The Role of Statistics in the Discovery of a Higgs Boson

David A. van Dyk

Statistics Section, Imperial College London, London SW7 2AZ, United Kingdom;
email: d.van-dyk@imperial.ac.uk

Annu. Rev. Stat. Appl. 2014. 1:5.1–5.19

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-062713-085841

Keywords

detection, exclusion, hypothesis testing, look-elsewhere effect, particle physics, Poisson models, sensitivity, upper limits

Abstract

The 2012–2013 discovery of a Higgs boson appears to have filled the final missing gap in the Standard Model of particle physics and was greeted with fanfare by the scientific community and by the public at large. Particle physicists have developed and rigorously tested a specialized statistical tool kit that is designed for the search for new physics. This tool kit was put to the test in a 40-year search that culminated in the discovery of a Higgs boson. This article reviews these statistical methods, the controversies that surround them, and how they led to this historic discovery.

1. THE SEARCH FOR THE GOD PARTICLE

The recent empirical confirmation of a Higgs boson (ATLAS Collab. 2012a,c; CMS Collab. 2012a,b) appears to be "the culmination of the experimental verification of the Standard Model" (Ellis et al. 2012, p. 2), which describes how the fundamental particles and forces between them (the electromagnetic, weak nuclear, and strong nuclear forces) give rise to all matter in the universe and most of its higher interactions. Within the Standard Model, the Higgs boson imparts mass to some fundamental particles that would otherwise be massless. It was the last of the fundamental particles predicted by the Standard Model to be experimentally verified. First predicted theoretically in 1964 (Higgs 1964), the Higgs boson was incorporated into the Standard Model in 1967 (Salam 1968, Weinberg 1967) in a unified theory for the weak and electromagnetic interactions. The 2012 discovery was the capstone of a 40-year international search and provided "closure on a half century of theoretical conjecture" (Ellis et al. 2012, p. 2). In his 1993 book, *The God Particle: If the Universe Is the Answer, What Is the Question?* (p. 22), Nobel Prize–winning physicist Leon M. Lederman explains that

> [t]his boson is so central to the state of physics today, so crucial to our final understanding of the structure of matter, yet so elusive, that I have given it a nickname: the God Particle.

Lederman goes on to say that the name "Goddamn Particle" might be a more appropriate name for the Higgs boson, "given its villainous nature and the expense it is causing."

The empirical discovery took place at the European Center for Nuclear Research (CERN) in Geneva using the Large Hadron Collider (LHC). Situated on the Swiss-French border, the LHC is housed in a circular tunnel that is 27 km in circumference and 100 m belowground. The LHC is a large particle accelerator in which bundles of protons circulate in opposite directions and collide within the LHC detectors, producing new particles by converting kinetic energy into mass. The detectors track the trajectories of all these particles, determine their momentum, and provide clues as to their identities. Each observed particle collision, perhaps subject to certain selection criteria on the observables, is referred to as an event. The detectors involved are enormous. The ATLAS detector at the LHC, for example, is as large as a seven-story building and is made up of 10^8 channels of electronics.

The particles produced in proton-proton collisions at the LHC may be unstable and may decay further before they are detected. The final detectable set of particles resulting from the collision is known as the final state. A Higgs boson is produced only very rarely and has numerous possible decay paths, known as decay channels, each with different observed final states. The Standard Model precisely predicts the probabilities of the various decay channels as a function of the mass of the Higgs boson, and in principle, the relative frequency of the observed final states can be used to estimate the Higgs boson's mass. In practice, it is much more efficient to estimate the mass directly: The mass of a particle that has decayed can be recovered (with uncertainty) from the energies and momenta of the particles that it decays into via Einstein's famous equation,

$$E^2 = p^2c^2 + m_0^2c^4, \qquad 1.$$

where E, p, and m_0 are energy, momentum, and rest mass, respectively, and c is the speed of light. (Mass increases with momentum, so letting m be the mass with momentum p yields the more familiar form of Einstein's equation: $E = mc^2$.)

Statistically, searches for new physics, that is, for new physical particles, typically take one of two forms. In the first, we simply count the total number of events. These events comprise those stemming from known physics, which are known as background events, and possibly additional events from new physics. In the second form, we consider the distribution of the so-called invariant

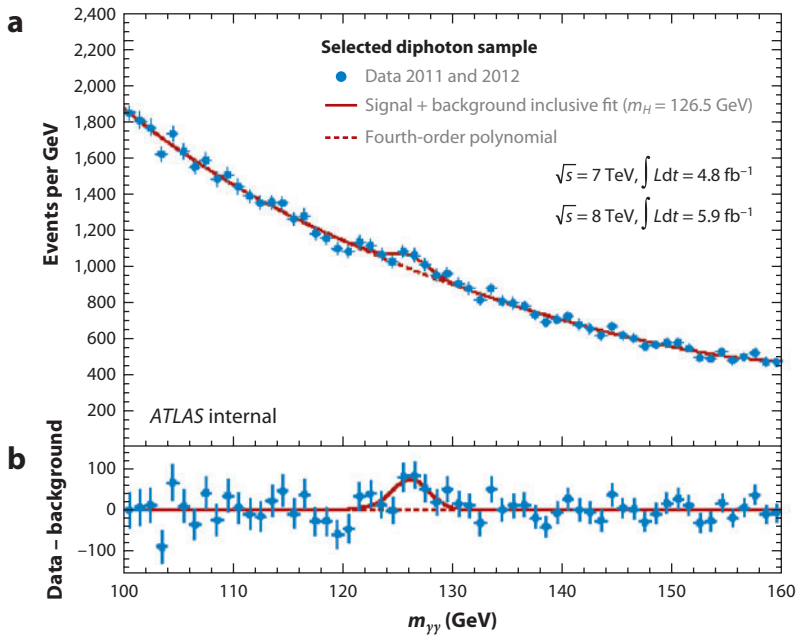

Figure 1

A subset of the 2011 and 2012 ATLAS event counts. (*a*) The observed event counts in the $\gamma\gamma$ decay channel in each invariant mass bin. (*b*) Residuals from the fitted background model. The excess counts with invariant mass near 125 GeV are apparent. The background models are disused in Section 3, and the source model is based on the Standard Model with Higgs mass (m_{H}) of 126.5 GeV. The quantity $\sqrt{s}$ is the energy of the collider, and $\int Ldt$ is the volume of data at each energy used in the plot. Copyright CERN.

mass of the particles formed in high-energy collisions. The invariant mass of the particles is the sum of their energies in their joint center of mass. When compared with known background physics, new physics is expected to result in excess events with invariant mass near that of the new physical particle. **Figure 1** illustrates this situation for the Higgs boson search. Either type of search can be formalized statistically as a hypothesis test: in the first form as a contaminated Poisson count and in the second as a search for a bump above a background distribution. In either case, the statistical search may involve one of several outcomes:

1. The conclusion that the data are inconsistent with the null hypothesis of the known background physics but consistent with the hypothesized new physics, resulting in the sought-after discovery of a new physical particle.
2. The conclusion that the data are inconsistent with the hypothesized new physics.
3. An upper limit on a possible signal strength generated by the new physics.
4. A determination that the experiment is not sensitive enough to distinguish between new physics and the background.

Outcome 4 involves experimental design and can be determined, to a certain extent, in advance of data collection but may depend on nuisance parameters in the background that are estimated from data. The upper limit in outcome 3 is the end point of an interval estimate for the signal strength. Although interval estimates have been well studied in statistics, a rather exacting set of requirements for the statistical properties of intervals and upper limits has led to a robust particle

THE SEARCH FOR THE PENTAQUARK

In the mid-2000s, a controversy arose around a number of experiments that reported strong evidence either for or against the discovery of the so-called pentaquark (e.g., Part. Data Group 2006, 2008). Strong evidence in favor came from the CLAS Collaboration (2003), which reported a p-value of less than 1 in 17 million, reaching the high threshold set by particle physicists for discovery. In the midst of both confirmatory and contradictory findings by other experiments, the CLAS Collaboration (2006) replicated their earlier experiment but collected six times more data. This time, the result was convincing evidence against the existence of the pentaquark. A follow-up Bayesian analysis of the 2003 experiment reported a Bayes factor that slightly favored the model without the pentaquark, an astonishing shift from the p-value of 1 in 17 million reported with the same data (CLAS Collab. 2008); a critical comment on the choice of prior, however, appears in Cousins (2008). The pentaquark discovery has now been completely discredited (Part. Data Group 2008), but the entire episode encouraged particle physicists (and their publishers) to be extraordinarily wary of claims of discovery, even with exceptionally extreme p-values.

physics literature (see Section 4.2). Together, outcomes 1 and 2 constitute a generalization of the standard statistical approach to hypothesis testing. Particle physicists are interested in the possible conclusion that the background model is sufficient, not simply in the inability to reject the background model. Evidence that there is no Higgs boson—or, more precisely, that should the Higgs boson exist its mass would be outside the range under consideration—would have been of direct scientific interest. As such, a statistical framework that allows not only for rejecting the null model but also for "rejecting the alternative" model is necessary. Moreover, rejection of the null model is not sufficient for discovery. It must also be shown that the alternative model adequately explains the data.

This article reviews the specialized statistical methods used in the discovery of a Higgs boson that were developed by particle physicists, sometimes in collaboration with statisticians. In some cases, the same or similar methods have been used in other searches for new new physics, for example, in the 1995 discovery of the top quark (e.g., Campagnari & Franklin 1997) and in the 2003 false discovery of the pentaquark (see sidebar, The Search for the Pentaquark). At the same time, particle physicists have developed statistical methods that are not related to the search for new physics such as the Higgs boson, for example, methods designed to study the lifetime of particular particles. In this article, we focus exclusively on methods related to the search for a Higgs boson. Although these methods are somewhat idiosyncratic to high-energy physics, they have been adopted in related fields, such as particle astrophysics, in which they are used in the search for dark matter (e.g., Weniger 2012). As in the search for a Higgs boson, the particle astrophysics searches involve looking for a line above background in multibin Poisson data. Searches of this type are also common in high-energy astrophysics (e.g., Kashyap et al. 2010, Park et al. 2008, Protassov et al. 2002).

Our review begins in Section 2 with an overview of the data collection and the statistical philosophy that underlies its analysis. In Section 3, we discuss the preprocessing of the data and the flexible models used to describe events stemming from known physics that are used to constrain the expected background counts. The statistical methods used for assessing sensitivity, computing upper limits, and declaring a detection are first discussed in Section 4 in the context of a single-bin analysis. These methods form the building blocks of the methods described in Section 5 that are used for detection in a multibin, multichannel analysis and the ultimate estimation of the Higgs mass.

2. DATA COLLECTION AND STATISTICAL PHILOSOPHY

There are seven particle detector experiments at the LHC. Two of them, ATLAS and CMS, were involved in the discovery of a Higgs boson. Both ATLAS and CMS are capable of tracking the trajectory of the final-state particles resulting from proton-proton collisions, determining their momenta and/or energy, and establishing their identities. They identified this particle primarily through two decay channels: a Higgs boson decay into two photons ($H \rightarrow \gamma\gamma$) and a Higgs boson decay into two Z bosons ($H \rightarrow ZZ$), each of which decays, in turn, into two leptons (either electrons or muons). Both detectors have sufficient resolution to identify a narrow peak above background in the distribution of the invariant masses of the decay particles (e.g., Della Negra et al. 2012).

The current LHC experiments are following up on previous studies: the Large Electron Positron collider, which operated at CERN from 1989 until 2000, and the Tevatron, which operated at Fermi National Accelerator Laboratory, outside Chicago, between 1983 and 2011. Although these studies did not experimentally confirm the existence of the Higgs boson, together they excluded a range of masses and focused the search on a mass between 114 and 130 GeV (e.g., Della Negra et al. 2012). In 2012, the ATLAS and CMS experiments simultaneously announced the discovery of a previously unknown boson. Combining data from the two primary decay channels with other channels, both experiments found a mass near 125 GeV, with p-values of less than one in three million (ATLAS Collab. 2012c, CMS Collab. 2012a). The behavior of this particle broadly matches what is predicted by the Standard Model for the Higgs boson. It decays into the predicted channels and at the predicted rates, at least up to the experimental uncertainties. In an abundance of caution, however, the experiments refrained from calling the new particle "the Higgs boson" and instead referred to it as "a Higgs boson" and continue to study its properties and note their similarities with the Standard Model's predictions for the Higgs boson (Del Rosso 2012).

Thousands of researchers from scores of institutions collaborate in these high-profile experiments. Procedural decisions including the choice of statistical methods are made by committees that face significant scrutiny, particularly given the expense involved with the experiments. They are among the most costly in human history, and researchers are decidedly adverse to ambiguity in their findings, including those stemming from their choice of statistical method. The result is that high-energy physicists tend to be conservative in their choice of methods and demanding in their search for mathematical and statistical rigor. Procedures must be well understood, well defined, and fixed in advance. The subjective aspects of statistical analyses, such as the choice of models and methods, are highly scrutinized in, for example, a long-running series of workshops on statistical methods in particle physics known as PHYSTAT. The proceedings of these meetings include thoughtful investigations into foundational issues on such topics as model selection and the quantification of uncertainty (e.g., Lyons et al. 2004, 2008, Lyons & Unel 2005, Prosper & Lyons 2011). Broadly speaking, particle physicists are adverse to Bayesian methods owing both to the perception that they are subjective in nature and to an insistence on well-quantified frequency properties (e.g., Cousins 1995, Mandelkern 2002), despite advocates for Bayesian model selections (e.g., Berger 2008) and the provocative Bayesian analysis in the search for the pentaquark (see the sidebar). Bayesian procedures, however, are generally accepted for dealing with nuisance parameters (Cousins & Highland 1992, Heinrich et al. 2004), and some physicists advocate for a much more central role for Bayesian thinking in searches in high-energy physics (Anderson 1992).

As discussed below, strict adherence to frequency-based statistical properties in complex scientific problems that involve a certain number of systematic errors and potential model misspecification leads to practical difficulties. Even when unusual outcomes occur more frequently

than is plausible under a postulated model, particle physicists are loath to dismiss a model or an expensive observation a posteriori for fear of biasing results. To avoid experimenter bias, an unknown constant, known as a secret offset, is typically added to observations (e.g., to the masses) and not revealed until analyses are complete (e.g., Mandelkern 2002). How to deal with manifest disagreement between data and model without sacrificing data, reducing efficiency, or incurring bias remains an important open question.

3. DATA PROCESSING AND BACKGROUND MODELS

The LHC produces millions of proton collisions per second, but most of these are uninteresting in that they involve well-understood physics. Rather than storing all of these events, the experiments have triggers that aim to make very fast decisions as to whether each event is interesting and save only approximately 100 per second. Although this is a small fraction of the total events, it still could result in 10^{10} saved events for each experiment over the expected 15-year life span of the LHC (Lyons 2008). The experiments make further cuts to data within each decay channel before they are formally analyzed. On the basis of the characteristics of the observed final states of the events, these cuts aim to prune the data, reduce the background events, and focus the analysis on a subset of the data wherein new physics is more likely to be observed. Indeed, the fraction of stored events (that survive the trigger) that involve new physics can be as low as 10^{-8} (Lyons 2008). In the actual Higgs boson discovery, for example, there were only a few hundred events in the two primary decay channels above the background rates near the estimated Higgs mass (**Figure 1**). Data pruning may begin with simple cuts based on the values of individual variables, such as the number of tracked particles associated with the event, and proceed with more formal supervised learning algorithms (e.g., Friedman 2005, Roe et al. 2005). Simulated background and signal events or manually classified real events are used to train the learning algorithms.

Once a subset of events in each detection channel is identified for analysis, it is stratified into relatively homogeneous categories. In analogy to stratified sampling, the homogeneity of the signal-to-background ratios (and invariant mass resolutions) within each category increases the statistical power for identifying possible excess events above background that are due to new physics (ATLAS Collab. 2008). Each of the five channels is split into a number of subchannels on the basis of the directly observable characteristics of the decay, knows as tags. The subchannels are then further divided into categories. In most cases, this classification is based on the type or flavor of the final state. In the $H \rightarrow ZZ$ channel, for example, the categories are determined by the decay path of each of the Z bosons into either electrons or muons. In a few cases, boosted decision trees are used to stratify the subchannels into categories. Using simulated background and signal events, physicists employ boosted decision trees to predict which simulated events involve a Higgs boson decay on the basis of the characteristics of the momenta and energy of the observed particles and the presence of particular particle types in the final state (e.g., ATLAS Collab. 2012b). Cut points on the prediction of the fitted tree are then used to separate events into categories. The number and location of the cuts are chosen to minimize the expected upper limit on the signal, subject to constraints on the size of the categories.

Parametric background models are used to quantify the distribution of invariant masses due to known physics (**Figure 1**). The primary goal is to search for excess events above this background that can be attributed to new physics, such as a Higgs boson. Although the Standard Model predicts the shape of the background distribution, empirical models are preferred in part because of the extreme and ad hoc nature of the cuts—only events deep in the tails of their distribution are used in the data analysis. The cuts also mean that a different background distribution is expected in each category (stratum) of each channel. Thus, different functional forms as well as

fitted parameters are used for the several category-channel combinations. In cases in which there are relatively few observations, simpler background models are used. The functional forms are typically determined by simulating large data sets under the model, applying the same cuts used in the real data, and comparing the resulting fits. Various models are considered. For example, the ATLAS Collaboration (2012b) examined single- and double-exponential functions, Bernstein polynomials up to the seventh order, exponentials of second- and third-order polynomials, and exponentials with modified turn-on behavior. For low-count categories, the ATLAS Collaboration found exponential functions to be sufficient, whereas for larger-count categories exponential functions of second-order polynomials or fourth-order Bernstein polynomials were used (**Figure 1**). The CMS Collaboration employed similar background models. Once the parametric models were determined, they were fit to data, as described in Section 5.1.

The distribution of background events underlies the importance of the $H \to \gamma\gamma$ and $H \to ZZ$ channels in the discovery of a Higgs boson. Together, they have a Higgs decay probability of less than 1%, meaning that only approximately 1% of Higgs decay paths involve these channels. Because of their advantageous signal-to-background rates, however, these rare events are much easier to identify above background.

4. DETECTION AND EXCLUSION IN A SINGLE-BIN ANALYSIS

4.1. A Simplified Model

Consider a single-bin analysis, a simplification of the actual detectors that have multiple bins in each category of each channel. We introduce this simplification to focus attention on several statistical issues and postpone the complexities of the full model until Section 5. The single-bin detector records a number of background events and perhaps additional events above background. We model the observed event count as

$$N \sim \text{Poisson}(\beta + \kappa\mu), \qquad 2.$$

where β is the expected background count and κ is the expected Higgs boson count under the Standard Model, so that $\mu = 0$ corresponds to a background-only model and $\mu = 1$ to the model with a Higgs boson. For the moment, we assume that both β and κ are known. In Section 5, we consider the situation in which neither is known and must be specified in terms of several fitted nuisance parameters. Although we expect that μ is either zero or one, it is generally treated as a continuous parameter as a hedge against minor systematic or unknown experimental errors and to allow for the discovery of new physics that does not adhere to prior expectation.

Here we discuss the four possible search outcomes described in Section 1, which involves three statistical tasks. First, the possible detection in outcome 1 is formulated in terms of a hypothesis test that compares $\mu = 0$ with either $\mu = 1$ or, more generally, $\mu > 0$. In the absence of a detection, the upper limit, μ_{UL}, in outcome 3 is computed on μ. If μ_{UL} is small enough that any possible deviation from background is too small to be consistent with what is predicted for new physics, we can conclude that there is evidence to exclude new physics, as in outcome 2. This outcome can also be formulated as a hypothesis test with the roles of the hypotheses reversed. Finally, as described in outcome 4, we can analyze the power of the hypothesis test to distinguish between new physics and background.

4.2. Intervals and Upper Limits

In a *Statistical Science* discussion paper, Mandelkern (2002) reviews several possible confidence intervals and upper limits for μ, along with those for the related problem of estimating bounds

on a Gaussian mean that is known to be nonnegative. Mandelkern focuses on the situation in which the classical Neyman–Pearson confidence interval for μ may be very narrow or even empty because the observed count, N, is smaller than the expected background. In this situation, there may be few or no values of μ that are consistent with the small value of N, resulting in a small or empty interval. Although relatively unlikely if Equation 2 is correctly specified, difficulties in estimating and incorporating statistical and/or systematic uncertainties compound the problem. The narrowness of such intervals falsely conveys a lack of experimental uncertainty when, in actuality, the discrepancy between data and model suggests substantial uncertainty. Handling this discrepancy from a frequency-based perspective poses significant challenges. Either altering the model after observing the data or discarding data because of a lack of fit upsets frequency properties. Bayesian methods tend to avoid empty intervals but can be quite narrow, and many particle physicists find the reliance on prior distributions undesirable (e.g., Cousins 1995). The choice of prior distribution is especially influential in the important case of upper limits for weak signals. Mandelkern (2002) describes an understandable if somewhat unrealistic goal of obtaining robust statistical methods that can be applied uniformly, that largely maintain the power obtained with a correctly specified model, and that yield reasonable results even with slight model misspecification.

Particle physicists take these challenges quite seriously and, along with their statistical collaborators, have proposed several approaches to address them. Because these methods generally do not alter the Poisson model formulated in Equation 2, they do not address the potential misspecification that worried Mandelkern. Instead, they take advantage of flexibility in the construction of confidence intervals and limits to obtain quantities with more desirable features. The most important of these is the so-called unified approach of Feldman & Cousins (1998). These authors note that the common procedure of deciding whether to report a one- or two-sided interval (i.e., an upper limit or a central confidence interval) based on the observed data destroys the frequency properties of the intervals. One might, for example, provide an upper limit in the absence of a detection and provide a central confidence interval if the source is detected, a practice Feldman & Cousins call flip-flopping. To avoid the need for flip-flopping, they propose what they call a unified approach that smoothly transitions from upper limits to confidence intervals as the source becomes statistically significant. The method is based on inverting the likelihood-ratio test and, unlike the classical construction, greatly reduces the troublesome occurrence of empty intervals or upper limits equal to zero.

Mandelkern (2002) observes that empty or very short intervals arise when the observed data are a priori unlikely, but the method for computing intervals or limits does not make use of this fact. If, for example, we observe $N = 0$, we know that the background count, N_B, is zero and that N equals the source count. Thus, we can replace Equation 2 with $N \sim \text{Poisson}(\kappa\mu)$. More generally, Roe & Woodroofe (1999) propose conditioning on $N_B \leq N$ in the construction of intervals with conditional frequency coverage. The resulting intervals are never empty and have the appealing property that they do not depend on β when $N = 0$. Of course, from a Bayesian perspective it is sensible to condition on the observed data, whether or not they are a priori unlikely. The beauty of Roe & Woodroofe's approach is that it delivers some of the advantages of a Bayesian procedure in a way that is palatable to researchers unaccustomed to the use of prior distributions in scientific inference.

There are numerous other frequency, Bayesian, and hybrid approaches in the physics literature. Mandelkern (2002) provides a good introduction.

4.3. Detection and Sensitivity

The likelihood-ratio test is the basis for formal detection and, under the unified approach, the construction of upper limits and intervals. Consider the null hypothesis, $H_0 : \mu = \mu_0$, where μ_0

is typically zero but other values may be used when inverting the test to construct intervals or limits. As discussed above, we may be interested in either the sharp alternative, $H_A : \mu = 1$, or the composite alternative, $H_A : \mu > 0$. In either case, however, the composite is typically used to formulate the likelihood ratio test statistic:

$$T = -2 \ln \left(\frac{L(\mu_0|N)}{\sup_{\mu \geq 0} L(\mu|N)} \right). \quad 3.$$

Other forms have also been also considered, however, and an extensive review is provided by Cowan et al. (2011a). In more complex models involving nuisance parameters (see Section 5), both the numerator and the denominator of Equation 3 are profiled over the nuisance parameters.

In the spirt of a power calculation, it is desirable to quantify the sensitivity of the test without reference to the observed data. Particle physicists typically summarize sensitivity by using the median value of the sampling distribution of μ_{UL} under H_0 (Lyons 2008). As an alternative, Punzi (2004) suggests the smallest value of μ that obtains a given statistical power for an α-level test. That is, Punzi's sensitivity, which we denote μ_{sen}, is the smallest value of μ that results in an α-level detection at least a given percentage, γ_{crit}, of the time. Confusingly, high-energy astrophysicists refer to μ_{sen} as the upper limit (Kashyap et al. 2010).

4.4. Treating the Standard Model as the Null Hypothesis

Because the Standard Model predicts $\mu = 1$, we are especially interested in whether μ_{UL} and/or μ_{sen} is greater than or less than one. Thus, particle physicists have designed a set of tools that aim to investigate both the possibility of excluding $\mu = 1$ and the sensitivity of the test to this specific alternative value. Consider the hypothesis test that compares the two precise hypotheses, namely $H_0 : \mu = 0$ versus $H_A : \mu = 1$. Noting that larger values of T in Equation 3 are more significant under H_0, we conventionally compute the p-value, $p_0 = \Pr(T \geq t_{obs}|H_0)$, where t_{obs} is the observed valued of T, and reject H_0 if p_0 is sufficiently small—that is, if $p_0 \leq \alpha$. Less conventionally, particle physicists also interchange the roles of the null and alternative hypotheses and compute the corresponding p-value, namely $p_A = \Pr(T \geq t_{obs}|H_A)$ (**Figure 2*b***). [Sometimes p_A is defined as one minus this tail probability (e.g., Lyons 2008).] Because large values of T are more significant under H_0, smaller values of T are more significant under H_A, and a small value of $1 - p_A$ is potential evidence for "rejecting" H_A in favor of H_0, that is, excluding the hypothesized new physics. Particle physicists are more concerned about false discovery than with missing a signal and use a much more stringent threshold for "rejecting" H_0 than for "rejecting" H_A; we label these thresholds α_0 and α_A, respectively.

Table 1 describes four possible cases based on the values of p_0 and $1 - p_A$. For example, the standard procedure is to reject H_0 in favor of detection if $p_0 \leq \alpha_0$. However, if we find that $1 - p_A$ is also small, we are in the awkward position of being suspicious of both the models under consideration. This problem is illustrated with $t_{obs} = t_{III}$ in **Figure 2*c***, which shows why declaring a detection in this case would raise questions: The data are unlikely with or without the Higgs boson. (We use subscripted Roman numbers to refer to the four cases in **Table 1**.) If neither p_0 nor $1 - p_A$ is small, the data are consistent with both models, and again we can neither exclude new physics nor declare a detection (t_{II} in **Figure 2*a***). The more straightforward cases arise when one of p_0 and $1 - p_A$ is small and the other is large and we can either declare a detection or exclude the possibility of new physics.

Figure 2 also illustrates how the distribution of T may differ under H_0 and H_A. The ideal case appears in **Figure 2*c***, where the distributions are well separated. In principle, the observed value of T should easily distinguish between the hypotheses leading to either a detection or an

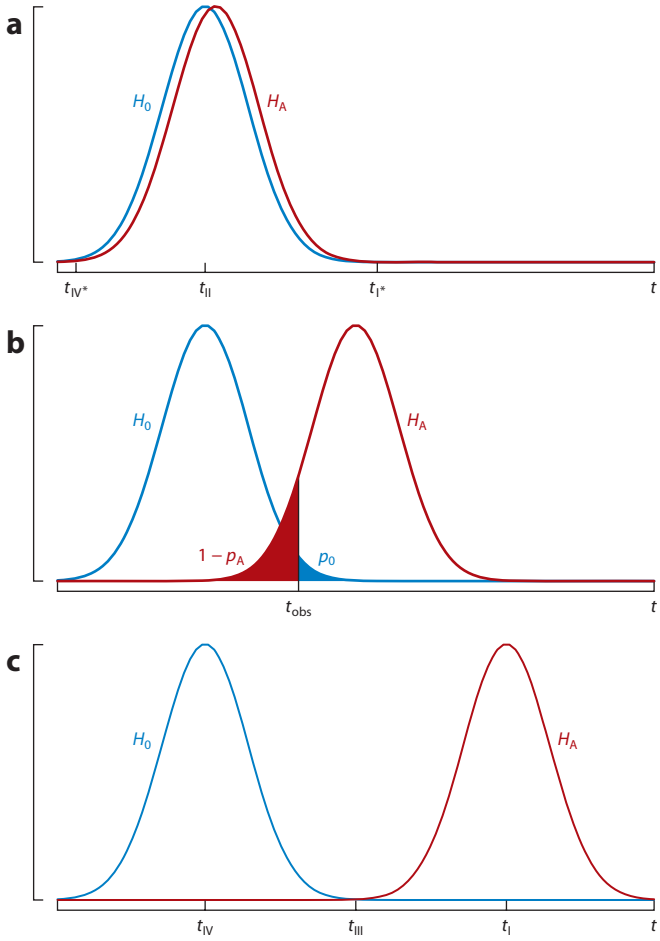

Figure 2

The distribution of T under H_0 and H_A. The definitions of p_0 and p_A are illustrated in panel *b*. The test shown in panel *a* is insensitive: Although t_{I^*} results in a small p_0, it is also unlikely under H_A, drawing a "detection" into question. The CL_S criterion aims to address the similarly problematic exclusion illustrated with t_{IV^*} (see Section 4.5). In panel *b*, the test is more sensitive and the distributions of T differ under the two models. In this case, there is a range of values of t_{obs} that are likely under H_A, but unlikely under H_0, which correspond to clear detections. Similarly clear exclusions occur when t_{obs} is likely only under H_0. Finally, in panel *c* the distributions are still more separated, which corresponds to a very sensitive test but permits intermediate values of t_{obs} that are unlikely under both models. The values t_I, t_{II}, t_{III}, and t_{IV} correspond to the four cases listed in **Table 1**. These plots are based on figure 2 of Lyons (2008).

Table 1 The four actions in a simple versus simple hypothesis test

	$p_0 \leq \alpha_0$	$p_0 > \alpha_0$
$1 - p_A > \alpha_A$	Case I: detection	Case II: both models consistent with data
$1 - p_A \leq \alpha_A$	Case III: neither model consistent with data	Case IV: exclusion

exclusion, as with the illustrated values of $t_{obs} = t_{I}$ and $t_{obs} = t_{IV}$, respectively. If an intermediate value, $t_{obs} = t_{III}$, is observed in which $p_0 \leq \alpha_0$ and $1 - p_A \leq \alpha_A$, we can make no decision but might look for systematic errors or problems in the model or data. The other extreme appears in **Figure 2*a***, where the distribution of T is hardly distinguishable under the two hypotheses; we say such a test is insensitive to the hypotheses.

Two important special cases of those listed in **Table 1** occur with an insensitive test. Although a small value of p_0 formally leads to a detection, a value such as $t_{obs} = t_{I^*}$ in **Figure 2*a*** is nearly as unlikely under H_A as it is under H_0. We refer to this special instance of case 1 as case 1*. It occurs when there are excess counts beyond what is predicted by the Standard Model with the Higgs boson. Although this means that neither hypothesis may be satisfactory, it points to problems beyond the formal choice between H_0 and H_A. Case I* is quite unlikely because particle physicists tend to use very stringent detection thresholds. Of more concern is case IV*, which is illustrated by $t_{obs} = t_{IV^*}$. Although this scenario leads to a small value of $1 - p_A$, exclusion is questionable because $t_{obs} = t_{IV^*}$ is nearly as unlikely under H_0. In case IV*, there are fewer counts than expected even without the Higgs boson. This situation may arise if there is a significant downward fluctuation in the background, and this case is considered in detail in Section 4.5. The most likely outcome is a value of $t_{obs} = t_{II}$ that is not extreme under either hypothesis. The real problem in **Figure 2*a*** is the insensitivity of the distribution of T to the choice between H_0 and H_A. The sensitivity, μ_{sen}, is the minimum alternative value of μ that, in some measure, avoids this problem. Thus, ideally $\mu_{sen} < 1$.

Lyons (2013) introduces a plot designed to illustrate the sensitivity of a test when comparing two simple hypotheses. Because larger values of T are more significant under H_0, both p_0 and p_A are monotone decreasing functions of t_{obs}; thus, as t_{obs} increases, p_0 decreases and $1 - p_A$ increases. If we vary t_{obs} over the support of T, the ordered pair $\{p_0, 1 - p_A\}$ forms a curve in the unit square. **Figure 3** shows how the shape of the curve describes the sensitivity of the test. As p_0 increases from zero to one, $1 - p_A$ falls more quickly for a more sensitive test, passing through the "neither model consistent with data" region near the origin for a test as sensitive as the one in **Figure 2*b***. (The test in **Figure 2*c*** is too sensitive to meaningfully plot in **Figure 3**.) At the other extreme, the test in **Figure 2*a*** passes quickly into the "both models consistent with data" region.

4.5. Avoiding Exclusion Under an Insensitive Test (Case IV*)

High-energy physicists are concerned that unwarranted exclusion may occur if the distribution of T is similar under $H_0 : \mu = 0$ and $H_A : \mu = 1$, that is, if the test is relatively insensitive. As illustrated by $t_{obs} = t_{IV^*}$ in **Figure 2*a***, an extreme value of $1 - p_A$ may be attributed to a downward fluctuation in the background rather than to inadequacy of H_A relative to H_0. Unlikely background counts are less of a concern for discovery than for exclusion because a much more stringent criterion is used for detection: $\alpha_0 \ll \alpha_A$. We can avoid unwarranted exclusion by refraining from excluding $\mu = 1$ even in $1 - p_A \leq \alpha_A$ if $1 - p_0$ is also small. In particular, Read (2000) proposes computing

$$CL_S = \frac{1 - p_A}{1 - p_0} \qquad 4.$$

and excluding μ_A only if $CL_S \leq \alpha_A$ (also see Junk 1999 and Read 2002). This criterion is more conservative in terms of exclusion because $CL_S \geq 1 - p_A$ (**Figure 3**). **Figure 3*b*** uses the CL_S criterion, which reduces the size of the yellow exclusion region and (nearly) eliminates the purple-shaded region that corresponds to unwarranted exclusions (case IV*). A more direct approach would be to exclude $\mu = 1$ only if $1 - p_A \leq \alpha_A$ and $1 - p_0$ is larger than some threshold. The drawback of the direct approach is that it requires an additional explicit threshold. Of course, CL_S defines this threshold implicitly.

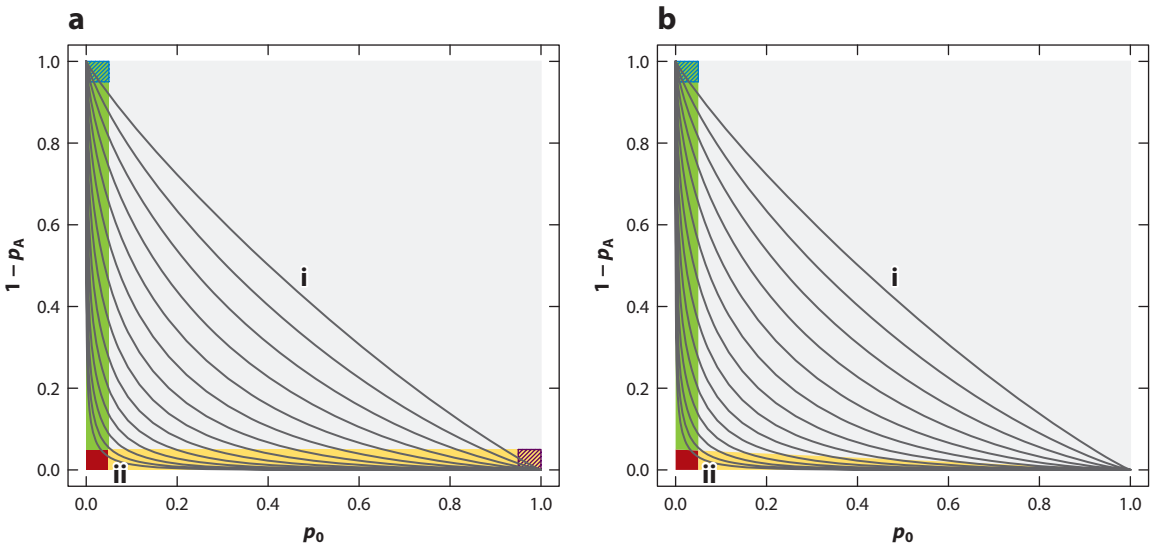

Figure 3

A comparison of the sensitivity of various tests. The curves marked *i* in each panel plot $\{p_0, 1 - p_A\}$ for the test described in **Figure 2*a***. The other curves correspond to cases in which the distributions of the statistic are increasingly separated under H_0 and H_A, culminating in the curve in each panel marked *ii*, which corresponds to the case illustrated in **Figure 2*b***. The colored regions correspond to the four cases listed in **Table 1**, plotted with $\alpha_0 = \alpha_A = 0.05$; panel *b* uses the CL_S criterion described in Section 4.5. Case I, the detection region, is colored green; case II, in which both models are consistent with the data, is gray; case III, in which neither model is consistent with the data, is red; and case IV, the exclusion region, is yellow. Curve *ii* enters the troublesome red region. Special cases I* and IV* are shaded in blue and purple, respectively. They also indicate that the data are inconsistent with both models in that there are either more (case I*, shaded blue) or fewer (case IV*, shaded purple) counts than can be explained under either model. The latter case may be explained by a downward fluctuation in the background and does not justify exclusion. As shown in panel *b*, this region is (nearly) eliminated by the CL_S criterion.

The primary use of CL_S is in the computation of upper limits by inverting the exclusion test. Consider the test that compares $H_0 : \mu = 0$ with $H_A : \mu = \mu_0$, where we are interested in whether or not we can exclude μ_0. Applying the CL_S criterion, we exclude μ_0 if

$$\frac{\Pr(T < t_{obs}|\mu = \mu_0)}{\Pr(T < t_{obs}|\mu = 0)} \leq \alpha_A, \qquad 5.$$

where T is the likelihood-ratio test statistic based on Equation 3. That is, we exclude values of μ_0 that make extreme values of T significantly less likely than they are under H_0. (Recall that for exclusion the roles of the hypotheses are interchanged and smaller values of T are more extreme.) **Figure 4** depicts the CL_S upper limits for ATLAS.

Another method that aims to reduce unwarranted exclusion is the power constrained limit, which reports the larger of μ_{UL} and μ_{sen} (Cowan et al. 2011b). This method precludes the exclusion of any value of μ that is lower than the sensitivity of the test. In the interest of full disclosure, van Dyk (2011) proposed that both μ_{UL} and μ_{sen} be reported.

4.6. The 5σ Detection Threshold

Particle physicists generally transform *p*-values into the number of standard deviations a test statistics would be from zero if its null distribution were standard normal under a one-sided test. For example, a *p*-value of 0.025 would be referred to as a 1.96σ result. The 2012 detection

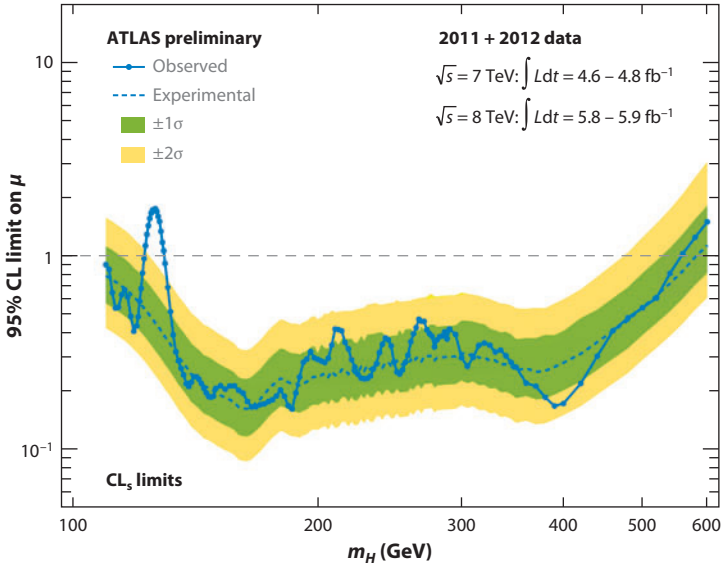

Figure 4

CL_S upper limits as a function of the Higgs mass, m_H, for ATLAS. The CL_S upper limits are compared with experimental sensitivity that is quantified by the distribution of μ_{UL} under H_0. The upper limits depart from their null distribution and are greater than one near $m_H \approx 125$ GeV. The plot is based on preliminary data but is similar to published results (figure 10 of ATLAS Collab. 2012c). Copyright CERN.

of "a previously unknown boson" with a *p*-values less than one in three million (quoted in Section 2) corresponded, more precisely, to 6.0σ and 5.0σ detections for ATLAS and CMS, respectively (ATLAS Collab. 2012c, CMS Collab. 2012a).

At face value, these values seem to be highly significant results, and editors of particle physics journals generally require significance levels of 5σ to claim a detection. This requirement is in part a response to high-profile false discoveries that predate even the pentaquark snafu described in the sidebar. These nominal significance levels, however, are computed under several assumptions that tend to attenuate the actual significance. For example, they are computed assuming the Higgs mass is known, when in fact it is not. Particle physicists refer to these as local significance levels, and these levels ignore the multiple testing that is conducted at multiple masses in the analysis. It is common, in fact, to plot the local significance as a function of the Higgs mass, m_H, as in **Figure 5** (also see figure 7 of Della Negra et al. 2012).

Particle physicists are well aware of the effect of multiple testing but lack a consensus on how to best handle it largely because it is not clear what constitutes a test. Because previous experiments excluded Higgs masses outside the range from 114 to 130 GeV and because ATLAS and CMS focused their searches on this mass range, it may seem that the scope of the multiple-testing problem is well defined. But suppose that in a formal search between 114 and 130 GeV, the experimenters noticed a strong Higgs boson–like signature at 113 GeV. There is no doubt that if the signal were strong enough a detection would be declared. Thus, correcting for multiple testing between 114 and 130 GeV is not sufficient. A troublesome feature of frequency-based methods is that they require a fully specified protocol of what one would do if one had different data than what was actually observed, and strict compliance to this protocol. To avoid these issues, physicists typically report local significances, but with the stringent 5σ threshold for detection.

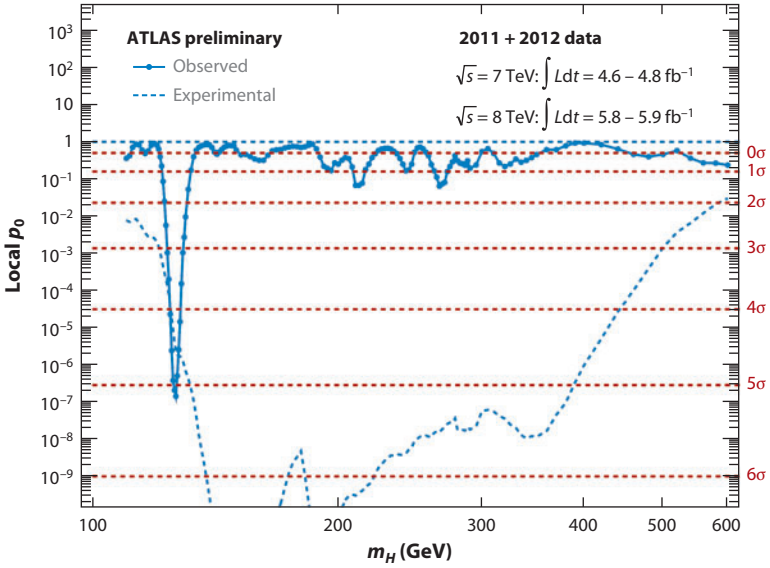

Figure 5

Local significance as a function of the Higgs mass, m_{H}. Values are based on preliminary results from ATLAS and are compared with the expected local significance under the $\mu = 1$ hypothesis. Notice the peak in significance at $m_{\mathrm{H}} \approx 125$ GeV. Published ATLAS and CMS results appear in figure 7 of Della Negra et al. (2012). Copyright CERN.

In its review of the 2012 CMS and ATLAS discoveries, the journal *Science* included a glossary that described the 5σ threshold. It noted (Am. Assoc. Adv. Sci. 2012, p. 1559) that "in particle physics, this criterion has become a convention to claim discovery but should not be interpreted literally." Indeed, the motivation of the 5σ detection threshold is not to keep the false detection rate below 1 in 3.4 million tests. Rather it is an attempt to account for concerns associated with multiple testing, calibration, and/or systematic errors, and statistical error rates that are not well calibrated due to general model misspecification (Cox 2011; Lyons 2008, 2012). Unfortunately, reducing α_0 does not adequately address these concerns. Model misspecification and systematic errors probably induce both increased bias and variance. Unfortunately, a more stringent threshold for detection does not address bias and is a uncalibrated response to variance. Of course, statistical practice always involves compromises of this sort. The irony here is that the strict adherence to frequency-based procedures that prevents postdata model checking does not, in the final analysis, deliver a well-calibrated frequentist procedure.

5. DETECTION AND EXCLUSION WITH UNKNOWN MASS

5.1. Mass-by-Mass Analysis

Generalizing from the single-bin analysis described in the model in Equation 2, the observed counts in a Higgs boson detection experiment can be written as N_{msc}, where c indexes the decay channels, s indexes the categories (strata) within each channel, and m indexes the recorded invariant masses of events within each channel-category pair. Recall that the number of categories varies among the several channels and that invariant masses are binned.

The search for the Higgs boson is conducted separately for each potential Higgs mass on a fine grid of values of m_H. The primary reason is a concern that an integrated search may overlook potential Higgs masses to which the experiment is relatively insensitive. A mass-by-mass analysis allows the sensitivity to be computed separately for each mass on the grid. A secondary advantage of this strategy is that the null distribution of the likelihood-ratio test statistic is simpler to derive for fixed m_H. The standard asymptotic results do not apply, because H_0 is on the boundary of the parameter space. In the spirit of Chernoff (1954), Cowan et al. (2011a) derive the null distribution for several variants of the likelihood-ratio test, including Equation 3.

Even when searching for a Higgs boson with a given mass, m_H, there is a range of possible recorded Higgs masses because the invariant mass of each candidate Higgs boson is computed using Einstein's equation (Equation 1). This computation involves a certain amount of stochastic error that blurs the recorded invariant masses of both Higgs and background events. There is also a small intrinsic variance in m_H, as described by Heisenberg's uncertainty principle. Although these two effects result in a relatively narrow mass spread (**Figure 1**), we must look at events counts over a wide range of masses to fit the background, as described in Section 3. Thus, all of the counts, N_{msc}, are used for each mass-specific search.

Returning to the model in Equation 2, both the background and source terms must be estimated using the multibin data. Doing so introduces both numerous nuisance parameters and sensitivity of results to the choice of parameterization. The model specification for the background is discussed in Section 3. Formally, we let $\beta_{sc}(\phi_{sc}, m)$ be the expected background count in invariant mass bin m of category (stratum) s of channel c, which depends on the unknown parameter, ϕ_{sc}, and the mass, m, associated with the bin. The subscripts on β and ϕ emphasize that the choice of both the background models and their parameters is different for each category-channel pair. In particular, the dimensions of the ϕ_{sc} are not all the same. However, they are fixed in advance along with the functional forms of the β_{sc}.

The source models describe the expected Higgs count under the Standard Model in each mass bin. This count depends weakly on physical parameters that are not precisely determined and, more importantly, on instrumental effects such as the inexact assignment of events to mass bins. In addition, the ATLAS Collaboration includes a term that allows for a spurious signal that in effect reduces the significance of any potential discovery. External data are available for some of the source-model parameters, and evidence-based prior distributions are used to incorporate this information. We write the source model as $\kappa_{sc}(\phi_{sc}, m)$, where ϕ_{sc} is a category-channel-specific nuisance parameter. Finally, the multibin data model used in the search for a Higgs boson of a particular mass can be written as

$$N_{msc} \sim \text{Poisson}[\beta_{sc}(\theta_{sc}, m) + \kappa_{sc}(\phi_{sc}, m)\mu]. \quad 6.$$

Taken together, the parameters (θ_{sc}, ϕ_{sc}) for each category-channel pair form a large dimensional nuisance parameter that complicates both the detection problem and the setting of upper limits. Physicists have considered a wide range of methods to deal with nuisance parameters and have come across many of the same roadblocks known to statisticians for both limits (Cousins & Highland 1992, Heinrich et al. 2004, Rolke et al. 2005) and testing (e.g., Demortier 2008). In the search for a Higgs boson, a pragmatic approach is taken. The likelihood-ratio test statistic is used both in setting limits and in detection (Cowan et al. 2011a). Because it is asymptotically ancillary, its distribution should be relatively free of nuisance parameters for large data sets.

5.2. The Look-Elsewhere Effect

The strategy of conducting separate detection tests on a fine grid of m_H leads to multiple dependent tests that complicate the interpretation of the multiple p-values. (The dependency of the tests can

be observed in how p_0 varies smoothly with m_H in **Figure 5**.) Physicists refer to this multiple-testing problem as the look-elsewhere effect (Demortier 2008, Lyons 2008) and the m_H-specific p-values as the local p-values. The dependency of the m_H-specific tests means that simple corrections such as the Bonferroni correction are overly conservative. Because m_H is unidentified under H_0, the asymptotic distribution of the global likelihood-ratio test statistics is unknown. In principle, its null distribution could be simulated, perhaps with nuisance parameters fixed at their fitted values, but doing so is infeasible given the stringent detection criteria used in the Higgs boson search. However, it has been attempted by, for instance, the CMS Collaboration (2007).

Formally, let $T(m_H)$ be the value of the likelihood-ratio test statistic for an m_H-specific test and assume that its null distribution is χ^2_s. Davies (1987) shows that

$$\Pr\left(\max_{m_H} T(m_H) > c\right) \leq \Pr(\chi^2_s > c) + E(M(c) \mid H_0), \qquad 7.$$

where $M(c)$ is the number of upcrossings, that is, the number of times $T(m_H)$ increases from below to above c as m_H increases. Because such stringent detection criteria are used in the Higgs boson search, we are interested in large c, where the bound in Equation 7 becomes exact. Although a direct Monte Carlo evaluation of the expected number of upcrossings is infeasible for such large c, Gross & Vitells (2010) propose an elegant solution by noting that

$$E(M(c)|H_0) = E(M(c_0) \mid H_0)\left(\frac{c}{c_0}\right)^{(s-1)/2} \exp\left(-\frac{(c-c_0)}{2}\right), \qquad 8.$$

where $c_0 \ll c$ and $E(M(c_0) \mid H_0)$ can be computed efficiently via Monte Carlo (also see Vitells 2011). Evaluating Equation 7 at $c = \max_{m_H} t_{\text{obs}}(m_H)$ yields the global p-value. Despite the ease with which Gross & Vitells's solution allows local p-values to be converted into global p-values, detection results still are typically quoted as local p-values (e.g., Della Negra et al. 2012). The assumption is that the stringent detection criterion guards against false detections, even when the look-elsewhere effect and other systematic errors in the analysis are ignored. The local significances of 6.0σ and 5.0σ for the ATLAS and CMS Higgs boson detections described above correspond to global significances of 5.1σ and 4.6σ, respectively (ATLAS Collab. 2012c, CMS Collab. 2012a).

5.3. Estimating the Higgs Mass

Once a Higgs boson is discovered, its mass is estimated by incorporating m_H into Equation 6 and conducting a unified analysis, rather than a mass-by-mass analysis. In particular, the source model, $\kappa(\phi_{sc}, m)$ in Equation 6, depends on the posited Higgs mass. Accounting for this dependence explicitly, the full model is

$$N_{msc} \sim \text{Poisson}[\beta_{sc}(\theta_{sc}, m) + \kappa_{sc}(\phi_{sc}, m, m_H)\mu]. \qquad 9.$$

Fitting Equation 9 leads to the reported estimates and statistical errors for m_H.

6. SUMMARY

The search for the Higgs boson illustrates a generalization of the standard hypothesis-testing framework. The goal is to determine which—if either—of the two hypotheses is consistent with the data, rather than to simply accept or reject the null model. This goal is achieved while maintaining an a priori preference for the null model: Much stronger evidence is required to reject the null hypothesis than to reject the alternative. To a statistician, a decision theoretic framework may seem better suited to this choice (van Dyk 2011) and more able to avoid seemingly ad hoc decision

criterion such as CL_S. Employing a Bayesian framework, however, would avoid the well-known bias of the *p*-value toward false discovery with a sharp null, but would also introduce the challenge of prior specification (Berger & Delampady 1987). Particle physicists (or their publishers) have instead chosen to stick with tail probabilities, with the hope that they can guard against model misspecification, systematic errors, and other possible biases by using an ultraconservative detection threshold. Although statisticians—and many particle physicists—find this solution unsatisfactory, hoping for a more principled approach may be unrealistic. Any methodological shift away from tail probabilities would require a corresponding cultural shift that is unlikely in a scientific endeavor as costly and high profile as the search for a Higgs boson.

Although this article focuses on the statistical challenges involved in the search for the Higgs boson, it is important to remember that the compromises and ad hoc solutions involved in the search are not unique. All real data analyses involve trade-offs. The Higgs boson search is simply a high-profile problem, and the researchers have insisted on much higher standards for their analyses than are typically encountered in practice. Overall, they should be commended for the caliber of their methods, and their discovery should be recognized as an excellent example of the dynamic interplay between modern statistical methods and a complex real-world applied problem.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I thank Louis Lyons, Paul Dauncey, Nicholas Wardle, and the participants in the CERN 2013 miniworkshop on "statistical issues involved in the search for the Higgs boson" for many helpful and educational conversations. Two anonymous referees also provided a number of constructive and helpful suggestions. My work was supported in part by the National Science Foundation (DMS-12-08791), the Royal Society (Wolfson Merit Award), and the European Commission (Marie Curie Career Integration Grant).

LITERATURE CITED

Anderson PW. 1992. The Reverend Thomas Bayes, needles in haystacks, and the fifth force. *Phys. Today* 45:9–11

Am. Assoc. Adv. Sci. 2012. The Higgs boson. *Science* 338:1558–59

ATLAS Collab. 2008. *Expected Performance of the ATLAS Experiment: Detector, Trigger and Physics. CERN-OPEN-2008-020*. Geneva, Switz.: CERN

ATLAS Collab. 2012a. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* 716:1–29

ATLAS Collab. 2012b. *Observation of an Excess of Events in the Search for the Standard Model Higgs Boson in the $\gamma\gamma$ Channel with the ATLAS Detector. ATLAS-CONF-2012-091*. Geneva, Switz.: CERN

ATLAS Collab. 2012c. A particle consistent with the Higgs boson observed with the ATLAS detector at the Large Hadron Collider. *Science* 338:1576–82

Berger J. 2008. A comparison of testing methodologies. See Lyons et al. 2008, pp. 8–19

Berger JO, Delampady M. 1987. Testing precise hypotheses (with discussion). *Stat. Sci.* 2:317–52

Campagnari C, Franklin M. 1997. The discovery of the top quark. *Rev. Mod. Phys.* 69:137–211

Chernoff H. 1954. On the distribution of the likelihood ratio. *Ann. Math. Stat.* 25:573–78

CLAS Collab. 2003. Observation of an exotic $S = +1$ baryon in exclusive photoproduction from the deuteron. *Phys. Rev. Lett.* 91:252001

CLAS Collab. 2006. Search for the θ^+ pentaquark in the reaction $\gamma d \to p k^- k^+ n$. *Phys. Rev. Lett*. 96:212001

CLAS Collab. 2008. Bayesian analysis of pentaquark signals from CLAS data. *Phys. Rev. Lett.* 100:052001

CMS Collab. 2007. CMS physics technical design report. Volume II: Physics performance. *J. Phys. G* 34:995–1579

CMS Collab. 2012a. A new boson with a mass of 125 GeV observed with the CMS experiment at the Large Hadron Collider. *Science* 338:1569–75

CMS Collab. 2012b. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B* 716:30–61

Cousins RD. 1995. Why isn't every physicist a Bayesian? *Am. J. Phys.* 63:398–410

Cousins RD. 2008. Comment on "Bayesian analysis of pentaquark signals from CLAS data." *Phys. Rev. Lett.* 101:029101

Cousins RD, Highland VL. 1992. Incorporating systematic uncertainties into an upper limit. *Nucl. Instrum. Methods A* 320:331–35

Cowan G, Cranmer K, Gross E, Vitells O. 2011a. Asymptotic formulae for likelihood-based tests for new physics. *Eur. Phys. J. C* 71:1554

Cowan G, Cranmer K, Gross E, Vitells O. 2011b. Power-constrained limits. arXiv:1105.3166 [physics.data-an]

Cox DR. 2011. Discovery: a statistical perspective. See Prosper & Lyons 2011, pp. 12–16

Davies RB. 1987. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74:33–43

DelRosso A. 2012. Higgs: the beginning of the exploration. *CERN Wkly. Bull.* 47–48/2012

Della Negra M, Jenni P, Virdee TS. 2012. Journey in the search for the Higgs boson: the ATLAS and CMS experiments at the Large Hadron Collider. *Science* 338:1560–68

Demortier L. 2008. *P* values and nuisance parameters. See Lyons et al. 2008, pp. 23–33

Ellis J, Gaillard MK, Nanopoulos DV. 2012. A historical profile of the Higgs boson. arXiv:1201.6045v1 [hep-ph]

Feldman GJ, Cousins RD. 1998. Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D* 57:3873–89

Friedman JH. 2005. Separating signal from background using ensembles of rules. See Lyons & Unel 2005, p. 10

Gross E, Vitells O. 2010. Trial factors for the look elsewhere effect in high energy physics. *Eur. Phys. J. C* 70:525–30

Heinrich J, Blocker C, Conway J, Demortier L, Lyons L, et al. 2004 . *Interval Estimation in the Presence of Nuisance Parameters. 1. Bayesian Approach. CDF note 7117.* Batavia, IL: Fermilab. 30 pp.

Higgs PW. 1964. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.* 13:508–9

Junk T. 1999. Confidence level computation for combining searches with small statistics. *Nucl. Instrum. Methods A* 434:435–43

Kashyap VL, van Dyk DA, Connors A, Freeman PE, Siemiginowska A, et al. 2010. On computing upper limits to source intensities. *Astrophys. J.* 719:900–14

Lederman LM. 1993. *The God Particle: If the Universe Is the Answer, What Is the Question?* Boston: Houghton Mifflin

Lyons L. 2008. Open statistical issues in particle physics. *Ann. Appl. Stat.* 2:887–915

Lyons L. 2012. Discovery or fluke: statistics in particle physics. *Phys. Today* 65:45–51

Lyons L. 2013. Bayes and frequentism: a particle physicist's perspective. *Contemp. Phys.* 54:1–16

Lyons L, Mount R, Reitmeyer R, eds. 2004. *Proceedings of the Conference on Statistical Problems in Particle Physics, Astrophysics, and Cosmology* (*PHYSTAT 2003*), *Stanford Linear Accelerator Center, Stanford, California September 8–11, 2003*. Menlo Park, CA: SLAC Tech. Publ.

Lyons L, Prosper HB, de Roeck A, eds. 2008. *Proceedings of the PHYSTAT 2007 Workshop on Statistical Issues for LHC Physics, CERN, Geneva, Switzerland, 27–29 June 2007. CERN-2008-001*. Geneva, Switz.: CERN

Lyons L, Unel MK, eds. 2005. *Proceedings of the PHYSTAT 2005 Workshop on Statistical Problems in Particle Physics, Astrophysics and Cosmology, Oxford, UK, 12–15 September 2005*. London: Imperial Coll. Press

Mandelkern M. 2002. Setting confidence intervals for bounded parameters (with discussion). *Stat. Sci.* 17:149–72

Park T, van Dyk DA, Siemiginowska A. 2008. Searching for narrow emission lines in X-ray spectra: computation and methods. *Astrophys. J.* 688:807–25

Part. Data Group. 2006. Review of particle physics. *J. Phys. G* 33:1–1232

Part. Data Group. 2008. Review of particle physics. *Phys. Lett. B* 667:1–1340

Prosper HB, Lyons L, eds. 2011. *Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva, Switzerland, 17–20 January 2011. CERN-2011-006*. Geneva, Switz.: CERN

Protassov R, van Dyk DA, Connors A, Kashyap V, Siemiginowska A. 2002. Statistics: Handle with care—detecting multiple model components with the likelihood ratio test. *Astrophys. J.* 571:545–59

Punzi G. 2004. Sensitivity of searches for new signals and its optimization. See Lyons et al. 2004, pp. 79–83

Read AL. 2000. Modified frequentist analysis of search results (the CL_s method). **http://cds.cern.ch/record/451614**

Read AL. 2002. Presentation of search results: the CL_S technique. *J. Phys. G* 10:2693–704

Roe BP, Woodroofe MB. 1999. Improved probability method for estimating signal in the presence of background. *Phys. Rev. D* 60:053009

Roe BP, Yang H-J, Zhu J, Liu Y, Stancu I, McGregor G. 2005. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nucl. Instrum. Methods A* 543:577–84

Rolke WA, Lopez AM, Conrad J. 2005. Limits and confidence intervals in the presence of nuisance parameters. *Nucl. Instrum. Methods A* 551:493–503

Salam A. 1968. Weak and electromagnetic interactions. In *Proceedings of the Eighth Nobel Symposium, Lerum, Sweden, 19–25 May 1968*, ed. N Svartholm, pp. 367–77. Stockholm: Almqvist & Wiksell

van Dyk DA. 2011. Setting limits, computing intervals, and detection. See Prosper & Lyons 2011, pp. 149–57

Vitells O. 2011. Estimating the "look elsewhere effect" when searching for a signal. See Prosper & Lyons 2011, pp. 183–89

Weinberg S. 1967. A model of leptons. *Phys. Rev. Lett.* 19:1264–66

Weniger C. 2012. A tentative γ-ray line from dark matter annihilation at the Fermi Large Area Telescope. *J. Cosmol. Astropart. Phys.* 1208:007