a result of our data; and \( P(\text{param}|\text{data}) \) is the Bayesian posterior, combining the information from previous knowledge with that from our data. The problem is that equation (2.3) involves probability densities for parameter values which, as explained in section 2.1.1, are not acceptable for

---

\(^5\) LL’s granddaughter has a good example to illustrate the difference. She estimates the probability of a person having bread for breakfast, given that they are a murderer, as being over 90%. However, the probability of being a murderer, given that the person eats bread for breakfast, is (thankfully) very much smaller.
frequentists. Bayesians would retort that frequentists use too restrictive a definition of probability.

In addition, the choice of Bayesian priors can be a problem, especially in situations where we are sensitive to parameter values in a range where no or little previous information exists. The idea of choosing a constant prior, in order not to favour any particular value of the parameter \( \mu \), is fallacious; as well as being ignorant about \( \mu \), we could equally well say that we know nothing about \( \mu^2, 1/\mu, \ln(\mu) \), etc, and priors which are constant in these different variables are not the same.

2.1.3. Frequentist approach. The frequentist approach to parameter determination is very different, and uses only \( P(\text{data}|\text{parameter}) \), the probability to observe the data given a specific parameter value, and avoids \( P(\text{parameter}|\text{data}) \). Often, the data are summarised by a single quantity known as the ‘test statistic’ (see section 2.5.1) and the probability to obtain the observed test statistic value is used for parameter determination. For example, consider the case of estimating the temperature \( T \) of the fusion reactor at the centre of the Sun by using an estimate of the solar neutrino flux \( \phi \) from one month’s data from a large solar neutrino detector. The Neyman construction uses \( P(\phi|T) \) at a given value of \( T \) to choose a region\(^6\) in \( \phi \) for which the fraction of outcomes in that region is a pre-determined level \( 1 - \alpha \), for example 90\% (see figure 1). This is repeated for all values of \( T \) to produce the shaded confidence band.

\(^6\) Note that this is a region, rather than the region, as there are many possible regions containing the required fraction of outcomes. Thus the Neyman construction can be used to produce central regions, upper limits, lower limits, Feldman–Cousins intervals [1], etc, simply by changing the ‘ordering rule’ used for accumulating different test statistic values up to the required confidence level.
showing the likely values\(^7\) of the data for any value of the parameter of interest. Then we collect one month’s data, from which we deduce a flux \(\phi_d\). The intersection of a vertical line at \(\phi_d\) with the confidence band then gives the frequentist range \(T_l \leq T \leq T_u\) for the parameter \(T\) at the chosen confidence level.

The Feldman–Cousins approach \([1]\) is a frequentist method which uses an ordering rule for deciding which of the infinitely many possible data regions of probability \(1 - \alpha\) is chosen for the confidence band. It has many useful properties and is commonly used in HEP.

2.1.4. Final result. For the determination of a parameter \(\mu\), Bayesians and frequentists can end up with a statement of the form

\[
\text{Prob} (\mu_l \leq \mu \leq \mu_u) = 1 - \alpha
\]

(sometimes even with the same values of \(\mu_l\), \(\mu_u\) and \(\alpha\), but their interpretations are very different. For Bayesians, the measurement has been performed once, and from the data, \(\mu_l\) and \(\mu_u\) have been determined and are constants. Although the physics parameter \(\mu\) presumably has a unique value, this is unknown and they ascribe a probability (or credible) distribution to it. Then a fraction \(1 - \alpha\) of this distribution is in the range \(\mu_l\) to \(\mu_u\). In contrast, frequentists regard \(\mu_l\) and \(\mu_u\) as part of an ensemble of ranges for \(\mu\) that would be obtained if the measurement were to be repeated many times. Then equation (2.4) is a statement that a fraction \(1 - \alpha\) of these ranges should contain the true value of the parameter.

2.2. Blind analyses

If during a physics analysis we are looking at the result as we change the event selection, apply corrections, etc, it is possible that even subconsciously we could choose procedures which enhance the result we are hoping for. It is often possible to protect against this potential bias by performing a blind analysis, in which we choose the analysis procedure without being aware of its affect on the result.

One way of doing this is to devise the analysis technique by using Monte Carlo simulation of the data. In some cases however, the simulation might not be a good representation of the data. Better methods allow the analyser to look at the real data as much as possible without being able to deduce the actual result. This could involve the computer that is performing the analysis adding a random number to its output (which will be subtracted after all the selections, corrections and checks have been performed); adjusting the analysis while looking at the first, say, 15% of the data, and then freezing the analysis decisions; keeping events in the signal region in multi-variate space (‘the box’) hidden, but the rest of events visible; for the histogram being used for the final analysis, having visible only a different random fraction of each bin; etc.

Blinding procedures are important for analyses that expect only a few events; making personal decisions about whether or not to accept individual events can seriously bias the results. On the other hand, it may be difficult to implement a blind procedure in a search for unexpected phenomena.

After unblinding, the result should be presented without modification of the analysis, unless it would be obviously stupid not to. For example, if in a dark matter search 10 candidates were obtained, but all occurred at exactly midnight on a Saturday night, further investigation would be required. Any post-blinding modifications should be explicitly reported in a publication.

\(^7\) We refer to the chosen values of \(\phi\) as ‘likely values’, as they are the ones selected for a particular ordering rule.
If within a given collaboration several analyses are performed with the same physics aim, the procedure for how to combine them, or which to choose as the result (with the others being regarded as just confirmatory), should be decided in advance, without looking at the individual results.

Anyone embarking on a blind analysis should consider carefully what the most appropriate method is [2].

2.3. Examples of hypotheses

From a statistics point of view, hypotheses $H$ are classified as simple or composite. The former have the expected distribution of the data statistic being completely defined (e.g. the spin-parity of the Higgs boson being $0^+$, but see the next paragraph), while composite hypotheses include free parameters (e.g. the SM plus an extra particle of unknown mass). In general their best values must be determined in order to perform a sensible comparison of the hypotheses; or the result is shown as a function of the parameter (e.g. exclusion of $H1$ is shown as a function of mass—see section 2.5 and figure 5).

In practice, hypotheses in Particle Physics are rarely simple. Even if there are no physics parameters involved, there are almost always systematic nuisance parameters of an experimental nature (e.g. energy calibrations, electron identification efficiency, etc). However, if these systematic effects have a small effect on the analysis, it may be an adequate approximation to treat the hypothesis as simple.

Typically, ‘nothing new’ is chosen as the null hypothesis $H0$, while ‘possibly something exciting’ is the alternative hypothesis $H1$. The level for rejecting $H0$ (corresponding to possible discovery) is usually chosen more stringently than for rejecting $H1$ (exclusion).

In deciding whether or not to reject $H0$, there are two different sorts of mistakes we can make:

- Error of First Kind: We reject $H0$ when it is in fact true. For a confidence level $1 - \alpha$ for accepting $H0$, this should occur in a fraction $\alpha$ of the tests, assuming $H0$ is true.
- Error of Second Kind: We fail to reject $H0$, when some other hypothesis is true. The rate at which this happens depends on the details of the alternative hypotheses, including how similar they are to $H0$.

For the case where both $H0$ and $H1$ are simple, there is a lemma due to Neyman and Pearson [3] which basically states that the likelihood ratio statistic (LR) provides the best way of choosing between the hypotheses. More specifically, for a given value of $\alpha$ (the rate at which $H0$ is rejected when it is true), the value of $\beta$, the rate at which true examples of $H1$ are incorrectly accepted as $H0$, is a minimum when decisions are based on LR. This implies that for a given efficiency of selecting $H0$, the contamination will be a minimum$^8$. Even when the hypotheses are not simple, the LR can still provide a useful way of distinguishing between hypotheses.

In physics analyses, there are two different situations in which hypothesis testing is used: for the result of the experiment and, at an earlier stage of an analysis, for event selection.

---

$^8$ The power of a test is $1 - \beta$; it is the probability of rejecting $H0$ when $H1$ is true. The power depends on the separation of the probability density functions (pdfs), and on the value of $\alpha$. 
2.3.1. Result of experiment. This situation is where we are dealing with the result of the analysis e.g. Do we have convincing evidence for the existence of a SUSY particle? Is the parity of the Higgs boson positive (as expected) or negative? Is the hierarchy of neutrino masses normal or inverted? etc.

The different possible outcomes are summarised in table 1. A notable difference from event selector situations (see section 2.3.2) is that here we can make no decision regarding the two hypotheses. Also an incorrect answer is now more serious. In an Error of First Kind we reject $H_0$ when it is in fact true. This results in a false discovery claim, which is extremely serious, and warrants a very stringent confidence level (e.g. the ’$5\sigma$ criterion’ for discovery claims). An Error of Second Kind means that we have missed a possible discovery, which is unfortunate, but is regarded as not so terrible as false discovery claims.

<table>
<thead>
<tr>
<th>Consistent with $H_0$</th>
<th>Consistent with $H_1$</th>
<th>Outcome</th>
<th>If $H_0$ true</th>
<th>If $H_1$ true</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
<td>Exclude $H_1$</td>
<td>Correct</td>
<td>False exclusion (E2)</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Discovery</td>
<td>False discovery (E1)</td>
<td>Correct</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>No decision</td>
<td>Not enough data</td>
<td>Not enough data (E2)</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>?</td>
<td>? (E1)</td>
<td>?</td>
</tr>
</tbody>
</table>

2.3.2. Event selector. Hypothesis testing is used at an early stage of many analyses, where an attempt is made to select wanted events for further analysis, while having a strong rejection of the usually more copious background. Here the relevant hypotheses are ‘signal’ as $H_0$ and ‘background’ as $H_1$. An example could be that we want to select events containing $b$-quarks in a particular analysis. In these cases, there is no option of not making a decision; the events are either accepted or rejected. A multivariate method (for example: decision trees (DTs), neural networks or support vector machines) is usually adopted for this (see sections 2.3.3 and 4.5). Unlike the situation where our hypothesis testing is for the result of the experiment, incorrect classification is not a disaster, provided we can make a reliable estimate of the levels of loss of efficiency and background contamination, and correctly allow for them in the analysis. This will inevitably increase the statistical uncertainty on the result of the experiment as compared with an ideal situation of 100% efficiency and no background. An unfortunate possibility is that the result will be dominated by systematic effects associated with uncertainties in these estimates; or, even worse, that the systematics will be underestimated.

Multivariate methods usually allow the signal acceptance efficiency to be varied; thus with neural networks, the acceptance cut on the network’s output can be altered. As the cut is loosened and the efficiency is increased (reducing Errors of First Kind), the background generally rises (more Errors of Second Kind) and the purity of the selected sample decreases. Some compromise is therefore needed in deciding where to make the cut. This is usually done by optimising the expected accuracy of the result (although some approaches build this optimisation into the multivariate method itself). Estimating the background is generally much more difficult than obtaining the signal efficiency, because the background could come from many sources, and it is often the hard-to-simulate tails of distributions that creep into the
It may thus be prudent to reduce the systematic uncertainty at the expense of a somewhat larger statistical one.

2.3.3. Decision trees for event selection. The simplest, and one of the most common, multivariate techniques is a DT. In HEP, an event is typically summarised using a pertinent subset of the observables that are measured in the detector, or can be calculated from the measurements. A DT is essentially a sequence of thresholds which can be used to define a region of the observable space for which background events are rejected more strongly than those of the potential signal. A basic DT, with only two observables $x$ and $y$, would look something like that shown in figure 2. The logic of this DT is shown in figure 2 (left). At the first node, the input value of $x$ is compared to the threshold $x_1$. If the input value is smaller than $x_1$, the input value pair ($x$, $y$) is classified as background usually meaning an event with these input values would be rejected. If not, then the logic continues by following the direction indicated by the arrow to the second node. At the second node, the input value of $y$ is compared to the threshold $y_1$ and a similar decision is made, this time based on the input value of $y$. Once the path ends, the input values ($x$, $y$) can be categorised into those which are selected as signal and those which are rejected as background. For each pair of input values,

Figure 2. (Left) DT logic for separating the signal and the background. At each node, a decision is taken for which path to follow based on the input value $(x, y)$ compared to a particular threshold. (Right) Observable space $x$, $y$ and the regions for which the value would be selected as signal or rejected as background in the DT. The dashed vertical and horizontal lines indicate the thresholds at each node of the DT.

9 Backgrounds in HEP are typically estimated either by Monte Carlo simulations of the relevant processes or by examining real data in control regions (regions which do not overlap with the signal region) which are similar in some way to the signal region but are expected to be relatively free from signal. In either case, careful studies must be performed to determine how well these approximate the background in the signal region.
there is only a single, unique decision (or output), for example assigning $-1$ or $+1$ for background-like or signal-like, respectively. The values of $x$ and $y$ which end up as either signal or background are shown as the shaded blue or solid green regions of the observable space in figure 2 (right). The dashed lines which bound these regions are the values of the thresholds at each node in the DT.

By following this chain of logic for every event, those which are more likely to be a signal can be separated from those which are most likely to be a background. The threshold values chosen for $x_1$, $x_2$, $y_1$ and $y_2$ depend on how the DT is optimised. Typically in particle physics, one wants to reject as much of the background as possible while ensuring that the efficiency for selecting signal is large. While of course, in any real data event we do not know whether the event is a signal or a background, we can use simulated events, or data from samples not used in the analysis, to calculate the background and signal efficiencies of the selection. The threshold values can be optimised to yield both a high signal efficiency and background rejection. This is known as ‘training’ the DT. Typically the samples used to assess the performance of the DT are different from those used to train it, to ensure that it has not learned merely to distinguish signal and background fluctuations in the particular training samples used. In this example, there are only a few nodes; however more nodes can be added to the tree which can pick out finer structure of the signal and background in the multivariate observable space. The number of nodes in a DT is usually referred to as its depth.

In particle physics, it is often the case that rather than using a single DT, multiple DTs (sometimes referred to as forests) are trained and their outputs are combined, for example by averaging the individual outputs. For a large enough forest, this can yield a nearly continuous output so that rather than having the decision of either signal or background as the output, the output is a real number between $-1$ and $1$ indicating that the input $(x, y)$ is more ‘background like’ or more ‘signal like’ respectively. A final cut(s) can be placed on this value to define one or more categories of events which will be used in the statistical analysis. It should be obvious that the DT shown in figure 2 (left) is not unique in defining this particular division of the observable space. A number of similar DTs could lead to the same division of the signal and background like regions. Part of the optimisation in terms of the efficiency of the DT would be to construct the most simple one which gives the same outcomes.

In a similar way, the procedure by which a neural network can select a particular region of multi-dimensional observable space can be readily understood in terms of the responses of the nodes of the network to their inputs. As for DTs, there is no need to regard a neural network as a mysterious black box.

The procedure for using either a forest of DTs or a neural network is as follows:
- Define samples for the signal and background.
- Use these samples for training.
- Use independent samples to measure the signal and background efficiencies, and potentially optimise one or more thresholds to select events.
- Run real data through the forest or neural network.

For a thorough review of the algorithms used in neural networks and DTs, see [4, 5], respectively. Fortunately, there are several software packages available to train and use DTs and neural networks. Those commonly used in HEP are detailed in [6, 7].
2.4. H0, or H0 versus H1

One approach to looking for new physics is to compare the data with just the SM. This involves a goodness of fit test; if it fails, that may be an indication of new physics. An alternative is to compare the data both with the prediction of the SM (H0) and with that for the specific form of new physics (H1); this is hypothesis testing, and may involve the LR of the two hypotheses. The advantage of the first approach is that it may be able to detect any form of new physics, including theories not yet formulated. In contrast, comparing H0 and a specific H1 (e.g. a particular leptoquark) will probably be more sensitive to detecting that leptoquark than just testing whether the data are consistent with the SM, but it may well be useless for discovering some other form of new physics (e.g. extra dimensions).

2.5. Methods for H0 versus H1

2.5.1. Test statistic. A test statistic \( t \) is a way of summarising the data, and is used in deciding between H0 and H1. In a counting experiment (e.g. searching for dark matter), it could be simply the number of observed events, but in analyses using more information for each event, a likelihood ratio, perhaps profiled over systematic nuisance parameters, might well be more appropriate (as described in the penultimate paragraph of section 2.5.6).

2.5.2. Parameter confidence interval as method of hypothesis testing. Some searches for new physics involve precision measurements of a parameter (e.g. the anomalous magnetic moment of the muon), which is then compared with the SM prediction. Since we are looking for evidence of corrections from the effect of virtual new particles, we would regard this as an indirect approach, rather than searching for their direct production. We do not consider such analyses further in this article.

2.5.3. Direct production of new particle. As in the previous section, here we are also involved in parameter determination, but in this case the parameter \( \sigma \) is directly related to the production of the new particle or effect.

Searches for new particles may define special regions of observable space, where the new effects give rise to an excess of events e.g.

\[
\kappa = A \epsilon L \sigma + b,
\]

where \( \kappa \) is the expected number of events; \( A \) and \( \epsilon \) are the geometrical and kinematic acceptance, and the efficiency for observing signal events, respectively; \( L \) is the integrated luminosity describing the intensities of the incident beams and the running time; \( \sigma \) is the cross-section of the process we are interested in, and usually constrained to be non-negative; and \( b \) is the expected background. In reality there will be statistical and/or systematic uncertainties associated with \( A \), \( \epsilon \), \( L \), and \( b \) but for simplicity we will regard them as precisely known. In the absence of a signal, the observed number of events \( n_{\text{obs}} \) will be Poisson distributed with mean \( b \), because of statistical fluctuations, \( n_{\text{obs}} \) can be smaller than \( b \). For a given \( n_{\text{obs}} \), we might hope to find that \( \sigma \) was conclusively larger than zero, and hence claim a discovery. A more modest aim is to set an upper limit on possible values of \( \sigma \) i.e.

---

10 While it is true that a goodness of fit test does not explicitly involve the possible alternative hypotheses, the choice of which particular goodness of fit method to use does depend on the likely alternatives. For example, the Kolmogorov–Smirnov test [8] can be used for comparing a small number of events in some kinematic variable \( x \) with the predictions from H0. However, if the expected possible alternatives are likely to produce deviations at large \( x \), then the Anderson–Darling version [9], which is more sensitive to deviations in tails of distributions, is preferable. A number of different goodness of fit tests and their performance are detailed in [10, 11].
larger values of $\sigma$ are excluded because they would likely have resulted in significantly more events than we saw.

The above description was in terms of a single region of interest for the data. This could be generalised to several bins of data, with the likelihood function using the number of events $n_i$ in each bin, and the expected signal and background rates $s_i$ and $b_i$. A special case is the use of just two bins, where one (the ‘control region’) is designed to contain no signal and hence can be used to estimate the background; and the other contains background and possibly signal too. This is sometimes referred to as the ‘on-off’ problem. A parameter determination method can be used to find a range for $\sigma$.

To check for discovery ($\sigma = 0$) when there is one parameter of interest $\sigma$, we need to use some sort of 2-sided interval. Upper limits are unsatisfactory in that they always include zero, and lower limits because they almost always exclude it. Maximum probability density intervals are not ideal because they are metric-dependent. In problems with one or more dimensions, the Feldman–Cousins approach [1] in principle provides a sensible way of defining the acceptance region; however, computing time may be an issue in several dimensions. By changing the confidence level for the interval until $\sigma = 0$ is just excluded, we can determine the significance of the observed effect (see section 2.7).

Analyses often do involve more than one physics parameter of interest. For example, for a new particle its mass $M$ and cross-section $\sigma$ could be the parameters. Similarly for a 2-neutrino oscillation analysis, the parameters are $\sin^2 2\theta$ and $\Delta m^2$. The procedure in such cases could use a single 2-dimensional approach; or alternatively a ‘Raster Scan’, where statements are made about one parameter for a variety of fixed values of the other (e.g. conclusions about $\sigma$ for fixed values of $M$). Which is better depends on the circumstances [12]. For parameter determination, the 2-dimensional approach is preferable; the Raster Scan would provide a range for $\sigma$ at each $M$, regardless of whether the fit to the data for those values of $M$ and $\sigma$ was acceptable or not. In contrast, the Raster Scan is recommended for hypothesis testing. It can find more than one mass peak; and it will not result in exclusion at other masses $M_2$ just because there is a peak at $M_1$. This would be especially undesirable if the analysis had little or no sensitivity to the presence of a peak at $M_2$. Finally Wilks’ theorem (see section 2.10.1) does not apply when $(M, \sigma)$ or $(\sin^2 2\theta, \Delta m^2)$ are the variables, whereas it may if a series of fixed values of $M$ or $\Delta m^2$ respectively are tested as in a Raster Scan. Thus in the Higgs search culminating in its discovery at 125 GeV (see section 4.6.1), and in analyses resulting in the exclusion of other Higgs-like bosons at higher masses, Raster Scans were performed.

2.5.4. $\Delta S$. In comparing two hypotheses, the weighted sums of squared deviations $S_0$ and $S_1$ may be calculated by comparing the data with predictions of each hypothesis in turn (see equation (2.7)). In deciding whether or not it is possible to choose between the two hypotheses, it is generally better to use $\Delta S = S_0 - S_1$, rather than the individual $S_0$ and $S_1$, especially if the numbers of degrees of freedom involved in calculating $S_0$ and $S_1$ are large.

2.5.5. $p$-values. Many methods of distinguishing between hypotheses involve their $p$-values, which are the probabilities of obtaining a test statistic as deviant from the predicted value or
more so, i.e. they are the tail probabilities of the relevant pdf of the test statistic (see figure 3). These are discussed in more detail in section 2.6.

2.5.6. ln(likelihood-ratio). It is important to realise that there are several different types of likelihood ratio that are used in particle physics analyses:

- Comparison with Saturated Model [13]: The maximum value of an unbinned likelihood is not a measure of goodness of fit. However for a histogram, −2 times the Poisson ln(LR)
\[-2 \, \ln \text{LR}_\text{Poisson} = -2 \, \sum \ln \left( \frac{P(n_i|\kappa_i)}{P(n_i|n_i)} \right) \]
\[= -2 \, \sum \left[ n_i \ln \left( \frac{\kappa_i}{n_i} \right) + n_i - \kappa_i \right] \]  
(2.6)

asymptotically approximates to the weighted sum of squares (see equation (2.7) below). Here, \(P(n_i|\kappa_i)\) is the Poisson probability for observing \(n_i\) events in the \(i\)th bin, when the predicted value is \(\kappa_i\); \(P(n_i|n_i)\) is the Poisson probability, evaluated for the best value of \(\kappa_i\) i.e. \(n_i\); and the summation is over the bins of the histogram. This provides a likelihood method of estimating goodness of fit for histograms. For small values of the \(\kappa_i\), it is better than the standard weighted sum of squares

\[S = \sum (\kappa_i - n_i)^2 / n_i \quad \text{or} \quad \sum (\kappa_i - n_i)^2 / \kappa_i. \]  
(2.7)

These are respectively the Neyman or Pearson versions. They rely on the Gaussian approximation for the Poisson.

- The Feldman–Cousins method [1] also makes use of the likelihood ratio \(\mathcal{L}(\kappa|x) / \mathcal{L}(\kappa_{\text{best}}|x)\) as its ordering rule for the Neyman construction of the confidence belt for the likely values of the observable \(x\) for each value of the parameter \(\kappa\). Here \(\kappa_{\text{best}}\) is the best physical value of the parameter \(\kappa\) for the particular \(x\).

- \(\mathcal{L}_i / \mathcal{L}_0\): For choosing between two hypotheses \(H_0\) and \(H_1\), the ratio of their corresponding likelihoods \(\mathcal{L}_i / \mathcal{L}_0\) can be used as a test statistic. For example, in the search for the SM Higgs production at the Fermilab Tevatron, this was used by setting \(\mu = 0\) for \(H_0\) and \(\mu = 1\) for \(H_1\). Here \(\mu\) is the ratio between the observed cross-section and that predicted by the SM. Analyses trying to distinguish between different orderings of neutrino masses (normal and inverted hierarchies) also use this form of LR [14, 15]. In the absence of nuisance parameters, this LR is that used in implementing the Neyman–Pearson lemma for the optimal way of choosing between two simple hypotheses. When the expected distributions of the test statistic for the \(H_0\) and \(H_1\) are Gaussians of equal width, \(-2 \, \ln (\mathcal{L}_i / \mathcal{L}_0) = S_i - S_0\). There is an offset between them, however, when the Gaussian widths are different.

- In contrast, for the LHC Higgs search, two LR test statistics were used: \(\mathcal{L}(\mu = 0) / \mathcal{L}(\mu_{\text{best}})\) for testing discovery, and \(\mathcal{L}(\mu = 1) / \mathcal{L}(\mu_{\text{best}})\) for exclusion of \(H_1\) i.e. each hypothesis was tested separately, using its likelihood compared with the best possible one (as \(\mu\) was varied). There are also variants of these test statistics where the LR set equal to unity when the best value of \(\mu\) is in certain ranges. These are to avoid excluding \(H_0\) when there is a downward fluctuation causing \(\mu_{\text{best}}\) to be negative; or an upward fluctuation that might result in the exclusion of \(H_1\) [16]. An extension of these test statistics defined with a continuous value of \(\mu\) (i.e. \(\mathcal{L}(\mu) / \mathcal{L}(\mu_{\text{best}})\)) is used to determine an interval in the parameter \(\mu\). With \(\mu_{\text{best}}\) restricted to be greater than or equal to zero, the Feldman–Cousins method is reproduced.

An advantage of these test statistics (as compared with a generic ratio of likelihoods) is that their asymptotic distributions are easier to calculate, thus reducing the need for Monte Carlo simulation. The expected distributions of various log likelihood ratios are discussed in section 2.10.

As well as depending on the parameter(s) of interest \(\phi\), in the presence of systematic effects, each of these likelihoods also depends on the corresponding nuisance parameters \(\theta\). Within the likelihood paradigm, the profile likelihood is often used. This is obtained for each value of the physics parameter(s) by re-maximising the likelihood by suitable choice of \(\theta\), i.e. \(\mathcal{L}_{\text{profile}}(\phi) = \mathcal{L}(\phi, \theta_{\text{best}}(\phi))\). For the LHC version of the test statistic, an advantage is that asymptotically it is independent of any nuisance parameters \(\theta\), and so it is not necessary to
consider the effect of varying the $\theta$. In contrast the Tevatron approach deals with the systematics by integrating over them.

For a comparison of $p$-values and LR for hypothesis testing, see section 2.6.1. The method used in Particle Physics of quoting $p$-values with the test statistic being a LR can be regarded as either:

- A $p$-value method where the test statistic just happens to be a LR; or
- A LR method where $p$-values are used merely to calibrate the significance of the values of the LR.

2.5.7. Bayesian methods. There are various Bayesian methods available (e.g. posterior probability ratio, Bayes factor, AIC [17], BIC [18], etc). Largely because of their dependence on the choice of priors occurring in one hypothesis and not in the other, these methods tend not to be used in Particle Physics searches for new phenomena.

2.5.8. Minimal ‘cost’. As already mentioned, this is not formally used much in searches for New Physics. It is, however, one of the reasons behind the different cut-offs on the $p$-values used for claiming a discovery (5$\sigma$) and for excluding some form of New Physics (5%). Here the relevant ‘costs’ in the discovery case include the embarrassment of making a discovery claim which turns out to be incorrect, as compared with the reward for making the discovery when it is true.

2.6. Features of $p$-values

A $p$-value is the probability under some assumed hypothesis $H$ of obtaining data at least as discrepant as ours from the expected value (see figure 3). It is unfortunately common for $p$-values to be mis-interpreted as the probability that $H$ is true, given our data. This confuses the probability of data, given a particular theory, with the probability of theory given the data13. It also conflates ‘data we have’ with ‘our data, or more discrepant data’.

A similar mistake is to assume that of all tests of a hypothesis giving $p < 0.05$, only 5% of the hypotheses should be wrong. This is fallacious. If we tested energy conservation in events produced at the LHC with a $p$-value cut-off of 0.05, we would expect 5% of these to fail the cut, but all of them would be false positives (i.e. Errors of the First Kind) at the expected rate.

2.6.1. Why $p$ is not equal to LR. There is sometimes discussion of why an approach using the $LR = \frac{L_1}{L_0}$ can give a very different numerical answer from a $p$-value calculation. A reason some agreement might be expected is that they are both addressing the question of whether there is evidence in the data for new physics.

In fact they measure very different things. Thus $p_0$ (see figure 3 (b)) simply measures the consistency with the null hypothesis, while the LR takes both hypotheses into account. Furthermore a $p$-value is the tail area of a pdf beyond the observed value of the test statistic, while a likelihood uses the value of the pdf at the observed test statistic. There is thus no reason to expect them to bear any particular relationship to each other. Figure 4(b) shows that even at a given value of $p_0$, the LR can have any value within a range.

13 This has even lead to a journal trying to ban the use of $p$-values [19]. The solution, however, should be better education rather than a ban. It would be ridiculous to exclude the use of matrices, because some people do not understand them.
2.6.2. Combining p-values. Sometimes the effect of New Physics could appear in several different reactions in a given experiment; or in different experiments. It would then be sensible to combine the information, in order to improve the sensitivity of the search. The best way to do this is to perform a joint analysis, but this is not always possible, so an alternative is to combine the individual p-values. Problems with this are:

- Different effects: Two analyses might each have small p-values because they both disagree with the SM, but have different discrepancies e.g. peaks at quite different mass values.
- Non-uniqueness: As reviewed by Cousins [21], in the combination of n p-values, we are trying to find a transformation from the n (presumed uniform) 1-dimensional distributions to just one uniform distribution; this can clearly be achieved in many ways, thus yielding a variety of different possible combined p-values. Which is best requires extra information about the possible alternative hypotheses we think might be relevant; and

**Figure 4.** (a) Plot of \( p_0 \) against \( p_1 \) for comparing a test statistic \( t \) with two hypotheses \( H0 \) and \( H1 \), whose expected distributions for \( t \) are assumed to be two Gaussians of peak separation \( \Delta \mu \), and of equal width \( \sigma \) (see figure 3). For a given pair of distributions for \( t \), the allowed values of \( (p_0, p_1) \) lie on a curve or straight line (shown solid in the diagram). As the separation \( \Delta \mu \) increases, the curves approach the \( p_0 \) and \( p_1 \) axes. Rejection of \( H0 \) is for \( p_0 \) less than, say, \( 3 \times 10^{-3} \); here it is shown as 0.05 for ease of visualisation. Similarly exclusion of \( H1 \) is shown as \( p_1 < 0.1 \). Thus the \( (p_0, p_1) \) square is divided into four regions: the largest rectangle is when there is no decision, the long one above the \( p_0 \)-axis is for exclusion of \( H1 \), the high one beside the \( p_1 \)-axis is for rejection of \( H0 \), and the smallest rectangle near the origin is when the data lie between the two distributions. For \( \Delta \mu / \sigma = 3.33 \), there are no values of \( (p_0, p_1) \) in the ‘no decision’ region. In the CLs procedure, rejection of \( H1 \) is when the \( t \) statistic is such that \( (p_0, p_1) \) lies below the diagonal dotted straight line. (b) Contours of the likelihood ratio (here labelled LR) on the \( (p_0, p_1) \) plot. For a fixed value of \( p_0 \), a range of likelihood ratios is possible, depending on the separation of the \( H0 \) and \( H1 \) pdfs. The likelihood ratio is unity when the test statistic is half way between the Gaussian distributions (‘No evidence’); or when the distributions lie on top of each other (‘No sensitivity’). Further details can be found in [20].
more details of the individual analyses beyond just their p-values. Again the desirability of a combined analysis is demonstrated.

- Selection bias: Care must be taken to combine all relevant analyses, and not just the ones which give small individual p-values.

One possibility is to calculate the product $\pi_{\text{obs}}$ of the n individual p-values. Then the probability $P$ of the product of n independent uniformly distributed p-values being smaller than $\pi_{\text{obs}}$ is

$$P = \pi_{\text{obs}} \sum (-\ln \pi_{\text{obs}})^j / j!$$

where the summation extends from $j = 0$ to $n - 1$. Thus $P$ is larger than $\pi_{\text{obs}}$. For two measurements $P = p_1 p_2 (1 - \ln(p_1 p_2))$.

If the $p_i$-values were obtained from weighted sums of squares $S_i$ which are expected to have $\chi^2$ distributions with numbers of degrees of freedom $\nu_i$, an alternative is to use $S_{\text{comb}} = \sum S_i$ and $\nu_{\text{comb}} = \sum \nu_i$ to obtain the overall $P$. In the special case where the individual $\nu_i$ are all 2, this becomes equivalent to using equation (2.8).

Another, known as the Stouffer method [22], uses

$$z_{\text{comb}} = \sum z_i / \sqrt{n},$$

where $z_i$ are the significances14 or signed z-scores (i.e. the number of standard deviations corresponding to the one-sided Gaussian p-value, with $p = 0.5$ corresponding to $z = 0$) and $z_{\text{comb}}$ is the combined value.

### 2.7. Significance

When an analysis looking for New Physics determines the p-value for the null hypothesis of ‘nothing new’, it is conventional to convert this to a z-score as defined in the previous section. For example a value of $p = 3 \times 10^{-7}$ is equivalent to 5$\sigma$ or a z-score of 5. This does not imply that the test statistic is Gaussian distributed, but merely that a Gaussian is used in the conversion15.

However it is common to find approximations to the significance being used, e.g. $S/\sqrt{B}$, where $B$ is the expected background in a region, and $S$ is the excess of observed events above the expected background. This is to be discouraged because:

- The approximation is based on the Poisson distribution with large mean $\mu$ being similar to a Gaussian distribution with mean and variance $\mu$. However, this approximation is worse in the tails than for the bulk of the distribution.
- Maximising $S/\sqrt{B}$ to determine the best way of conducting a search can be very unsatisfactory. For example, it could result in very stringent cuts being used that would reduce $B$ to $10^{-4}$ of an event, but the expected $S$ could be down to 0.1 event. Then the expected $S/\sqrt{B} = 10$, which sounds very good, but unfortunately 9 times out of 10 we would observe no events at all, which is not an ideal way to look for New Physics.

It is also worth noting that the formula makes no allowance for any uncertainties on $B$ and it provides an approximation only to the local p-value rather than the global one; it does not account for any look elsewhere effects. Other approximations used in HEP perform better than $S/\sqrt{B}$, for example $\sqrt{(S + B) \ln(1 + S/B) - S}$ (mentioned in [23]), but it is still advisable to use a full calculation of the significance.

---

14 This usage of the word ‘significance’ seems to be specific to Particle Physicists. Statisticians use ‘z-scores’.
15 Given that there is a one-to-one relationship between $p$ and the z-score, the only reason for using the latter is that it is easier to remember a number like 5 rather than $3 \times 10^{-7}$. 

2.8. Look elsewhere effect (LEE)

The significance of an interesting peak in a mass spectrum is defined in terms of the probability of a background fluctuation producing an effect at least as large as the one actually seen. If the background fluctuation is required to be at the observed mass, this probability is called the local p-value. If, however, the restriction on the location of the background fluctuation is removed, this is instead a global p-value, and is larger than the local one (i.e. less significant) because of the LEE\(^{16}\).

Ambiguity in the global p-value arises because of the choice of requirements that are imposed on the location\(^{17}\) of potential background fluctuations. Options include (a) any reasonable mass; (b) anywhere in the current analysis; (c) anywhere in analyses by members of the experimental collaboration; etc. Possibilities of different event selections, choice of which distributions to look at, etc may also contribute to the LEE. Thus which ‘anywhere’ to choose depends on who you are; the one for the Spokesperson of a large collaboration might be wider than that for a graduate student doing a specific analysis.

In addition to the local p-value, it is desirable to quote at least the global one corresponding to option (a). It is clearly important to specify in any publication exactly how the global p-values are defined.

One way of calculating the global p-value is the brute-force method of generating Monte Carlo simulations of the data, and seeing how often there are fluctuations ‘anywhere’ in the simulations. The trouble is that, to check for effects at the 5\(\sigma\) level, of order \(10^8\) simulated spectra have to be generated, and then fitted. A useful alternative approach involves looking for fluctuations at a lower level (e.g. 2\(\sigma\)), and then extrapolating to the required significance level; further details can be found in [25].

2.9. Why 5 sigma?

Statisticians who hear about the 5\(\sigma\) criterion for discovery are very skeptical about it being sensible. In particular, they claim that it is very uncommon that statistical probability distributions accurately describe reality so far out in the tails. Particle Physics does differ from other fields in that we do believe that in many situations the numbers of observed events do follow a Poisson distribution, and so small tail probabilities may well be meaningful. However, certainly as far as systematic effects are concerned, even in Particle Physics there is often a degree of subjectivity in specifying the exact magnitude of the effect, let alone believing in its assumed distribution far into the tails of the observed effect. Thus the Statisticians’ criticism may well be valid.

The traditional arguments for the 5\(\sigma\) criterion are:

- **History:**
  In the past there have been many ‘phenomena’ that corresponded to 3 or 4\(\sigma\) effects that have gone away when more data were collected. The 5\(\sigma\) criterion is designed to reduce the number of such false claims.

- **Look elsewhere effect:**
  This allows for a statistical fluctuation to occur ‘anywhere’, rather than just at the location of a peak actually seen in the data. This has the effect of reducing the significance (see section 2.8).

\(^{16}\) The LEE in HEP is equivalent to what statisticians refer to as multiple trials [24].

\(^{17}\) If the width etc of the possible signal are regarded as free parameters, they also contribute to the LEE factor.
• Subconscious Bayes’ factor:
  When searching for a discovery, the test statistic that is used to discriminate between ‘just background’ (the null hypothesis \( H_0 \)) and ‘background plus signal’ (\( H_1 \)) is often the likelihood ratio \( \mathcal{L}_1 / \mathcal{L}_0 \) for the two hypotheses (see section 2.5.6). The 5\( \sigma \) criterion is applied to the observed value of this ratio, as compared with its expected distribution assuming just background. However, more relevant to the discovery claim is the ratio of the Bayesian posterior probabilities for the two hypotheses. These are related by Bayes theorem:

\[
P(H_1|\text{data}) = \frac{p(\text{data}|H_1)}{p(\text{data}|H_0)} \times \frac{\pi_1}{\pi_0}
\]

where \( p(\text{data}|H_i) \) and \( \pi_i \) are respectively the likelihood and the prior probability for hypothesis \( H_i \); and \( P(H_i|\text{data}) \) is its posterior probability. Hence, even if the likelihood ratio favours \( H_1 \), we would still prefer \( H_0 \) if our prior belief in \( H_1 \) was very low. Thus, if an experiment seemed to show that neutrinos travel faster than light, we might well have a tendency to believe that there was some undiscovered systematic, rather than that neutrinos were a potential source for violating causality.

The above argument is clearly a Bayesian application of Bayes theorem, while analyses in Particle Physics usually have a more frequentist flavour. Nevertheless, this type of reasoning does and should play a role in requiring a high standard of evidence before we reject well-established theories. There is sense to the oft-quoted maxim ‘extraordinary claims require extraordinary evidence’.

• Systematics:
  It is in general more difficult to estimate systematic uncertainties than statistical ones. Thus a \( n \sigma \) effect in an analysis where the statistical uncertainties are dominant may be more convincing than one where the \( n \sigma \) claim is dominated by systematic uncertainties. In the latter case, a 5\( \sigma \) claim would be reduced to merely 2.5\( \sigma \) if the systematic uncertainties had been underestimated by a factor of 2; this corresponds to the \( p \)-value increasing from \( 3 \times 10^{-7} \) by a dramatic factor of \( 2 \times 10^4 \). The current 5\( \sigma \) criterion is partially motivated by a cautious approach to somewhat vague estimates of systematic uncertainties.

Of course not all experiments are equally affected by the above considerations, and it is not really sensible to have the same discovery criterion for all experiments [26]. In particular, for a very general non-blind search for anything unusual, 5\( \sigma \) may well be not stringent enough, whereas for as yet undiscovered but nevertheless expected effects (e.g. the \( H^0 \) decay to \( \mu^+\mu^- \)), it is unnecessarily strict. However, there seems to be little chance of having a varying criterion.

2.10. Distributions of test statistics for \( H_0 \) and \( H_1 \)

In order to claim discovery of new physics, or its exclusion in various parameter ranges, it is necessary to know the expected distributions of the test statistics e.g. the LR discussed in section 2.5.6. These can always be obtained by Monte Carlo simulation, but it is useful if they can be determined analytically. This is sometimes possible if there is enough data to be in the asymptotic regime.

For simple hypotheses (i.e. completely specified, without any nuisance parameters) the logarithm of a likelihood, or of a likelihood ratio, involves a summation over the independent events. Hence, assuming the central limit theorem is applicable, the distribution approaches a Gaussian. The locations and widths of these distributions for \( H_0 \) and \( H_1 \) require more information.
Some analyses involve the comparison of two hypotheses that are simple. Even if there are some systematic nuisance parameters, if their effects are not too large, the simple versus simple criterion might be applicable. Examples of this could be the different neutrino mass hierarchies; or different spin-parity possibilities for the Higgs boson. For the neutrino example, it is common practice to assume that the Gaussian distributions of $-2$ times the log-likelihood ratio for $H_0$ and $H_1$ are symmetrically situated at $\pm T$, with each having a standard deviation $2\sqrt{T}$. This is not in general true for two simple hypotheses. It may apply in the neutrino case because the distributions of the data are very similar for the two hierarchies.

Sometimes the hypotheses being tested are nested i.e. We can reduce the larger hypothesis $H_1$ to be the same as $H_0$ by choosing special values of the $\nu$ extra parameters of $H_1$. Because of cancellations between the correlated nested log-likelihoods, the central limit theorem in general does not apply [27]. The expected distributions are discussed in the next section.

The expected asymptotic distributions for the test statistics used in the LHC Higgs searches (see section 2.5.6) are discussed in detail in [16].

### 2.10.1. Wilks’ theorem

When comparing data with two hypotheses, a common test statistic $t$ is $-2$ times the logarithm of the likelihood ratio, or almost equivalently the difference in $S_{\text{min}}$ between the two hypotheses, where $S_{\text{min}}$ is the usual weighted sum of squared discrepancies between theory and data, minimised with respect to any free parameters in each case. Wilks’ theorem [28] states that under special circumstances, the distribution of $t$ is the mathematical $\chi^2$, with the number of degrees of freedom $\nu$ equal to the difference in the numbers of free parameters in the two hypotheses.

The conditions are:

- The null hypothesis must be true.
- The hypotheses are nested.
- When we reduce $H_1$ to $H_0$, the values of all the extra parameters must be well defined, and not at an extreme end of their possible ranges.
- There are enough data for asymptotic approximations to apply.

When $H_0$ is true, we still expect its $S_{\text{min}}$ to be larger than that for $H_1$ (because $H_1$ has extra parameters), but the difference should be small. However, when $H_1$ is true, the difference can be large. Wilks’ theorem provides an easy way of assessing how small is small: the value of $t$ can immediately be turned into a p-value for the null hypothesis, via the known distribution of $t$. When the theorem does not apply, $t$ may well still be a useful test statistic, but it will usually be necessary to calculate its expected distribution under $H_0$ by simulation.

The examples below illustrate when the conditions do or do not apply:

- Polynomials: e.g. $H_0$ is a straight line, and $H_1$ is a polynomial of degree 3. These are nested since by setting the squared and cubic terms in $H_1$ to zero, we obtain $H_0$. Furthermore, these special values are not at the end of mathematically or physically allowed ranges. Wilks’ theorem applies, with $\nu$ equal to the difference in the number of free parameters in the two hypotheses; in this example $\nu = 2$.

- Presence of signal. Here $H_0$ is just a background distribution, while $H_1$ also has a signal bump of mass $M$ and cross-section $\sigma$. $H_1$ becomes $H_0$ if $\sigma$ is set to zero, but in that case $M$ is undefined. Also unless $\sigma$ is allowed to be negative, its special value is at the end of its physical range. Thus the theorem does not apply. However, if $M$ is kept fixed (and negative values for $\sigma$ are allowed), either because its value is known or because we are performing a Raster Scan, Wilks’ theorem may apply.
• Spin of resonance: We try to determine whether the spin of a resonance is 1 or 2. These hypotheses are not nested, so the theorem does not apply.

3. Upper limits

It is (unfortunately) common for searches for new effects to yield null results. There are two possible reasons for this: (a) the sought after phenomenon does not exist; or (b) the effect may exist, but our experiment was not sensitive enough to detect its effect. The former can result in some progress in our knowledge, but the second merely means we should try to perform a better experiment. In order to decide which applies, we need to know the sensitivity of our experiment. The most famous example of a null-result influencing physics was the Michelson–Morley experiment [29]. This set out to measure the speed of the Earth as it travelled through the aether, the all-pervasive medium that had been invented as the carrier for electromagnetic radiation (e.g. the light and heat reaching earth from the Sun). The experiment saw no effect, and was able to place a stringent upper limit on the possible speed. This was lower than that expected from the revolution of the Earth about the Sun, and of the solar system around the Milky Way. This null measurement thus resulted in the demise of the aether theory; special relativity soon followed.

Contemporary theoretical particle physicists have fertile imaginations, and come up with many ideas for new types of particles or interactions, which may yield events with special characteristics, e.g. a peak in a mass spectrum; an excess of events with large transverse momentum; undetected transverse momentum due to unseen new particles; etc. Examples of contemporary direct searches include those for dark matter and for SUSY particles. These experiments have not resulted in discoveries and, subject to specific model assumptions, various mass ranges for them have been excluded. However, this has not yet resulted in the demise of these theories.

Even for simple counting experiments, there are several methods that can be used to extract an upper limit. These include:

• $p$-values: The upper limit on $\sigma$ is the smallest value for which $p_1$ is less than a specified level e.g. 5%.

• Likelihood: The parameter is increased until the logarithm of the likelihood decreases from its largest value by a specific amount.

• $\chi^2$: Here the weighted sum of squares is required to increase by a specific amount from its smallest value.

• Bayesian methods: For example, the posterior probability is integrated up to a parameter value such that the integral is equal to the required credible level. This upper limit will be sensitive to the Bayesian prior used for extracting the posterior.

• Neyman construction: The procedure described in figure 1 is used, with the modification that the confidence belt is defined to run from infinity down to some possible data value such that the chosen range for the data contains the required probability. This then is equivalent to the $p$-value method described above.

• Feldman–Cousins: This is the version of the Neyman construction where a likelihood-ratio ordering rule is used for constructing the confidence belt. Whether the resulting region for the parameter is two-sided or simply an upper limit depends on the value of the test statistic $t$, and is not up to the choice of the analyser.

Once systematic uncertainties on the background are taken into account, the possibilities multiply even further. For example, the likelihood method could be replaced by the profile
likelihood, in which the profiling is performed over the nuisance parameters as described in section 2.5.6.

A comparison of several methods can be found in [30]. The largest differences among the methods arise when the observed number of events $n$ is less than the expected background $b$, presumably because of a downward statistical fluctuation. Determining upper limits is thus not like calculating $2 + 3$, for which there is one preferred answer so it is important to specify precisely the procedure used for obtaining any upper limit.

More details on common procedures can be found in the first two of the PHYSTAT series of meetings on statistical issues in Particle Physics, which were devoted to upper limits [31, 32].

3.1. CL$_s$

The CL$_s$ method [33] is a modification of the standard ‘exclusion of $H_1$’ procedure. It aims to provide protection against a downward fluctuation of the test statistic $t$ (see figure 3(a)) resulting in a claim of exclusion in a situation where the experiment has little or no sensitivity to the production of the new particle; this could happen in up to 5% of such experiments. It achieves this by defining

$${\text{CL}}_s = p_1/(1 - p_0), \quad (3.1)$$

i.e. CL$_s$ is the ratio of the left-hand tail areas of $H_1$ and $H_0$ beyond the test statistic in figure 3(b). Thus CL$_s$ cannot be smaller than $p_1$. It tends to $p_1$ when $t$ lies above the $H_0$ distribution, and to unity when the $H_0$ and $H_1$ distributions are very similar. The reduced CL$_s$ exclusion region is shown by the dotted diagonal line in figure 4; the price to pay for the protection provided by CL$_s$ is that there is built-in conservatism when $p_1$ is small but $p_0$ has intermediate values i.e. there are more cases in which there is no exclusion of $H_1$.

Most statisticians are appalled by the use of CL$_s$, because they consider that it is meaningless to take the ratio of two $p$-values. We regard CL$_s$ as a conservative version of a frequentist procedure.

3.2. Sensitivity

As well as using experimental data to determine an upper limit on a cross-section (or the related lower limit on the mass of the hypothesised particle), it is also important to determine the sensitivity of the experiment for these quantities. This is also true for the significance of a search for a possible discovery. These sensitivities are the expected values, and depend on the details of the detector, the amount of data and the model being investigated, but not on the actual data acquired. They can thus be calculated before the experiment is performed.

There are several options for the ‘expected’ value. It could be the median or the mean of the expected distribution. The median has the advantage of being invariant with respect to different choices of the functional form for the variable of interest (e.g. lifetime or decay rate; $m$, $m^2$ or $\ln m$; etc), and to be less sensitive to outliers. However, the median value of a distribution can be zero, which is fairly uninformative, so when such a situation is anticipated, the mean would be a better choice. Another possibility is to determine the Asimov value. This

---

$^{18}$ Given the fact that CL$_s$ is the ratio of two $p$-values, the choice of symbol CL$_s$ (standing for ‘confidence level of signal’) is not optimal. Another source of confusion is that in definitions of CL$_s$ the ways the $p$-values are defined vary, so the formulae can look different but the underlying concept is the same. A subtlety with equation (3.1) is that $p_0$ there is the probability of obtaining a measurement greater than the observed one, rather than the usual ‘greater than or equal to’. This is to make $1 - p_0$, the probability of a value smaller than or equal to the observed one. It makes a difference when the observation is a small discrete number.
essentially assumes that the data agrees exactly with the prediction of the relevant hypothesis, without any fluctuations\footnote{In [16], this was named in honour of Isaac Asimov’s suggestion [34] that elections would be simpler if, instead of the whole population being eligible to vote, the result were to be determined by the opinion of a single person, who was deemed to be the most typical voter.}.

The reasons for determining the sensitivity include

- **Funding:** If two or more experiments are competing for funding, the relevant agency may well take into account their expected sensitivities.
- **Optimising a search:** If there are several ways of conducting a specific search, it is bad statistical practice to try them all, and to select the one which gives the answer you like best. It is better to determine which one provides the best sensitivity.
- **Comparing experiments:** Because of the different statistical fluctuations in separate experiments, the analysis which provides the tightest limit or the largest significance is not necessarily the one with the best sensitivity.
- **Checking that actual limits or significance are reasonable:** The expected limit and its distribution enable us to determine whether the observed value is within the expected range. Of course, it is entirely possible that the observed upper limit on a cross-section is much larger than the value expected assuming $H_0$, if the signal actually is there.

### 3.3. Example of an upper limit: search for excited quarks

A specific example is the ATLAS search for an excited quark $q^*$ [35], which decays quickly to a quark and a gluon; the experimental signature would be two jets (two distinct streams of collimated charged and neutral particles). A spectrum of the di-jet effective mass was used to look for a peak, which would be at the mass of the $q^*$. The absence of a peak can be evidence against the production of a $q^*$, provided that it would have been produced with a large enough cross-section, and that we have enough data.

For each possible mass $M_{q^*}$ in a Raster Scan, and using the particular model for $q^*$ production and decay, an upper limit on the cross-section $\sigma$ was determined at the 95% level, by integrating the Bayesian posterior probability distribution for $\sigma$. This is then compared with a theoretical prediction of the cross-section, also as a function of $M_{q^*}$. At masses below about 6 TeV, the predicted cross-section is larger than the experimental upper limit, and so, within the model assumptions, $q^*$ of these masses are excluded (see figure 5). Thus the theoretically predicted cross-section allows us to convert upper limits on cross-sections into a lower limit on the allowed $q^*$ mass. Because the data were collected at a centre of mass energy of 13 TeV, the search was sensitive to high mass $q^*$.

Also shown on the plot are the median expected upper limits on the cross-section, and the 68% and 95% bands for its expected distribution, assuming that no $q^*$ is produced ($H_0$). It is satisfactory that the observed and the predicted limits are close.

### 4. An example of discovery: the search for the Higgs boson

In the SM, the masses of the fundamental particles, the leptons, quarks and bosons, are a consequence of their interactions with the Higgs boson. The Higgs boson is a particularly special particle in that it is the only known fundamental particle which has zero spin and, until its discovery in 2012, remained the only piece of the SM which had yet to be discovered. One of the reasons for this elusiveness is that the SM makes no prediction as to what the mass of
the Higgs boson should have been. The range of possibilities, taking certain theoretical considerations into account, spanned from roughly the mass of the electron (i.e. around 0.6 MeV) to over 1000 times the mass of the proton (≈1 TeV). Furthermore, the production of the Higgs boson is comparatively rare compared to other SM processes which occur at particle colliders like the LHC and the Tevatron. Despite this challenge, on the 4th of July in 2012, the ATLAS and CMS collaborations announced their discovery of the Higgs boson with a mass of around 125 GeV, roughly 130 times the mass of the proton. For the discovery of the Higgs boson, around 10 fb$^{-1}$ of data, or one thousand trillion proton-proton collisions, were used per experiment. Now with about six times that amount of data, the measurements of the Higgs boson’s properties are becoming more varied and ever more precise.

Naturally, the LHC analyses have evolved and improved over time and thus the contents of this section is a composite of the procedures used up to, at the time of and after the discovery, rather than an exact historic account of the analysis performed in the summer of 2012.

\[ \text{Figure 5. Results from an ATLAS search}\]

\[ \text{[35] showing the 95\% credibility-level upper limits obtained from the dijet distribution on cross-section } \sigma \text{ times acceptance } A \text{ and branching fraction BR for an excited quark } q^* \text{, shown as a function of the } q^* \text{ mass. The jagged solid line is the observed upper limit, while the dotted curve is the median expected one, assuming that no } q^* \text{ is present (H0); the inner green and outer yellow bands show the 68\% and 95\% regions for the expected limit. Larger values of the cross-section are excluded because that would have resulted in more events than were observed. The predicted cross-section for } q^* \text{ production is shown as the dashed almost straight line. Below 6 TeV, this cross-section is larger than the upper limit, and so, for the given assumptions, } q^* \text{ below this mass are excluded. © 2017 CERN, for the ATLAS Collaboration. CC BY 4.0.} \]
4.1. Decay modes of the Higgs boson

As the Higgs boson is unstable, almost immediately after it is produced, it will decay to other particles. In the SM, for a particular value of the Higgs boson mass, the rate at which the Higgs boson will decay into different SM particles, or ‘final states’ is predicted to a very high precision. The fractions of Higgs boson decays via the different modes, known as the branching fractions, vary greatly depending on the mass of the Higgs boson $m_{H0}$. If the Higgs boson were very heavy, more than 300 GeV, then most of the time it would decay into a pair of W or Z bosons, which themselves as unstable particles will decay into leptons or quarks. In nature, it turns out that the Higgs boson has a mass of 125 GeV. This is great news for experimentalists as it means there are a number of different final states which can be exploited to study its properties.

The first two decay modes are the two photon ($H^0 \rightarrow \gamma\gamma$) and the four lepton ($H^0 \rightarrow ZZ^* \rightarrow llll$) channels. The photons ($\gamma$) in the $H^0 \rightarrow \gamma\gamma$ channel are stable particles while the two Z bosons in the $H^0 \rightarrow ZZ^* \rightarrow llll$ are unstable meaning they further decay to two charged leptons ($l$) each. As the Higgs boson’s mass is not large enough to produce two real Z bosons, one of the Z bosons is produced ‘off-shell’ (usually written $Z^*$) sometimes referred to as a virtual particle. The final states (leptons and photons) in these decays consist of particles which can be reconstructed in particle detectors with a high resolution. Figure 6 shows the distributions in data of the invariant mass of the two photons in the $H^0 \rightarrow \gamma\gamma$ search from CMS and of the four leptons in the $H^0 \rightarrow ZZ^* \rightarrow llll$ search from ATLAS, at the time of the discovery. The invariant mass of the decay products was used in the search since the signal that a Higgs boson is produced appears as a clear, narrow bump on top of other,
background processes with the same final state. Despite the rather small branching fraction in these two modes they provided the greatest sensitivity in the search.

The next most sensitive decay mode is $n \rightarrow lH\ell\ellWW$. In this case, the presence of the neutrinos in the final state makes it impossible to determine the invariant mass of the decay products since the neutrinos escape detection. The design of ATLAS and CMS however, allows the determination of the transverse mass ($m_T$) of the event, which is defined as

$$m_T = \sqrt{2|p_T^1| + p_T^2||p_T^{miss}|(1 - \cos \phi)},$$

where $p_T^1$ and $p_T^2$ are the momentum vectors of the two leptons projected onto the transverse plane (perpendicular to the proton beamlines). Although we cannot measure the momentum of the neutrinos in the event, we can estimate their summed momenta in the transverse plane as ‘missing transverse momentum’ in the event, $p_T^{miss}$. Conservation of momentum tells us that the sum of all of the transverse momentum vectors should be zero. The missing transverse momentum is just the magnitude of the vector we would need to add in order to balance the event, that is to account for the momentum carried away by the neutrinos. Finally $\phi$ is the angle between the sum of the lepton transverse momentum vectors and the missing transverse momentum.

The signal will appear as a rather broad excess in the distribution of $m_T$ as opposed to a narrow peak. While the resolution in this decay mode is poor compared to the $H^0 \rightarrow \gamma\gamma$ and $H^0 \rightarrow ZZ^* \rightarrow llll$ modes, the branching fraction is much larger so that more events are expected in this search.

Finally, the least sensitive channels in the search for the Higgs boson were the $H^0 \rightarrow \tau\tau$ and $H^0 \rightarrow bb$ decays. Although the $\tau$-lepton or $b$-quark will decay or hadronise (and subsequently be reconstructed as sprays of particles, or ‘jets’), most of the products can be detected and the invariant mass can be calculated. Furthermore, the branching fractions in these two modes are relatively large. However, the challenge in these decay modes comes from the overwhelming backgrounds which produce the same final state. In the $H^0 \rightarrow bb$ mode, the backgrounds are so large that the dominant mechanism (gluon fusion) for producing the Higgs boson is completely masked. Instead, these channels target other, less common production mechanisms which help to reduce the background contributions. The events in this analysis are selected based on the presence of additional particles in the final state beyond the pair of $\tau$ leptons or $b$-jets, targeting the vector boson fusion mechanism or the Higgs-strahlung mechanism, in which a Higgs boson is radiated from a virtual $V$, where $V$ is a $W$ or $Z$ boson (see figure 7). While Higgs bosons produced via these mechanisms are
fewer by a factor of 10 than the dominant mechanism, the backgrounds are reduced by larger factors making the signal clearer. Furthermore, these production modes are extremely important to decipher the properties of the Higgs boson.

Given the various pros and cons in each of the decay modes, both ATLAS and CMS used a combination of the decay modes to yield the greatest sensitivity in the search for the Higgs boson. It was this combining of different decay modes which lead to the eventual discovery of the Higgs boson by ATLAS and CMS.

4.2. Detecting the Higgs boson

Detecting the Higgs boson’s decay products from these different channels requires the use of several different types of detector technologies. The two large general purpose detectors at the LHC, CMS and ATLAS, are both cylindrical and consist of concentric sub-detectors with different functions:

- **Tracker.** This is a high spatial resolution detector placed as close as possible to the interaction region. This detects charged particles (such as electrons and charged hadrons) and, due to the magnetic field of several Tesla, enables their helical paths to be determined. It is useful for finding charged particles that come from the decays of heavier particles, such as those containing b quarks, that have travelled millimetres before decaying.

- **Electromagnetic (EM) calorimeter.** Electrons and photons are identified and their energies measured by the showers they produce in the calorimeter’s material of high Z nuclei.

- **Hadron calorimeter.** Usually made from a very dense material (such as brass), this is used for measuring the energies of hadrons. This sub-detector is particularly important for reconstructing neutral hadrons, such as neutrons, which do not leave tracks in the tracker. These hadrons are commonly contained in collimated sprays of particles, known as jets.

- **Muon detector.** These are placed at the outside of the whole detector, so that almost the only charged particles penetrating as far are muons. Although the muons escape the detector, they leave a track in the muon detectors which can be used to identify the particle as a muon and provide extra information beyond that from the inner tracker to measure their momenta.

By combining the information from each of these components, ATLAS and CMS can reconstruct the proton collision events and identify whether or not they were likely to have been the products of a Higgs boson decay. Those events more likely to include a Higgs boson are selected for further analysis.

4.3. Data selection for a Higgs analysis

The LHC at CERN collides bunches of protons together at extremely large energies. For a given centre-of-mass energy, we can predict how often an inelastic collision, in which additional particles including Higgs bosons are produced, occurs. At a centre-of-mass energy around 7 or 8 TeV, the cross-section for a Higgs boson to be produced is roughly 21 pico-barns (assuming the mass to be what it is known to be today) \(^{39}\). With the amount of data

---

\(^{21}\) The ATLAS detector is 46 m long and 25 m in diameter while the CMS detector is a little smaller being only 21.6 m long and with a diameter of 15 m. Despite the CMS detector being smaller, it has a mass of 14 000 tonnes compared to the ATLAS detector’s 7000 tonnes due to the differences in detector component design between the two.

\(^{22}\) The LHC currently runs at a centre-of-mass energy of 13 TeV. At this energy, the cross-section for Higgs boson production in the SM is around 52 pico-barns.
collected by July 2012, approximately 400 thousand Higgs bosons were produced. However, the background cross-sections in most of the decay channels are considerably larger than that of the Higgs boson, making the Higgs boson hard to identify. The total hard collision cross-section, meaning the cross-section for processes in which the proton’s constituent quarks interact with enough energy to produce new particles (about 74 mb [40]), is over 3.5 billion times larger than that of Higgs production. This makes trying to spot the Higgs boson about 3 and a half times less likely than calling a mobile number at random while in the UK without knowing it beforehand, and finding that it belongs to Peter Higgs!23 In order to overcome these backgrounds, ATLAS and CMS physicists used a number of tricks to cut away as many background interactions as possible while keeping a large enough number of interactions in which a Higgs boson was produced.

4.4. Triggers

The rate of collisions at the LHC is designed so that extremely large enough datasets can be collected in a reasonable time to study comparatively rare processes such as Higgs boson production. The protons are collided in bunches of \(\sim 10^{11}\) protons at a rate of roughly 40 million bunches per second, meaning the time separation between the bunches crossing is only 25 ns.24 The first task in data selection is to reduce the number of events which are recorded to one which is manageable. At ATLAS and CMS, a rate of around 400 events per second is acceptable. A triggering system is used which selects those events which are more likely to contain signal than background and choose which ones to save to disk for analysis. The trigger needs to use fast algorithms to make the decision. In the \(gg \rightarrow H_0\) analysis, the decision is based on the presence of two energetic photons in the event which are ‘isolated’ (meaning they are not surrounded by other particles), which show up as two regions of activity in the EM calorimeters of the detector. Since the most common background processes do not produce such photons, a large fraction of the backgrounds can be rejected while keeping almost 100% of the signal events.

4.5. Event selection

Once a subset of the events has been recorded, they can be further analysed without time constraint. The next step is to use information about each event to decide whether or not to keep or reject it. The information is typically used to determine whether the event is more ‘signal-like’ (one in which a Higgs boson is produced) or ‘background-like’. A wide variety of different criteria are used for such a decision ranging from the properties of the particles in the event, such as their momenta, or event level quantities such as the number of leptons identified in the event. Often a number of different criteria are combined in order to select events by choosing a set of observables \(X = \{x_1, x_2, \ldots, x_N\}\) and selecting only those events for which the values of \(x_i\) are above, below or between some threshold values. These are known as ‘cuts’. These cuts are typically optimised so that one can expect the best sensitivity to the signal either by maximising the expected significance if the Higgs boson exists or minimising the expected upper limit one could place on its production cross-section. Commonly in HEP, the cut values are chosen by using one or more multi-variate analysis (MVA) techniques such

---

23 Mobile numbers in the UK are usually 07 followed by 9 digits meaning there are a total of 999 999 999 possible mobile telephone numbers in the UK. This assumes all 9 digit numbers are equally valid and that Professor Higgs even uses a mobile phone of course.

24 These numbers are for the design specifications of the LHC. These numbers vary as the operating parameters of the LHC change, for example during the first run, the average time between bunch crossings was never less than 50 ns.
as the DTs described in section 2.3.3. For all of the Higgs search channels, the various optimisations were performed using simulation and data from control samples since the signal region of the analysis was blinded to avoid introducing any bias in the final results.

The \( \text{gg} \rightarrow \gamma \gamma \) analysis from CMS uses a number of different DTs, trained for specific purposes. These range from DTs used to select the correct interaction vertex from the proton collisions, to identifying which of the reconstructed photons in the event are likely to be the decay products of a Higgs boson. For the latter, the kinematic properties of an event where a Higgs boson decays to two photons can be used to discriminate it from background events. These include quantities such as the overall magnitude of the momentum of the two photon system, in the plane transverse to the beam, and the direction in which the photons from the decay propagate. All of these properties are combined using a forest of DTs whose output, known as the classifier score, is then used to categorise (or classify) the events. Figure 8 shows the distribution of that output for events used in the \( \text{gg} \rightarrow \gamma \gamma \) analysis \[41\] which have been recorded by the trigger. Also shown are the expected contributions from the backgrounds (blue line) and the Higgs boson when it is produced through different production mechanisms (red and pink shaded). The signal is separated into the four dominant production modes. These are the gluon fusion (ggH), vector boson fusion (VBF) and Higgs-strahlung (WH/ZH) modes, which were discussed in section 4.1, and the ttH mode in which a top and an anti-top quark pair are produced. The signal processes have different distributions in the classifier score since the kinematic properties of the signal event, such as the transverse momentum of the photons to which the classifier is sensitive, will depend on the way in which the Higgs boson is produced.

For larger values of the classifier score, the ratio of the expected numbers of signal to background events increases. Events with a larger score are therefore more likely to contain a Higgs boson. The events are divided into six classes, based on their score. These divisions are indicated by the vertical dashed lines. These boundaries were optimised so that by statistically combining the different classes, the best overall sensitivity to the Higgs boson signal is obtained. The lowest scoring class (indicated by the shaded grey region) are rejected as the inclusion of these events does not yield any significant improvement in the sensitivity. This can be expected since events in this region are much more likely to be background events.

Figure 8. Distribution of the \( \text{H}^0 \rightarrow \gamma \gamma \) classifier score for the data (left-axis) and the contribution expected from signal events (right-axis) \[41\]. The vertical lines indicate the divisions between the different classes of events used in the analysis with the lowest scoring events being rejected altogether. © The Author(s) 2014. CC BY 4.0.
4.6. Local and global p-values

The discovery of the Higgs boson was based on the combination of data from several analyses targeting not only different decay modes of the Higgs boson but also divided into separate categories with different purities, as in the example from the $H^0 \rightarrow \gamma\gamma$ analysis in the previous section. In the end, we will want to quantify the extent of disagreement with $H^0$, the hypothesis of the SM without the Higgs boson, for all the data categories at once. If the agreement is very poor, we can reject $H^0$ and perhaps conclude that something must be added. For this, ATLAS and CMS make use of the likelihood ratio described in section 2.5.6. The analysis was performed as a Raster Scan i.e. at a series of closely-spaced fixed masses $m_{H^0}$.

4.6.1. Likelihood ratio and local p-values. We can think of each of these categories as having a likelihood function $\mathcal{L}_c(\mu, \theta)$; with $\mu$ defined as the ratio of the assumed value to the theoretical cross-section as in section 2.5.6. The statistical combination of the data in all of the independent categories, is performed by taking the product,

$$\mathcal{L}(\mu, \theta) = \prod_c \mathcal{L}_c(\mu, \theta), \quad (4.1)$$

where $\theta$ are the nuisance parameters which may be correlated between different categories. An important point to notice is that the parameter $\mu$ is continuous. One could imagine only being interested in the value $\mu = 1$ since that would yield the signal rates, in every category, as predicted by the SM. However, if nature contained some new physics which altered that rate in some way, or perhaps the SM calculations were incorrect in some fashion, we would not want to miss discovering something, just because it was not exactly the Higgs boson predicted by the SM. In fact, one can think of the $\mu$ parameter as representing a whole family of alternate hypotheses $H(\mu)$ which predict the existence of a ‘Higgs-like’ boson, with similar properties, decay rates etc, but being produced with a different rate. Instead then of comparing $\mu = 0$ with $\mu = 1$, we compare it with $\mu = \hat{\mu}$ which denotes the value of $\mu$ which maximises the likelihood $\mathcal{L}(\mu, \theta)$ for the data we observed, thereby giving the ‘most-likely’ value for $\mu$. The value of $m_{H^0}$ is implicitly included in the definition of $H(\mu)$ so typically, a comparison between $H(\mu)$ and $H^0$ is made for different values of $m_{H^0}$, within some reasonable range in which the search is expected to be sensitive.

The quantity $q_0 = -2 \ln(\mathcal{L}(\mu = 0, \theta^0_{\text{prod}})/\mathcal{L}(\mu_{\text{best}}, \theta_{\text{best}}))$, was used as the test statistic for the discovery of the Higgs boson. Here, $\theta^0_{\text{prod}}$ denotes the values of the nuisance parameters resulting from profiling the likelihood function when $\mu = 0$; similarly $\theta_{\text{best}}$ denotes the corresponding values for $\mu = \mu_{\text{best}}$. Clearly, large values of $q_0$ would indicate that the data would favour strongly the presence of a Higgs boson signal, which we would call a ‘significant’ result. By comparing the observed $q_0$ with its expected distribution (usually determined by simulation, but sometimes available analytically), we can determine the $p$-value $p_0$; this is the probability of a statistical fluctuation, consistent with a particular value of $m_{H^0}$, resulting in $q_0$ being at least as extreme as its observed value, assuming that $H^0$ is true. An example distribution of $q_0$ under $H^0$ is given in figure 9. The observed value of $q_0$ is indicated by the arrow.

The probability to observe a value of $q_0$, under $H^0$, which is larger than the one we observed in the real data is called the local $p$-value $p_0$, for the assumed value of $m_{H^0}$. It can be obtained from the fraction of these pseudo-data sets for which $q_0 > q^\text{obs}_0$. 


where \( f(q_0|H0) \) is the pdf for \( q_0 \) under \( H0 \), for the particular mass being investigated. As the significance of an observation increases, it becomes more and more computationally challenging to produce these distributions with pseudo-data. This issue is overcome through the use of asymptotic approximations which are valid when there are a large number of events. In figure 9, the expected distribution of \( q_0 \) using these approximations, shown as the green curve, agrees very well with that from the toys and hence the local \( p \)-value can be determined analytically. ATLAS and CMS use these asymptotic approximations when calculating \( p \)-values for different values of \( m_{H0} \). The local \( p \)-value for the data collected by CMS is shown in figure 10 using data from each decay mode separately and the combination of all modes. The smallest such value occurs at the large dip with a \( p \)-value of \( \approx 3 \times 10^{-7} \), corresponding to a \( z \)-score of about 5 standard deviations, the predefined requirement to announce a discovery. The dashed line in the plot shows what the median expected value of \( p_0 \) would be at each value of \( m_{H0} \) if the Higgs boson was \( m_{H0} \), according to the SM predictions.

4.6.2. Global \( p \)-value. As explained in section 2.8, we need to convert the local \( p \)-value into a global one; the former gives the chance of a statistical fluctuation giving an effect at 125 GeV at least as significant as the observed one, while the global one is for a range of possible Higgs masses. This is chosen to be a reasonable range over which all of the channels used in the combination are sensitive. Above this range, for example, the branching fractions for the Higgs decays to \( \gamma\gamma \), \( bb \) and \( tt \) channels become negligible compared to that for WW and ZZ. For some channels, like the high resolution channel \( H \rightarrow \gamma\gamma \), there are several wiggles in the local \( p \)-value. This indicates the fact that the peak could occur in a number of different places (or specifically, different values of \( m_{H0} \)) so that a large look elsewhere effect is expected in this channel. Conversely, for channels like the \( H \rightarrow WW^* \rightarrow 2/2\nu \), with
poorer resolution, the $p$-value is rather smooth with respect to $m_{H0}$ meaning if an excess is seen at one value of $m_{H0}$, the same excess will be compatible with other, nearby values of $m_{H0}$ too. Effectively, this suggests that there are fewer independent values of $m_{H0}$ which can be tested. For the combination, for which outcomes are more significant if the excess appears around the same mass in the two high-resolution channels, the look elsewhere effect is somewhere in between. The global $p$-value (using values of $m_{H0}$ in the range shown in figure 10) is $3.4 \times 10^{-6}$, corresponding to a $z$-score of 4.5.

**4.7. Systematics**

As mentioned in section 2.9, systematic uncertainties often result in reducing the overall sensitivity of an analysis and limit the precision with which a measurement can be made. One of the most important aspects of the search for the Higgs boson, or any search for new physics, is being able to estimate how much SM background should be expected after all of the data has been selected. This is because in the test for discovering a new particle, like the Higgs boson, one needs to know how much background, with some limited precision, is there before claiming that we see a signal.

**4.7.1. Background rates.** Many analyses at ATLAS and CMS rely on theoretical calculations and Monte Carlo simulation in order to calculate the number of background events, $b$, in equation (2.5), from a particular process. This could be determined using

$$b = A_b \epsilon_b L \sigma_b,$$

where $\sigma_b$ is the production cross-section for the background process, $A_b$ and $\epsilon_b$ are the overall acceptance and efficiency for background events to pass all of the selections and $L$ is the integrated luminosity. Typically, quantities such as the efficiency, acceptance and luminosity can be measured with an accuracy of a few percent at the LHC. The predicted cross-section can, in some cases, only be known to within 20% to 30%! In ATLAS and CMS, these kind of uncertainties are modelled assuming a log-normal distribution meaning the number of background events becomes
where $\theta$ is the nuisance parameter for the cross-section uncertainty\textsuperscript{25} (of size 20\%). The nuisance parameter is assigned a Gaussian prior probability (or in the frequentist case the likelihood is multiplied by a Gaussian term) centred at 0 with a width of 1. This means that a value of $\theta = 1$ corresponds to a positive shift of the nuisance parameter by one standard deviation and yields a background rate of $1.2b$. Instead, a value of $\theta = -1$ is a negative shift of one standard deviation and yields a value of $b/1.2 \approx 0.83b$. This seems a little odd in that the value of $b(\theta)$ is not symmetric about $b$. However, this is an intentional feature of using log-normals since the number of expected background events can never go below 0.

Suppose we have an experiment where $b = 900$, while the number of signal expected is 90. If we observe 990 events in our data, with no systematic uncertainty on the background, this would constitute a 3\(\sigma\) evidence for our signal. However, when considering a 20\% uncertainty, this excess of events could be explained by assuming $\theta \approx 0.52$, only around half a standard deviation from its assumed value. The significance of the excess would drop to roughly 0.5\(\sigma\) which is not worth getting excited over.

If the rate of expected signal is very small compared to the backgrounds as in this case, uncertainties of this size will make it very difficult to determine whether or not any excess seen in the data is some hint of new physics or just some inaccuracy in the background rate, which can easily be explained by shifting the nuisance parameters a little. Worse yet, adding more data will not solve the problem since typically systematic uncertainties are not statistical in nature so remain a constant frustration as the luminosity increases. In the Higgs boson search at ATLAS and CMS, the majority of the work is understanding these background rates and reducing their systematic uncertainties. Moreover, in real analyses, there are typically several contributing backgrounds, which must be summed to obtain the total, potentially having different sources of systematic uncertainties.

One common way to reduce these kinds of systematic uncertainties is to directly use the data to estimate the background rates.

### 4.7.2. Background shapes

For searches which use kinematic variables to separate the signal from the background, both the rate and the distribution (shape) of the background must be accurately modelled. This often means accounting for additional experimental uncertainties related to the reconstruction of the decay products, such as the momenta of the reconstructed particles in the event. Typically, searches are performed in the tails of the background distributions which make their estimation from MC unreliable.

In the ATLAS and CMS $\mathrm{H}^0 \rightarrow \gamma\gamma$ analyses, the backgrounds are determined by fitting functions to the invariant mass spectra directly in the data. This means that there are no uncertainties associated with the luminosity, cross-section or acceptance and efficiency of the backgrounds. However, there is no such thing as a (systematics) free lunch in particle physics. The problem still arises that one has to select the appropriate function to fit the background. Without relying heavily on theoretical calculations, a number of different smooth functions will provide good fits to the background in this analysis. This choice itself is a systematic uncertainty which can be handled in a number of different ways.

At the time of discovery, the CMS analysis used polynomial functions to model the background whose number of parameters were chosen such that any potential bias from choosing such a function would be much smaller than the statistical uncertainty, thereby rendering it negligible. Since then, the analysis has adopted a method in which this choice of

\textsuperscript{25} The log-normal expression can be thought of as interpreting an uncertainty of 20\% as saying that the cross-section is known to a factor of 1.2. See [43] for a discussion of the use of log-normal distributions.
function is a nuisance parameter which can take on discrete values, each of which represents a particular choice from a list of suitable functional forms [44]. This discrete nuisance parameter is profiled, along with all the other nuisance parameters, effectively leaving the choice of function itself free in the fit²⁶.

Accurately determining both the rate and the shape of the background contributions was not only an issue during the search for the Higgs boson but continues to present a challenge for measuring its properties. As the precision of these measurements increases, due to inclusion of new data, the issue of reducing systematic uncertainties becomes ever more difficult and the techniques used to overcome them become increasingly sophisticated.

4.8. Measuring the Higgs boson’s properties

Since the discovery of the Higgs boson, the focus of the experimental work has been towards measuring various properties of the Higgs boson and testing whether they agree with SM or possibly indicate the existence of new physics waiting to be discovered. In this section, we give just two examples, the spin of the Higgs boson and its mass.

4.8.1. Spin of the Higgs boson

An important property of the Higgs boson in the SM is that it should have zero intrinsic spin. If the spin of the discovered particle were something else, it could not be identified with the SM Higgs boson. Since the spin of a particle is not a continuous quantity, a hypothesis test is performed to determine whether or not the hypothesis of a spin-0 particle can be rejected in favour of an alternative hypothesis such as spin-2. Experimentally, these two hypotheses would yield different angular distributions of the Higgs decay products. For example, in the \( H^0 \to ZZ^* \to 4\ell \) channel, the angles defined using the directions of the leptons will depend on whether or not the decaying boson has spin-0 or spin-2. Using the distribution of these angles in the data the two hypotheses, spin-0 or spin-2, can be distinguished. A test statistic is defined using the ratio of the likelihoods to observe the distribution in the data under each of these two hypotheses. As usual, the convention is to take the log of this ratio so that the numbers are of the order 1, and multiply by a factor of \(-2\) so that the quantity asymptotically approximates to the difference of the weighted sum of squares for the two hypotheses (see section 2.5.6). Figure 11 shows the distribution of the test statistic under each of the two hypotheses, generated using pseudo-experiments. The shaded blue or solid yellow histograms show the distribution that we would expect if the decay products originated from a spin-2 boson or the SM Higgs boson (with spin-0), respectively.

The red arrow indicates the value of the test statistic evaluated on the actual angular distribution which is observed in the data collected by CMS [45]. From these distributions two \( p \)-values are calculated. The first (\( p_{2>0} \)) is the probability to observe a value which is at least as large as the one observed, assuming the spin-2 hypothesis. The value of \( p_{2>0} \) in this case is around 0.008, which corresponds to around 3 standard deviations from the mean of a normal distribution. The other \( p \)-value (\( p_{0<2} \)) is defined as the probability to observe a value which is smaller than the one observed, assuming the spin-0 hypothesis. In this case, \( p_{0<2} \) is around 0.62. From this we can conclude that the hypothesis of a spin-0 particle is certainly more favourable than a spin-2. Note however, that in this kind of test, had the observed data yielded a value of the test statistic around -20 or smaller, neither the spin-2, nor the spin-0 hypothesis would be a very good candidate and both could have been rejected. In such a case, one would have to revisit the list of possible models to test and it

²⁶ The method is an analogue of the continuous nuisance parameter case, for which Wilks’ theorem applies. However, the theorem does not apply to the discrete case, and so needs to be validated empirically for each particular analysis.
would have probably sparked a lot of interest in the particle physics community. When performing hypothesis tests of this kind, ATLAS and CMS use the CLs criterion (defined as $-2 \ln \left( \frac{L_{m}}{L_{b}} \right)$) to avoid excluding in cases where the two distributions lie almost on top of one another and the observed value of the test statistic lies somewhere in the tails of the distributions. In the example shown in figure 11, $CL_s = 0.02$ and so with the usual 95% confidence level criterion, this particular spin-2 model is excluded.

4.8.2. Mass of the Higgs boson. A major task was to precisely measure the mass of the Higgs boson since, in the SM, once this is known, all of the other properties of the Higgs boson can be predicted. In order to get the best measurement of the mass, ATLAS and CMS used their datasets from the two high resolution decay channels, $H \to \gamma\gamma$ and $H \to ZZ → 4l$. The likelihoods from ATLAS and CMS were combined to give an overall likelihood for the combined dataset [46]. Figure 12 shows the result of that combination. The figure shows $-2 \ln \Lambda(m_{H^0})$ (where $\Lambda(m_{H^0})$ is the ratio of the profile likelihood at a particular value of $m_{H^0}$ to the value obtained at $m_{H^0} = m_{H^0,\text{best}}$) for the sub-combinations of each decay mode separately and the overall combination. In each case the minimum of the curve indicates the value of $m_{H^0}$ which best fits that particular dataset. The widths of the curves indicate the level of uncertainty in each of the measurements. By appealing to Wilks’ theorem, the uncertainty of the measurement is estimated from the values at which the curve increases by 1 from the minimum i.e the regions in $m_{H^0}$ for which $-2 \ln \Lambda(m_{H^0}) \leq 1$. Naturally, the width of the combined measurement is smaller than in either of the two channels alone. This is expected since the data used in each of the two channels are statistically independent from one another. Since the statistical component of the uncertainties is much larger than the systematic component, as shown by the small differences between the dashed and solid lines, the systematics play a small role in the combination.
Many other properties of the Higgs boson are currently under study at ATLAS and CMS. As more and more data are collected, the precision to which these properties can be determined and the range in the types of properties which can be measured will increase. Up to this point however, the measured properties of the discovered particle are consistent with those that are predicted for the Higgs boson in the SM. However, particle physicists are ever hopeful that at some point we will see something we did not expect and that is when the real challenge and excitement will begin.

5. Conclusions

Some 50 years after its suggested existence, the search for the Higgs boson was successful. The analysis involved many interesting statistical techniques, including:

- Blind analysis, to ensure that the result was not subject to (subconscious) personal bias by the analysers.
- Multivariate methods for selecting data samples enhanced in the Higgs signal.
- Careful treatment of possible systematic effects.
- Combination of many event categories and decay modes of the Higgs.
- Calculation of the look elsewhere effect factor.
- Estimating the significance of the observed effect.

Searches for a host of other types of New Physics have not yet resulted in any discovery claims, and upper limits have been set on the production rates of the possible relevant particles or effects. This has often resulted in various models of beyond SM physics being ruled out in some regions of their parameter space. Maybe in time some of these models will be abandoned.

To date, the CMS collaboration has produced around 200 papers on upper limits for various processes.

We conclude with a list of some resources that are useful for understanding and/or implementing statistical techniques for searches.
• There are numerous books [47–53] written by High Energy Physicists specifically for a Physics audience.
• The Particle Data Group’s review of particle properties [23] contains short sections on probability, statistics and Monte Carlo simulation.
• The PhyStat series of workshops have been devoted to statistical issues that appear in HEP analyses (see [54, 55] for the latest workshops, from which the earlier ones can be found).
• The large experimental collaborations have Statistics Committees, which deal with statistical issues relevant to their collaboration. Their websites contain useful information; the CDF one [56] is the most accessible.
• A pedagogical overview of the practicalities of implementing the statistical prescriptions used at the LHC is given in [57] and detailed descriptions of the various test statistics and their asymptotic distributions as used in searches at the LHC can be found in [16].
• There are many different software packages available for statistical calculations, for example the package RooStats [58] is commonly used in HEP.

Finally good wishes for your analyses, and every success in your searches for New Physics.

Acknowledgments

We wish to thank our colleagues on the CMS Statistics Committee for many patient and enlightening discussions. Conversations and e-mail exchanges with Sara Algeri, Mattias Blennow, Bob Cousins, Glen Cowan, David van Dyk, Tom Junk, Xin Qian, Aixin Tan and Ofer Vitells are all gratefully acknowledged. Finally, we would like to thank the UK Science and Technologies Facilities Council who, in part, provided funding for this work.

ORCID iDs

Nicholas Wardle © https://orcid.org/0000-0003-1344-3356

References


37
[26] Lyons L 2013 Discovering the significance of 5 sigma (arXiv:1310.1284)
[31] Lyons L, Perrin Y and James F E 2000 Confidence Limits Workshop https://cds.cern.ch/record/411537

[43] Limpert E and Stahel W A 2017 The log-normal distribution Significance 14 8–9


[51] Lista L 2015 Statistical Methods for Data Analysis in Particle Physics (Berlin: Springer)


[58] RooStats https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome