Testing Hypotheses in Particle Physics: 
Plots of $p_0$ Versus $p_1$

Luc Demortier†, Louis Lyons‡

†Laboratory of Experimental High Energy Physics
The Rockefeller University, New York, NY 10065, USA
‡Blackett Laboratory
Imperial College, London SW7 2BW, UK

August 27, 2014

Abstract

For situations where we are trying to decide which of two hypotheses $H_0$ and $H_1$ provides a better description of some data, we discuss the usefulness of plots of $p_0$ versus $p_1$, where $p_i$ is the $p$-value for testing $H_i$. They provide an interesting way of understanding the difference between the standard way of excluding $H_1$ and the $CL_s$ approach; the Punzi definition of sensitivity; the relationship between $p$-values and likelihood ratios; and the probability of observing misleading evidence. They also help illustrate the Law of the Iterated Logarithm and the Jeffreys-Lindley paradox.

1 Introduction

Very often in particle physics we try to see whether some data are consistent with the standard model (SM) with the currently known particles (call this hypothesis $H_0$), or whether it favors a more or less specific form of new physics in addition to the SM background ($H_1$). This could be, for example, a particular form of leptoquark with a well-defined mass; or with a mass in some range (e.g. 50 to 1000 GeV). In the first case there are no free parameters and $H_1$ is described as being ‘simple’, while in the latter case, because of the unspecified leptoquark mass, $H_1$ is ‘composite’.

If the only free parameter in the alternative hypothesis $H_1$ is the mass of some new particle, we can test each mass in $H_1$ separately against $H_0$, in which case we are comparing two simple hypotheses. However, the ensemble of different possible masses in the overall procedure (known as a ‘raster scan’ [1]) makes $H_1$ composite. Insight into this type of situation is facilitated by two-dimensional ‘$p$-value plots’, where the significance of possible observations under the null hypothesis is plotted against their significance under various values of the free parameter in the alternative hypothesis [2]. The purpose of this article is to use such plots to explore various aspects of hypothesis testing in particle physics1.

---

1In this article we concentrate on hypothesis testing procedures pertaining to discovery claims in search experiments (not necessarily in particle physics). We do not consider other uses of hypothesis testing, such as in particle physics event selection for instance. The desiderata are slightly different.
We begin in section 2 by recapitulating the types of hypothesis testing familiar from
the statistics literature, and contrasting these with the practice in particle physics. Sec-

tion 3 introduces p-value plots and uses them to discuss the CLs criterion, upper limits,
fixed-hypothesis contours, and the Punzi definition of sensitivity. The probabilities for ob-
servations to fall into various regions of a p-value plot are derived in section 4, together with
the error rates and power of a particle physics test. Likelihood ratios form the subject of
section 5, where they are compared to p-values and used to plot contours and to compute
probabilities of misleading evidence. Two famous p-value puzzles are described in section 6.
Section 7 contains remarks on the effect of nuisance parameters, and our conclusions and
recommendations appear in section 8. An appendix provides technical details about the
relationship between CLs and Bayesian upper limits.

2 Types and outcomes of hypothesis testing

When using observed data to test one or more hypotheses, the first step is to design a test
statistic \( T \) that summarizes the relevant properties of the data. The observed value \( t \) of \( T \) is
then referred to its probability distribution under each specified hypothesis in order to assess
evidence. The form of the test statistic depends on the type of test one is interested in.

Comparisons of data with a single hypothesis are performed via ‘goodness of fit’ tests. An example of this is the \( \chi^2 \) test, which generally requires the data to be binned, and where
\( T \) is equal to the sum of the squares of the numbers of standard deviations between observed
and expected bin contents. Another well-known technique, which does not require binning,
is the Kolmogorov-Smirnov test, where \( T \) is constructed from the expected and observed
cumulative distributions of the data. There are many other techniques \([3, 4]\). The outcome
of a goodness-of-fit test is either ‘Reject’ or ‘Fail to reject’ the hypothesis of interest.

Comparison of the data with more than one hypothesis in order to decide which is
favored is known as ‘hypothesis testing’. If there are just two simple hypotheses \( H_0 \) and \( H_1 \),
the appropriate framework is Neyman-Pearson hypothesis testing. The optimal test statistic
\( T \) in this case is the likelihood ratio for the two hypotheses, or a one-to-one function of it \(^2\).
The outcome of a Neyman-Pearson test is either ‘Reject \( H_0 \) and accept \( H_1 \),’ or ‘Accept \( H_0 \)
and reject \( H_1 \).’

In particle physics it often happens that we need to consider additional possible outcomes
of a test. In the leptoquark example, an observed signal could be due to something entirely
different from a leptoquark: some new physics that we did not anticipate, or a systematic
bias that we did not model. Hence we may need to reject both \( H_0 \) and \( H_1 \) in favor of a third,
unspecified hypothesis. On the other hand it may also happen that the data sample does not
allow us to reject either \( H_0 \) or \( H_1 \) \([5]\). This leads to the formulation of a ‘double test’ of two
hypotheses \( H_0 \) and \( H_1 \), which are independently tested, resulting in four possible outcomes:

1. Fail to reject \( H_0 \), and reject \( H_1 \). This is referred to as ‘\( H_1 \) excluded,’ and in a frequentist
   approach the rejection of \( H_1 \) is valid at some level of confidence, typically 95%.

2. Fail to reject \( H_0 \) and fail to reject \( H_1 \) (‘No decision’).

\(^2\)In the case of a counting experiment, the number of observed counts \( n \) is typically a one-to-one function of
the likelihood ratio for the ‘signal+background’ and ‘background-only’ hypotheses (\( H_1 \) and \( H_0 \) respectively).
3. Reject $H_0$, and fail to reject $H_1$. This corresponds to ‘Discovery of $H_1$.’ In a frequentist approach the rejection of $H_0$ is valid at some confidence level, which in particle physics is usually much higher than the confidence level used for excluding $H_1$. Typically the significance level, defined as one minus the confidence level, is set at $2.87 \times 10^{-7}$ for rejecting $H_0$. This is the area under a Gaussian tail, starting five standard deviations away from the mean.

4. Reject both $H_0$ and $H_1$.

Often a likelihood ratio is used as the test statistic $T$ for a double test.

For given $H_0$ and for fixed values of the parameters in $H_1$, we can plot the probability density functions (pdf’s) of $T$, assuming (a) that hypothesis $H_0$ is true, or (b) that $H_1$ is true. Three possible situations are shown in figure 1. In (a), the two hypotheses are hard to distinguish as the pdf’s lie almost on top of each other; this could happen if $H_1$ involved a new particle that was only very weakly produced. In (b), the pdf’s still overlap to some extent, but distinguishing between the two hypotheses may be possible for some data sets. Finally (c) shows a situation where it is relatively easy to choose between the hypotheses.

3 p-Values

The degree to which the data are unexpected for a given hypothesis can be quantified via the $p$-value. This is the fractional area in the tail of the relevant pdf, with a value of $t$ at least as extreme as that in the data. In tests involving two hypotheses, it is conventional to use the one-sided tail in the direction of the other hypothesis. For the examples shown in figure 1, this corresponds to $p_0$ being the right-hand tail of $H_0$ and $p_1$ the left-hand tail of $H_1$. In the extreme case where $H_0$ and $H_1$ coincide (and where $t$ is continuous rather than discrete), $p_0 + p_1 = 1$.

3.1 Regions in the $(p_0, p_1)$ plane

Figure 2 contains a plot of $p_0$ versus $p_1$, with the regions for which the double test either rejects $H_0$ or fails to reject it; these depend solely on $p_0$. In the diagram, the critical value $\alpha_0$ for $p_0$ is shown at 0.05; this value is chosen here for clear visibility on the plot, rather than as a realistic choice.

In particle physics, when we fail to reject $H_0$, we want to see further whether we can exclude $H_1$. Although not as exciting as discovery, exclusion can be useful from a theoretical point of view and also for the purpose of planning the next measurement. The most famous example is the Michelson-Morley experiment, which excluded any significant velocity of the earth with respect to the aether and led to the demise of the aether theory. In figure 2, the region $p_1 \leq \alpha_1$ is used for excluding $H_1$. The critical value $\alpha_1$ is usually chosen to be larger than the $p_0$ cut-off $\alpha_0$; 0.05 is a typical value. In the figure $\alpha_1$ is shown at 0.10.

---

3In this paper we do not consider problems in which it is desired to reject $H_0$ when the data statistic $t$ falls in the extreme left-tail of the $H_0$ pdf (see figure 1), or to reject $H_1$ when $t$ is very large. Our $p$-values are one-sided and would therefore be close to unity in these cases.
If \( p_0 \) and \( p_1 \) fall in the large rectangle at the top right of the plot (\( p_0 > \alpha_0 \) and \( p_1 > \alpha_1 \)), we claim neither discovery of \( H_1 \) nor its exclusion: this is the no-decision region. The small rectangle near the origin corresponds to both \( p \)-values being below their cut-offs, and the data are unlikely under either hypothesis. It could correspond to the new physics occurring, but at a lower than expected rate.

### 3.2 The CL\(_s\) criterion

An alternative approach for exclusion of \( H_1 \) is the CL\(_s\) criterion [6]. Because exclusion levels are chosen to have modest values (say 95%), there is substantial probability (5%) that \( H_1 \) will be excluded even when the experiment has little sensitivity for distinguishing \( H_1 \) from \( H_0 \) (the situation shown in figure 1(a)). Although professional statisticians are not worried about this, in particle physics it is regarded as unsatisfactory. To protect against this, instead of rejecting \( H_1 \) on the basis of \( p_1 \) being small, a cut is made on

\[
CL_s \equiv \frac{p_1}{1 - p_0},
\]

i.e. on the ratio of the left-hand tails of the \( H_0 \) and \( H_1 \) pdf’s. Thus if the pdf’s are almost indistinguishable, the ratio will be close to unity, and \( H_1 \) will not be excluded. In figure 2, the region below the dashed line referred to as ‘CL\(_s\)’ shows where \( H_1 \) would be excluded. This is to be compared to the larger region below the horizontal line for the more conventional exclusion based on \( p_1 \) alone. The CL\(_s\) approach can thus be regarded as a conservative modification of the exact frequentist method; conservatism is the price to pay for the protection CL\(_s\) provides against exclusion when there is little or no sensitivity to \( H_1 \).

### 3.3 Upper limits

As pointed out in the introduction, the pdf of \( H_1 \) often contains one or more parameters of interest whose values are not specified (e.g. the mass of a new particle, the cross section of a new process, etc.). It is then useful to determine the subset of \( H_1 \) parameter space where, with significance threshold \( \alpha_1 \), each parameter value is excluded by the observations. In the frequentist paradigm, the complement of this subset is a CL = 1 − \( \alpha_1 \) confidence region.

For a simple and common example consider the case where the pdf of the data depends on the cross section \( \mu \) of a new physics process: then \( \mu > 0 \) if the process is present in the data (\( H_1 \) true), and \( \mu = 0 \) otherwise (\( H_0 \) true). Suppose that the test statistic \( T \) is stochastically increasing with \( \mu \), meaning that for fixed \( T \), increasing \( \mu \) reduces the \( p \)-value \( p_1 \). Then the set of \( \mu \) values that cannot be excluded by the observations has an upper limit, and that upper limit has confidence level 1 − \( \alpha_1 \).

If instead of rejecting \( H_1 \) with the standard frequentist criterion \( p_1 \leq \alpha_1 \), we use the CL\(_s\) criterion \( CL_s \leq \alpha_1 \), the above procedure yields a CL\(_s\) upper limit for \( \mu \), which is higher (i.e. weaker) than the standard frequentist upper limit.

In the previous example suppose that, instead of a cross section, \( \mu \) is a location parameter for the test statistic \( t \). More precisely, suppose that the pdf of \( t \) is of the form \( f(t - \mu) \), with \( f \) a continuous distribution. Then it can be shown that the upper limit using CL\(_s\) at the 1 − \( \alpha_1 \) level coincides exactly with the credibility 1 − \( \alpha_1 \) Bayesian upper limit obtained by
assuming a uniform prior for $\mu$ under $H_1$ (i.e., a prior that is a non-zero constant for $\mu > 0$, and zero elsewhere). This result extends to the discrete case where $t$ is a Poisson-distributed event count with mean $\mu$ (see Appendix A).

3.4 Fixed-hypothesis contours in the $(p_0, p_1)$ plane.

If we keep the hypotheses $H_0$ and $H_1$ fixed, but vary the observed data statistic $t$, the point $(p_{0,\text{obs}}, p_{1,\text{obs}})$ will trace a contour in the $(p_0, p_1)$ plane. In general this contour depends on the particular characteristics of each hypothesis, but useful simplifications may occur when the pdf of the test statistic is translation-invariant or enjoys other symmetries. Here we give four examples based on the pdf’s shown in figure 3. The corresponding contours are drawn in figure 4 and assume that the test is of the basic form $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$, with $\mu_1 > \mu_0$, and that the test statistic is $T$. The parameter $\mu$ could be related to the strength of a possible signal for a new particle with unknown mass. Increasing separation between $\mu_0$ and $\mu_1$ could then correspond to increasing amount of data; fixed amount of data and fixed particle mass, but increasing cross section; fixed amount of data and varying particle mass, with the cross section depending on the mass in a known way (i.e. raster scan).

Example 1: $\mu$ is the mean of a Gaussian distribution of known width $\sigma$:

$$T \sim e^{-\frac{1}{2} \left(\frac{t-\mu}{\sigma}\right)^2}. \quad (2)$$

In this case the fixed-hypothesis contours only depend on $\Delta \mu / \sigma$, with $\Delta \mu \equiv \mu_1 - \mu_0$, and have the form:

$$\text{erf}^{-1}(1 - 2 p_1) + \text{erf}^{-1}(1 - 2 p_0) = \frac{\Delta \mu}{\sqrt{2} \sigma}. \quad (3)$$

Figure 4(a) shows three examples of this, with $\Delta \mu / \sigma = 0$ (when the locus is the diagonal line $p_0 + p_1 = 1$), 1.67 and 3.33. As $\Delta \mu / \sigma$ increases, the curves pass closer to the origin.

Example 2: $\mu$ is the mode of a Cauchy distribution with known half-width at half-height $\gamma$:

$$T \sim \frac{\gamma}{\pi \left[\gamma^2 + (t-\mu)^2\right]}. \quad (4)$$

The contours have a simple expression that depends only on $\Delta \mu / \gamma$:

$$\tan\left[(1 - 2 p_1) \frac{\pi}{2}\right] + \tan\left[(1 - 2 p_0) \frac{\pi}{2}\right] = \frac{\Delta \mu}{\gamma}. \quad (5)$$

Example contours are shown in figure 4(b).

---

4Fixed-hypothesis contours on a $(p_0, p_1)$ plot are closely related to ROC (Receiver Operating Characteristic) curves, which have been used for many years in a variety of fields.

5Following standard convention we write $T \sim f(t)$ to indicate that $f$ is the pdf of $T$ (use of the ‘$\sim$’ symbol does not imply any kind of approximation).
Example 3: \( \mu \) is an exponential decay rate:

\[
T \sim \mu e^{-\mu t}
\]

Here the fixed-hypothesis contours depend only on the ratio of \( \mu_1 \) to \( \mu_0 \):

\[
\ln(p_1) = \frac{\mu_1}{\mu_0} \ln(1 - p_0).
\]

An interesting generalization is to perform the test on a combination of \( n \) independent decay time measurements \( T_i \). In this case the likelihood ratio statistic is a one-to-one function of the sum of the measurements, which we therefore take as our test statistic, \( T \equiv \sum_{i=1}^{n} T_i \). The distribution of \( T \) is Gamma\((n, \mu)\), with \( n \) the shape parameter and \( \mu \) the rate parameter:

\[
T \sim \frac{\mu^n t^{n-1} e^{-\mu t}}{\Gamma(n)} ;
\]

the fixed-hypothesis contours depend on \( n \) and on the ratio \( \mu_1/\mu_0 \):

\[
p_1 = 1 - P\left(n, \frac{1}{2} \frac{\mu_1}{\mu_0} \chi_{2n, p}^2 \right),
\]

where \( P(n, z) \) is the regularized incomplete gamma function and \( \chi_{2n, p}^2 \) is the \( p \)-quantile of a chi-squared distribution with \( 2n \) degrees of freedom. Some example contours are shown in figure 4(c). Unlike Gaussian or Cauchy contours, gamma contours are not symmetric around the main diagonal of the plot.

Example 4: \( \mu \) is a Poisson mean:

\[
T \sim \frac{\mu^t}{t!} e^{-\mu} \quad (t \text{ integer}).
\]

In this case the contours are discrete and must be computed numerically. Their dependence on \( \mu_0 \) and \( \mu_1 \) does not simplify. A few examples are plotted in figure 4(d).

A common feature of examples 1-3 above is that the region of the plot above the diagonal \( p_0 + p_1 = 1 \) is empty. This is a general consequence of our definition of one-sided \( p \)-values, and of the fact, suggested by the requirement \( \mu_1 > \mu_0 \), that the bulk of the pdf \( f_1(t) \) under \( H_1 \) lies to the right of the bulk of the pdf \( f_0(t) \) under \( H_0 \). In other words, for any \( t \), the area under \( f_0 \) and to the right of \( t \) is smaller than the corresponding area under \( f_1 \):

\[
\text{For all } t : \int_t^{\infty} f_0(u) \, du \leq \int_t^{\infty} f_1(u) \, du.
\]

On the left-hand side one recognizes \( p_0 \) and on the right-hand side \( 1 - p_1 \), whence the inequality \( p_0 + p_1 \leq 1 \) follows. However, this is only strictly true for continuous pdf’s. For discrete pdf’s, \( p_0 \) and \( p_1 \) both include the finite probability of the observation, so that it may happen that \( p_0 + p_1 > 1 \) when \( f_0 \) and \( f_1 \) are very close to each other. This is evident in figure 4(d) for the case \( \mu_0 = \mu_1 = 10 \).
3.5 Asimov data sets

In a given data analysis problem, any data set (real or artificial) for which the parameter estimators yield the true values is called an Asimov data set [7]. By evaluating a test statistic on an Asimov data set one usually obtains an approximation to the median of that test statistic, and the corresponding $p$-value will be the median $p$-value under the assumed hypothesis. Median $p$-values are used to characterize the sensitivity of an experiment.

A simple example of the use of the fixed-hypothesis contours is that they map the abscissa $p_0 = 0.5$ onto the median value of $p_1$ under $H_0$, and vice-versa, the value $p_1 = 0.5$ is mapped onto the median of $p_0$ under $H_1$. These medians can be directly read off the plot. For the Gaussian case with $\Delta \mu/\sigma = 0.0$, $1.67$, or $3.33$, the median $p_1$ under $H_0$ is $0.5$, $4.7 \times 10^{-2}$, or $4.3 \times 10^{-4}$, respectively. By symmetry of the Gaussian density, these values are also those of the median $p_0$ under $H_1$.

By the invariance of probability statements under one-to-one transformations of random variables, the median $CL_s$ under $H_0$ can be obtained by plugging $p_0 = 1/2$ into the definition of $CL_s$. This yields:

$$\text{Med}_{H_0}(CL_s) = CL_s\bigg|_{p_0=1/2} = \frac{p_1}{1-p_0}\bigg|_{p_0=1/2} = 2 \ p_1\bigg|_{p_0=1/2} = 2 \text{Med}_{H_0}(p_1).$$

(12)

Assuming $H_0$ is true, the median $CL_s$ for testing $H_1$ equals twice the median $p_1$.

3.6 Punzi sensitivity

For large enough separation of the pdf’s, the fixed-hypothesis contour will keep out of the no-decision region. Punzi [8] defines sensitivity as the expected signal strength required for there to be a probability of at least $1 - \alpha_1$ for claiming a discovery with significance $\alpha_0$ (e.g. a probability of 95% for discovery at the level of $2.87 \times 10^{-7}$). This has the advantage that above the sensitivity limit, the data are guaranteed to provide rejection of $H_0$ at the significance level $\alpha_0$, or exclusion of $H_1$ at the significance level $\alpha_1$, or both; the data cannot fall in the no-decision region. In figure 5, the Punzi sensitivity corresponds to a pdf separation for which the $(p_0, p_1)$ contour (not drawn) passes through the intersection of the vertical dot-dashed and horizontal dashed lines. In the following we refer to this intersection as the ‘Punzi point’.

3.7 Effect of one-to-one transformations of the test statistic

$P$-values are probabilities and therefore remain invariant under one-to-one transformations of the test statistic on which they are based. Plots of $p_0$ versus $p_1$ are similarly unaffected, but one must remember that these plots involve two hypotheses, and that effects of a transformation on the pdf’s of the test statistic under $H_0$ and $H_1$ are different. This is the reason that, for example, the $p_0$ versus $p_1$ plot for testing the mode of a Gaussian pdf is not identical to the plot for testing the mode of a Cauchy pdf, even though Gaussian and Cauchy variates are related by one-to-one transformations (see examples 1 and 2 in section 3.4). We take a closer look at this particular case here. Suppose that under $H_0$ ($H_1$) the test statistic $X$ is
Gaussian with mean \( \mu_0 (\mu_1) \) and width \( \sigma \). Then, if \( H_0 \) is true, the transformation

\[
X \rightarrow Y \equiv \mu_c + \gamma \tan \left[ \frac{\pi}{2} \text{erf} \left( \frac{X - \mu_0}{\sqrt{2} \sigma} \right) \right]
\]

(13)

maps \( X \) into a Cauchy variate \( Y \) with mode \( \mu_c \) and scale parameter \( \gamma \). If on the other hand \( H_1 \) is true, \( X \) will be mapped into a variate \( Y \) with pdf

\[
f_y(y) = \frac{\exp \left[ -\frac{1}{2} \left( \frac{\Delta \mu}{\sigma} \right)^2 + \frac{\sqrt{2} \Delta \mu}{\sigma} \text{erf}^{-1} \left( \frac{2}{\pi} \arctan \left( \frac{y - \mu_c}{\gamma} \right) \right) \right]}{\pi \gamma \left[ 1 + \left( \frac{y - \mu_c}{\gamma} \right)^2 \right]},
\]

(14)

which is an asymmetric density that depends on three parameters: \( \mu_c, \gamma, \) and \( \Delta \mu/\sigma \equiv |\mu_1 - \mu_0|/\sigma \); it reduces to a Cauchy density in the limit \( \Delta \mu/\sigma \rightarrow 0 \). Figure 6 compares the two pdf’s. Thus, whereas in \( X \) space we are testing a Gaussian hypothesis against a Gaussian hypothesis with a different mean, in \( Y \) space we are testing a Cauchy hypothesis against a hypothesis with a rather different, asymmetrical distribution. However the \( p_0 \) versus \( p_1 \) plot is the same in both spaces.

Another interesting property of one-to-one transformations of test statistics is that they preserve the likelihood ratio (since the Jacobian of the transformation cancels in the ratio). Thus, if \( f_i \) is the pdf of \( X \) under \( H_i, i = 0, 1 \), and we transform \( X \) into \( Y \) with pdf’s \( g_i \), we have:

\[
\frac{g_0(y)}{g_1(y)} = \frac{f_0(x)}{f_1(x)}.
\]

(15)

Suppose now that the \( X \rightarrow Y \) transformation is the likelihood ratio transformation: \( y \equiv f_0(x)/f_1(x) \). Then it follows from the above equation that

\[
g_0(y) = y g_1(y),
\]

(16)

a useful simplification. If one prefers to work with the logarithm of the likelihood ratio, \( q \equiv \ln y \), and \( h_i(q) \) is the pdf of \( q \) under \( H_i \), then one finds:

\[
h_0(q) = e^q h_1(q).
\]

(17)

Suppose for example that \( f_i(x) \) is Gaussian with mean \( \mu_i \) and width \( \sigma \). The pdf’s of the log-likelihood ratio are then:

\[
h_0(q) = \frac{1}{\sqrt{2\pi \Delta \mu/\sigma}} \exp \left[ -\frac{1}{2} \left( \frac{q - \frac{1}{2} (\Delta \mu/\sigma)^2}{\Delta \mu/\sigma} \right)^2 \right],
\]

(18)

\[
h_1(q) = \frac{1}{\sqrt{2\pi \Delta \mu/\sigma}} \exp \left[ -\frac{1}{2} \left( \frac{q + \frac{1}{2} (\Delta \mu/\sigma)^2}{\Delta \mu/\sigma} \right)^2 \right]
\]

(19)

and it is straightforward to verify equation (17).

The Gauss-versus-Gauss likelihood ratio \( f_0(x)/f_1(x) \) in the above example is invariant under translations and rescalings of the original pdf’s (i.e. under addition of a common
constant to $\mu_0$, $\mu_1$, and $x$; and under multiplication of $\mu_0$, $\mu_1$, $\sigma$, and $x$ by a common factor. These two invariances reduce the three numbers ($\mu_0$, $\mu_1$, and $\sigma$) required to specify the $f_i(x)$ to a single one ($\Delta \mu/\sigma$) for the $h_i(q)$. Note that $\Delta \mu/\sigma$ is the ratio of the difference in means to the standard deviation for the pair $(f_0, f_1)$ as well as $(h_0, h_1)$.

More generally, since the likelihood ratio transformation $x \to q$ is one-to-one, fixed-hypothesis contours obtained from the $h_i(q)$ are identical to those obtained from the $f_i(x)$.

## 4 Outcome probabilities and error rates

A useful feature of $(p_0, p_1)$ plots is that they help us map probabilities under $H_0$ to probabilities under $H_1$ and vice-versa, using a simple graphical method. Suppose for instance that we are interested in the outcome $p_0 \leq 0.3$. When $H_0$ is true this has probability 0.3, since $p_0$ is uniformly distributed under $H_0$. To find the probability under $H_1$ we map the interval $0 \leq p_0 \leq 0.3$ onto the $p_1$ axis using the appropriate contour on figure 5, say the one with $\Delta \mu/\sigma = 1.67$. This yields the interval $0.13 \leq p_1 \leq 1$. Since $p_1$ is uniform under $H_1$, we can conclude that the outcome $p_0 \leq 0.3$ has probability 0.87 under $H_1$. In a similar way, it can be read from the figure that rejection of $H_0$ (the outcome $p_0 \leq \alpha_0$) has probability $1 - \beta_0$ under $H_1$, where $\beta_0$ is the $p_1$ coordinate of the intersection of the line $p_0 = \alpha_0$ with the relevant contour. As for rejection of $H_1$ (the outcome $p_1 \leq \alpha_1$), this has probability $1 - \beta_1$ under $H_0$, where $\beta_1$ is the $p_0$ coordinate of the intersection of the line $p_1 = \alpha_1$ with the contour.

The graphical method allows one to derive the probabilities under $H_0$ and $H_1$ of the four possible outcomes of the double test (see Table 1). In computing these probabilities

| Double Test Outcome | Decision                  | Probability Under $H_0$ | Probability Under $H_1$ |
|---------------------|---------------------------|--------------------------|--------------------------|
| $p_0 \leq \alpha_0$ & $p_1 \leq \alpha_1$ | Reject $H_0$ <br> Reject $H_1$ | $\max(0, \alpha_0 - \beta_1)$ | $\max(0, \alpha_1 - \beta_0)$ |
| $p_0 \leq \alpha_0$ & $p_1 > \alpha_1$ | Reject $H_0$ <br> Fail to reject $H_1$ | $\min(\alpha_0, \beta_1)$ | $\min(1 - \alpha_1, 1 - \beta_0)$ |
| $p_0 > \alpha_0$ & $p_1 \leq \alpha_1$ | Fail to reject $H_0$ <br> Reject $H_1$ | $\min(1 - \alpha_0, 1 - \beta_1)$ | $\min(\alpha_1, \beta_0)$ |
| $p_0 > \alpha_0$ & $p_1 > \alpha_1$ | Fail to reject $H_0$ <br> Fail to reject $H_1$ | $\max(0, \beta_1 - \alpha_0)$ | $\max(0, \beta_0 - \alpha_1)$ |

Table 1: Possible outcomes of the double-test procedure, together with their probabilities under $H_0$ and $H_1$. As expected, the probabilities under a given hypothesis all add up to one.

...
the segment of contour above both the $\alpha_0$ and $\alpha_1$ thresholds has probability $\beta_1 - \alpha_0$ under $H_0$. The probabilities for these two cases can be summarized as $\max(0, \beta_1 - \alpha_0)$, as shown in the table. An important caveat about the table is that the double test allows for the possibility that an unspecified hypothesis other than $H_0$ and $H_1$ could be true, in which case a separate column of probabilities would be needed. It is nevertheless reasonable to use this table for performance optimization purposes, since $H_0$ and $H_1$ are the two main hypotheses of interest.

Using Table 1 one can compute various error rates as well as the power of the double test. In analogy with the nomenclature of Neyman-Pearson tests, we can say that there are two Type-I errors, wrong decisions that are made when $H_0$ is true:

**Type-Ia error:** Rejecting $H_0$ when $H_0$ is true. The probability of this error is

$$\mathbb{P}(p_0 \leq \alpha_0 \mid H_0) = \max(0, \alpha_0 - \beta_1) + \min(\alpha_0, \beta_1) = \alpha_0. \quad (20)$$

This is the Type-I error rate in a standard Neyman-Pearson test of $H_0$ against $H_1$.

**Type-Ib error:** Failing to reject $H_1$ when $H_0$ is true. This has probability

$$\mathbb{P}(p_1 > \alpha_1 \mid H_0) = \min(\alpha_0, \beta_1) + \max(0, \beta_1 - \alpha_0) = \beta_1. \quad (21)$$

It is of course possible to commit both a Type-Ia and a Type-Ib error on the same testing problem. The rate of such double errors is not the product of the individual rates $\alpha_0$ and $\beta_1$, but rather, as Table 1 indicates, their minimum, $\min(\alpha_0, \beta_1)$. Errors that are made when $H_1$ is true are called Type II:

**Type-IIa error:** Rejecting $H_1$ when $H_1$ is true. The probability is

$$\mathbb{P}(p_1 \leq \alpha_1 \mid H_1) = \max(0, \alpha_1 - \beta_0) + \min(\alpha_1, \beta_0) = \alpha_1. \quad (22)$$

**Type-IIb error:** Failing to reject $H_0$ when $H_1$ is true. The rate of this error is

$$\mathbb{P}(p_0 > \alpha_0 \mid H_1) = \min(\alpha_1, \beta_0) + \max(0, \beta_0 - \alpha_1) = \beta_0. \quad (23)$$

This is the Type-II error rate in a standard Neyman-Pearson test of $H_0$ against $H_1$.

The rate for committing both Type-II errors simultaneously is $\min(\alpha_1, \beta_0)$. Finally, there is a Type-III error, which has no equivalent in the Neyman-Pearson setup:

**Type-III error:** Failing to reject $H_0$ and $H_1$ when a third, unspecified hypothesis is true. Without additional information about this third hypothesis it is not possible to calculate the Type-III error rate.

Since there is more than one Type-II error, there is some arbitrariness in the definition of the power of the double test. One possibility is to define it as the probability of committing
neither of the two Type-II errors, that is, as the probability of rejecting $H_0$ and failing to reject $H_1$, when $H_1$ is true:

$$P(p_0 \leq \alpha_0 \& p_1 > \alpha_1 \mid H_1) = 1 - \min(\alpha_1, \beta_0) = \min(1 - \alpha_1, 1 - \beta_0). \quad (24)$$

This is different from the power of the Neyman-Pearson test, which is $1 - \beta_0$. Equation (24) has a simple interpretation if we look at it in terms of the separation between the $H_0$ and $H_1$ pdf’s (see figure 1). At low separation, $\beta_0$ is large, and the power is dominated by our ability to reject $H_0$. At high separation (figure 1c), $\beta_0$ is low, and the power is limited by our willingness to accept $H_1$ (as opposed to a third, unspecified hypothesis).

Instead of using $p$-values to decide between hypotheses, one can use likelihood ratios to evaluate the evidence against them. In this case error rates are replaced by probabilities of misleading evidence. The corresponding discussion can be found in Section 5.3.

5 Likelihood ratios

Rather than using $p$-values for discriminating between hypotheses, it is possible to make use of a likelihood ratio; this would also be the starting point for various Bayesian methods.

5.1 Likelihood-ratio contours

It is instructive to plot contours of constant likelihood ratio $\lambda_{01} \equiv L_0/L_1$ on the $p_0$ versus $p_1$ plot. This needs some thought however, since a likelihood ratio calculation requires three input numbers (the values $\mu_0$ and $\mu_1$ of the parameter $\mu$ under $H_0$ and $H_1$, and the observed value $t$ of the test statistic), whereas a point in the $(p_0, p_1)$ plane only yields two numbers. Our approach here is the following: for a set of contours with given $\lambda_{01}$, we fix the null hypothesis $\mu_0$ in order to map $p_0$ to $t$, then solve the likelihood-ratio constraint $\lambda_{01} = L_0(t, \mu_0)/L_1(t, \mu_1)$ for $\mu_1$, and finally use $t$ and $\mu_1$ to obtain $p_1$. In this way, both the likelihood ratio and the value of $\mu$ under $H_0$ are constant along our likelihood-ratio contours, but in general the value of $\mu$ under $H_1$ varies point by point.

If the test statistic $t$ itself is the likelihood ratio, the above procedure needs to be adjusted, since now the pdf’s of $t$ under $H_0$ and $H_1$ depend on both $\mu_0$ and $\mu_1$ (see for example equations (18) and (19) in section 3.7). There is no longer a likelihood-ratio constraint to solve. Instead, for pre-specified values of $\mu_0$ and $t \equiv \lambda_{01}$, one maps $p_0$ into $\mu_1$, and substitutes $t$, $\mu_0$ and $\mu_1$ into the expression for $p_1$.

Remarkably, for some of the simple cases examined in section 3.4 it turns out that the likelihood-ratio contours are independent of $\mu_0$ and $\mu_1$. The contours do depend on the family of pdf’s to which the data are believed to belong, but not on the particular family members specified by the hypotheses. For the examples of section 3.4, the likelihood-ratio contours take the following forms:

Example 1: $\mu$ is the mean of a Gaussian distribution of known width $\sigma$:

$$[\text{erf}^{-1}(1 - 2p_1)]^2 - [\text{erf}^{-1}(1 - 2p_0)]^2 = \ln(\lambda_{01}). \quad (25)$$

---

Note that a likelihood ratio can be used as a test statistic $T$ within a $p$-value method, or directly, without the calibration provided by $p$-values. It is the latter case that we are considering in this section.
5 LIKELIHOOD RATIOS

**Example 2:** $\mu$ is the mode of a Cauchy distribution with known half-width at half-height $\gamma$:

\[
\frac{1 + \left[ \tan \left( \left( 1 - 2 p_1 \right) \frac{\pi}{2} \right) \right]^2}{1 + \left[ \tan \left( \left( 1 - 2 p_0 \right) \frac{\pi}{2} \right) \right]^2} = \lambda_{01}. \tag{26}
\]

**Example 3:** $\mu$ is an exponential decay rate:

\[
\left[ \frac{P^{-1}(n,p_0)}{P^{-1}(n,1-p_1)} \right]^n e^{P^{-1}(n,1-p_1)-P^{-1}(n,p_0)} = \lambda_{01}, \tag{27}
\]

where $P^{-1}(n,x)$ is the inverse, with respect to the second argument, of the regularized incomplete gamma function (i.e., $y = P^{-1}(n,x)$ is equivalent to $x = P(n,y)$).

**Example 4:** $\mu$ is a Poisson mean:

There is no closed analytical expression, and the contours, which must be computed numerically, depend on $\mu_0$ and $\mu_1$ (as opposed to just their difference or their ratio).

Figure 7 shows the $\lambda_{01} = 0.37, 0.83, 1.0, 1.2$ and $2.7$ contours for these four cases. Along the diagonal $p_1 = 1 - p_0$ (or close to it in the Poisson case), the $H_0$ and $H_1$ pdf’s are identical and $\lambda_{01}$ is unity. For symmetric pdf’s such as the Gaussian and Cauchy, the likelihood ratio is also unity along the other diagonal line, $p_1 = p_0$. This is because the observed value of the test statistic is then situated midway between the pdf peaks. For asymmetric pdf’s such as the gamma and Poisson the likelihood ratio is no longer unity when $p_1 = p_0$, but there is still a $\lambda_{01} = 1$ contour that starts at the origin of the plot and rises toward its middle. Above and to the left of this curve, the likelihood ratio favors $H_1$; below it, $H_0$ is favored.

Loosely stated, the central limit theorem asserts that the distribution of the mean of $n$ measurements converges to a Gaussian as the sample size $n$ increases. When the test statistic is defined as such a mean, likelihood ratio contours will converge to their shape for a Gauss versus Gauss test. This is illustrated in figure 8 for the exponential/gamma case and in figure 9 for the Poisson case.

5.2 Comparison of p-values and likelihood ratios

A criticism against $p$-values is that they overstate the evidence against the null hypothesis [9, 10]. One aspect of this is that $p$-values tend to be impressively smaller than likelihood ratios. The fact that they are not identical is no surprise. Likelihoods are calculated as the height of the relevant pdf at the observed value of the statistic $T$, while $p$-values use the corresponding tail area. Furthermore a $p$-value uses the pdf of a single hypothesis, while a likelihood ratio requires the pdf’s of two hypotheses. As can be seen from figure 10, at constant $p_0$ (even if it is very small) $L_0/L_1$ can have a range of values, sometimes favoring $H_1$, sometimes $H_0$. This will depend on the separation of the pdf peaks. Thus for Gaussian pdf’s, a $p_0$ value of $3 \times 10^{-7}$ will favor $H_1$ provided $0 < \Delta \mu / \sigma < 10$, but for larger $\Delta \mu / \sigma$ the observed test statistic is closer to the $H_0$ peak than to $H_1$’s, and so even though the data are very inconsistent with $H_0$, the likelihood ratio still favors $H_0$ as compared with $H_1$.

Another example is given in Table 2; this uses simple Poisson hypotheses for both $H_0$ and $H_1$. It involves a counting experiment where the null hypothesis $H_0$ predicts 1.0 event
5.3 Probability of misleading evidence in likelihood ratio tests

When studying the evidence provided by the likelihood ratio \( L_0/L_1 \) in favor of hypothesis \( H_0 \), an important quantity is the probability of misleading evidence. This is defined by Royall [11] as the probability of observing \( L_0/L_1 > k \), for a given \( k > 1 \), when \( H_1 \) is true. Figure 10 shows how this probability can be determined by drawing the appropriate fixed-hypothesis contour (dashed line, here corresponding to \( \Delta \mu/\sigma = 1.67 \)) on top of the likelihood-ratio contour of interest (here \( L_0/L_1 = 1.2 \)). Larger likelihood-ratio contours intersect the dashed line at lower values of \( p_1 \). Therefore the probability of a larger likelihood ratio under \( H_1 \),

|                  | First data set | Second data set |
|------------------|----------------|-----------------|
| \( H_0 \)       | Poisson, \( \mu = 1 \) | Poisson, \( \mu = 10 \) |
| \( H_1 \)       | Poisson, \( \mu = 10 \) | Poisson, \( \mu = 100 \) |
| \( n_{\text{obs}} \) | 10             | 30              |
| \( p_0 \)       | \( 1.1 \times 10^{-7} \) | \( 2.5 \times 10^{-7} \) |
|                  | 5.2\( \sigma \)   | 5.0\( \sigma \)   |
| \( p_1 \)       | 0.58             | 2.2 \times 10^{-16} |
|                  | -0.2\( \sigma \)  | 8.1\( \sigma \)   |
| \( L_0/L_1 \)   | \( 8 \times 10^{-7} \) | \( 1.2 \times 10^{+9} \) |

Strongly favors \( H_1 \)  Strongly favors \( H_0 \)

Table 2: Comparing \( p \)-values and likelihood ratios

and the alternative \( H_1 \) predicts 10.0 events. In a first run 10 events are observed; both \( p_0 \) and the likelihood ratio disfavor \( H_0 \). Then the running time is increased by a factor of 10, so that the expected numbers according to \( H_0 \) and \( H_1 \) both increase by a factor of 10, to 10.0 and 100.0 respectively. With 30 observed events, \( p_0 \) corresponds to about 5\( \sigma \) as in the first run, but despite this the likelihood ratio now strongly favors \( H_0 \). This is simply because the 5\( \sigma \) \( n_{\text{obs}} = 10 \) in the first run was exactly the expected value for \( H_1 \), but with much more data the 5\( \sigma \) \( n_{\text{obs}} = 30 \) is way below the \( H_1 \) expectation. In fact, in the second run, the \( p \)-value approach rejects both \( H_0 \) and \( H_1 \).

More data corresponds to increasing pdf separation. Thus we are moving downwards on a line at constant \( p_0 \), resulting in a smaller \( p_1 \), and provided \( p_1 < 1/2 \), a larger \( L_0/L_1 \). This is one motivation for hypothesis selection criteria that employ a decreasing value for the rejection threshold \( \alpha_0 \) as the amount of data increases.

It is interesting to contrast the exclusion regions for \( H_1 \) provided by cuts on \( p_1 \) and on the likelihood ratio \( L_0/L_1 = \lambda_{01} \) (see figures 2 and 7 respectively). The main differences are at small and at large \( p_0 \), where the excluded region extends up to \( p_1 = \alpha_1 \) for \( p_1 \) cuts, but to much smaller \( p_1 \) values for cuts on the likelihood ratio. At large \( p_0 \), the likelihood cuts resemble more those provided by \( CL_s \) (see figure 2). At small \( p_0 \), the likelihood cuts correspond to the exclusion \( p_1 \) cut-off \( \alpha_1 \) effectively decreasing as the \( H_0 \) and \( H_1 \) pdf’s become more separated (e.g., as the amount of data collected increases).
i.e., the probability of misleading evidence, is given by the $p_1$-coordinate of the intersection point $X$.

It is of course also possible to calculate the probability of misleading evidence that favors $H_1$ when $H_0$ is actually true. For this we look at the intersection of a fixed-hypothesis contour with a likelihood-ratio contour for which $L_0/L_1 < 1$, and we are concerned about even smaller likelihood ratio values\(^7\). The probability of misleading evidence is then given by the $p_0$-coordinate of that intersection.

Careful inspection of the shape of the likelihood-ratio contours in figure 10 reveals that the probabilities of misleading evidence are small at small values of $\Delta \mu/\sigma$ (where there is little chance of obtaining strong evidence in favor of either hypothesis), then increase to a maximum, and finally become small again at large $\Delta \mu/\sigma$.

The determination of probabilities of misleading evidence from $p_0$ and $p_1$ coordinates may give the impression that these probabilities could be calculated from the observed likelihood ratio and reported ‘post-data’. According to the likelihoodist paradigm of statistics, this view is incorrect. As emphasized in ref. [11], all the relevant evidence about the hypotheses is contained in the likelihood ratio. The probabilities of misleading evidence should only be used for experiment-planning purposes, by calculating them for standard likelihood ratio values. By convention, a value of $L_0/L_1 = 8$ is defined as ‘fairly strong’ evidence in favor of $H_0$, whereas $L_0/L_1 = 32$ is said to be ‘strong’ evidence. Likelihood-ratio contours for these values would not be visible on a linear plot such as figure 10. As shown in figure 11, a log-log plot gives much better visualization.

6 Famous puzzles in statistics

The topic of $p$-values has generated many controversies in the statistics literature. In this section we use $p_0$ versus $p_1$ plots to discuss a couple of famous puzzles that initiated some of these controversies.

6.1 Sampling to a foregone conclusion

Suppose that in searching for a new physics phenomenon we adopt the following procedure:

1. Choose a discovery threshold $\alpha_0$, and let $\mathcal{E}$ be a set of candidate events, initially empty.

2. Add one candidate event to $\mathcal{E}$ and compute $p_0$, the $p$-value to test the background-only hypothesis $H_0$ based on all events in $\mathcal{E}$.

3. If $p_0 \leq \alpha_0$, reject $H_0$, claim discovery, and stop; otherwise go back to step 2.

If the new physics phenomenon can be modeled by a simple hypothesis $H_1$, we can also compute the $p$-value $p_1$ at step 2, and the whole procedure can be represented by a random walk in the $p_0$ versus $p_1$ plane. At each step of the walk, the $p$-values are updated with the

\(^7\)Just as the cut-offs $\alpha_0$ and $\alpha_1$ for $p_0$ and $p_1$ are usually taken to be (very) different, similarly when using likelihood ratio cuts there is generally no necessity for one to be the reciprocal of the other.
addition of a random new event. Four examples of such random walks are shown in figure 12, two assuming that \( H_0 \) is true, and two assuming that \( H_1 \) is true.

What is the chance of the above procedure stopping when \( H_0 \) is true? In other words, what is the probability of incorrectly claiming discovery with this procedure? The answer, perhaps surprisingly, is 100%, due to a result from probability theory known as the Law of the Iterated Logarithm (LIL). The latter applies to any sequence of random variables \( \{X_1, X_2, X_3, \ldots \} \) that are independent and identically distributed with finite mean \( \mu_0 \) and variance \( \sigma^2 \). Consider the Z-values constructed from partial sums of the \( X_i \):

\[
Z_n = \frac{\frac{1}{n} \sum_{i=1}^{n} X_i - \mu_0}{\sigma / \sqrt{n}}, \quad \text{for } n = 1, 2, 3, \ldots
\]  

(28)

The LIL states that with probability 100% the inequality

\[
|Z_n| \geq (1 + \delta) \sqrt{2 \ln \ln n}
\]  

(29)

holds for only finitely many values of \( n \) when \( \delta > 0 \) and for infinitely many values of \( n \) when \( \delta < 0 \). At large \( n \) the \( Z_n \) will be approximately standard normal and correspond to the \( p \)-values

\[
p_0(n) = \int_{|Z_n|}^{\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}} \, dt = \frac{1}{2} \left[ 1 - \text{erf} \left( \frac{|Z_n|}{\sqrt{2}} \right) \right],
\]

(30)

so that the LIL of eqn. 29 can be rephrased as stating that, as \( n \) increases, the inequality

\[
p_0(n) \leq \frac{1}{2} \left[ 1 - \text{erf} \left( (1 + \delta) \sqrt{\ln \ln n} \right) \right]
\]

(31)

occurs infinitely many times if \( \delta < 0 \). In particular, regardless of how small \( \alpha_0 \) is, at large \( n \) the right-hand side of (31) will become even smaller; therefore, if \( \delta < 0 \) the LIL guarantees that \( p_0(n) \) will cross the discovery threshold at some \( n \), allowing the search procedure to stop with a discovery claim. Crucial to this guarantee is the fact that inequality (31) occurs infinitely many times for \( \delta < 0 \); it will then certainly occur at \( n \) large enough to force a crossing of the discovery threshold. In contrast, for \( \delta > 0 \) there is a value of \( n \) beyond which there are no crossings (and there may indeed be none at all for any \( n \)); rejection of \( H_0 \) is not guaranteed to occur.

In terms of designing a coherent search procedure, one can view the LIL as defining an \( n \)-dependent boundary

\[
\alpha_{\text{LIL}}(n) = \frac{1}{2} \left[ 1 - \text{erf} \left( \sqrt{\ln \ln n} \right) \right].
\]

(32)

Any discovery threshold with an \( n \)-dependence that causes it to exceed this boundary at large \( n \) is unsatisfactory since it is guaranteed to be crossed. It is instructive to draw the LIL boundary on a \( p_0 \) versus \( p_1 \) plot. To each value of \( n \) there corresponds a fixed-hypothesis contour on the plot (see figure 13). When testing \( H_0 \), one point on the LIL boundary is then given by the intersection of that contour with the line \( p_0 = \alpha_{\text{LIL}}(n) \). By connecting all such points across contours one obtains the blue lines drawn in figure 13 and in figure 12(a) and (c) (note that \( n = 2 \) is the smallest integer for which \( \alpha_{\text{LIL}}(n) \) can be computed). When testing \( H_1 \), the LIL boundary is given by the intersections of the contours with the lines \( p_1 = \alpha_{\text{LIL}}(n) \), as shown in figure 12(b) and (d).
Focusing on plots (a) and (c) of figure 12, we note that when \( H_0 \) is true, the \( p_1 \) coordinate of random walks tends to decrease very rapidly as a function of \( n \). The \( p_0 \) coordinate is more stable, but it does exhibit occasional excursions towards low \( p_0 \) values. The LIL states that the number of such excursions to the left of the blue line is finite (not infinite) as \( n \) goes to infinity. However, any threshold curve to the right of the blue line will be crossed infinitely many times. A constant threshold of the form \( p_0 = \alpha_0 \) will be to the right of the blue line at large \( n \) and is therefore unsatisfactory, in contrast with a threshold curve in the form of a likelihood ratio contour (see figure 12) or with an \( n \) dependence of the form \( \alpha_0/\sqrt{n} \) (see figure 13).

In particle physics we have constant thresholds of \( 3\sigma \left( \alpha_0 = 1.35 \times 10^{-3} \right) \) and \( 5\sigma \left( \alpha_0 = 2.87 \times 10^{-7} \right) \). Due to the iteration of logarithms in the LIL, it takes an enormously large value of \( n \) for the blue line to cross these thresholds, so that the problem is not practically relevant. The statistician I. J. Good once remarked that a statistician could “cheat by claiming at a suitable point in a sequential experiment that he has a train to catch […] But note that the iterated logarithm increases with fabulous slowness, so that this particular objection to the use of tail-area probabilities is theoretical rather than practical. To be reasonably sure of getting \( 3\sigma \) one would need to go sampling for billions of years, by which time there might not be any trains to catch.” [12]

The LIL provides the weakest known constraint on the \( n \)-dependence of discovery thresholds. It is a purely probabilistic characterization of tail probabilities under a single hypothesis. Much more stringent constraints can be obtained by introducing an alternative hypothesis and using statistical arguments (see for example [13]).

### 6.2 The Jeffreys-Lindley paradox

The Jeffreys-Lindley paradox occurs in tests of a simple \( H_0 \) versus a composite \( H_1 \), for example:

\[
H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0.
\]  

The paradox is that for some values of the observed test statistic \( t \), the value of \( p_0 \) can be small enough to cause rejection of \( H_0 \) while the Bayes factor favors \( H_0 \). Writing \( L_0, L_1(\mu) \) for the likelihood under \( H_0 \), respectively \( H_1 \), the Bayes factor is defined by

\[
B_{01} \equiv \frac{L_0}{\int L_1(\mu) \pi_1(\mu) \, d\mu},
\]

where \( \pi_1(\mu) \) is a prior density for \( \mu \) under \( H_1 \). In order to understand the origin of the paradox, it helps to note that this Bayes factor can be rewritten as a weighted harmonic average of likelihood ratios:

\[
B_{01} = \left[ \int \frac{1}{\lambda_{01}(\mu)} \pi_1(\mu) \, d\mu \right]^{-1},
\]

with \( \lambda_{01}(\mu) \equiv L_0/L_1(\mu) \). As this formula suggests, it will prove advantageous to look at the composite \( H_1 \) as a collection of simple hypotheses about the value of \( \mu \), each with its own simple-to-simple likelihood ratio \( \lambda_{01}(\mu) \) to the null hypothesis \( H_0 \).
In the following subsection we use this idea to develop basic insight into the origin of the Jeffreys-Lindley paradox. Later subsections take a deeper look at the conditions under which the paradox appears and at possible solutions.

### 6.2.1 Basic insight

Figure 14 illustrates the paradox for the case where the pdf of the test statistic \( t \) is Gaussian with mean \( \mu \) and standard deviation \( \sigma = 1 \). As in Section 5.2, consider a vertical line at the relevant \( p_0 \) in plot (a); this crosses a series of different \( \lambda_{01} \) contours. At point \( b \), the \( H_0 \) and \( H_1 \) pdf’s are identical (see plot (b)), and the likelihood ratio is unity. Point \( c \) is at \( p_1 = 0.5 \), with the \( H_1 \) pdf having its maximum exactly at the position of the data statistic \( t \). The likelihood ratio now favors \( H_1 \), and is in agreement with the small \( p_0 \) value in rejecting \( H_0 \). Plots (d) and (e) show even larger separations between \( H_0 \) and \( H_1 \). In plot (d), corresponding to point \( d \) in plot (a), the position of the \( H_1 \) pdf is such that the data statistic \( t \) is midway between the \( H_0 \) and \( H_1 \) peaks. Thus \( p_0 = p_1 \) and point \( d \) lies on the diagonal of plot (a), with the likelihood ratio again unity. Finally, with the larger separation of plot (e), the likelihood ratio now favors \( H_0 \), even though \( p_0 \) is small; the likelihood ratio and \( p_0 \) lead to opposite conclusions.

To go from the series of simple \( H_1 \)’s to the composite \( H_1 \) with unspecified \( \mu \) in the Jeffreys-Lindley paradox, we take the weighted harmonic average of the likelihood ratios \( \lambda_{01} \), with the weighting given by the prior \( \pi_1(\mu) \) as in equation (35). As we integrate along the vertical line in plot (a), the contributions between points \( b \) and \( d \) favor \( H_1 \). Lower down, from \( d \) to \( e \) and beyond, \( H_0 \) is favored. The Bayes factor will thus end up favoring \( H_0 \) if the integration range is wide enough\(^8\) and suitably weighted by the prior \( \pi_1(\mu) \). This explains the mechanism by which the Jeffreys-Lindley paradox can occur.

### 6.2.2 Regions in the plane of \( p_0 \) versus prior-predictive \( p_1 \)

To visualize the conditions under which the Jeffreys-Lindley paradox appears, we generalize the \( p_0 \) versus \( p_1 \) plot to the case of a composite \( H_1 \) by making use of the prior-predictive \( p \)-value [14]; this is a prior-weighted average \( p \)-value over \( H_1 \):

\[
p_{1pp} \equiv \int \pi_1(\mu) \int_{-\infty}^{t_0} f(t \mid \mu) dt d\mu,
\]

where \( f(t \mid \mu) \) is the pdf of \( T \) and \( t_0 \) its observed value. To fix ideas, assume that \( f \) is Gaussian with mean \( \mu \) and standard deviation \( \sigma \) (not necessarily equal to 1), and that the prior \( \pi_1(\mu) \) is the indicator function of the interval \([\mu_0, \mu_0 + \tau]\) for some positive \( \tau \):

\[
\pi_1(\mu) \equiv \pi_1(\mu \mid \tau) = \frac{1}{\tau} \mathbf{1}_{[\mu_0, \mu_0 + \tau]}(\mu) = \begin{cases} \frac{1}{\tau} & \text{if } \mu_0 < \mu \leq \mu_0 + \tau, \\ 0 & \text{otherwise}. \end{cases}
\]

\(^8\)The value of \( \mu \), which determines the separation between the corresponding simple \( H_1 \) and \( H_0 \), varies non-linearly with distance along the line \( bcde \), such that there is generally a far wider range of \( \mu \) values below the diagonal than above it.
In the absence of detailed prior information about $\mu$, one could think of this prior as modeling the range of $\mu$ values deemed to be theoretically and/or experimentally relevant for the search experiment of interest. In any case the exact shape of $\pi_1$ is not material to the paradox, only the ratio of length scales $\tau/\sigma$ is. Reference [15] discusses the choice of $\tau$ in several particle physics experiments.

Use of $p_{1pp}$ calls for a couple of caveats. First, a small value of $p_{1pp}$ does not imply that all values of $\mu$ under $H_1$ are disfavored. In general it only provides evidence against the overall model (prior plus pdf) under $H_1$. However with the particular choice of prior (37), and assuming that $\tau/\sigma$ is sufficiently large, small $p_{1pp}$ implies that the vast majority of $\mu$ values under $H_1$ are unable to explain the data. Second, the distribution of $p_{1pp}$ under a fixed value of $\mu$ in $H_1$ is not uniform. Hence, in a linear plot of $p_0$ versus $p_{1pp}$, distances along the $p_{1pp}$ axis cannot be interpreted as probabilities under a fixed $\mu$ in $H_1$ (contrast Section 4). However, such distances can still be interpreted as prior-predictive probabilities, with pdf given by the integral of $\int t \, dt$ over $\pi_1(\mu)$.

Figure 15 shows fixed-hypothesis contours (fixed $\mu_0$, $\sigma$, and $\tau$) and constant Bayes factor contours in the $p_0$ versus $p_{1pp}$ plane. For the testing situation examined here, fixed-hypothesis contours only depend on the ratio $\tau/\sigma$ and are labeled accordingly. The constant Bayes factor contours are labeled by the value of $B_{01}$. For $\tau/\sigma = 0$, $H_1$ coincides with $H_0$ and the resulting fixed-hypothesis contour is a subset of the $B_{01} = 1$ contour. As $\tau/\sigma$ increases, the ability of the test to distinguish between $H_0$ and $H_1$ also increases. Figure 16 presents a log-log version of the same plot. This allows the drawing of contours with a wider range of Bayes factor values, $B_{01} = 1, 3, 20$, and 150. According to ref. [16], a Bayes factor between 1 and 3 represents evidence “not worth more than a bare mention;” between 3 and 20, “positive;” between 20 and 150, “strong;” and greater than 150, “very strong.” One can identify the following regions in the plot:

**Upper Left:** At small values of $p_0$ and small values of $\tau/\sigma$ (red contour region), the Bayes factors disfavor $H_0$. There is agreement between Bayes factors and $p$-values.

**Lower Left:** At small values of $p_0$ and large values of $\tau/\sigma$ (green contour region), the Bayes factors favor $H_0$. There is disagreement between Bayes factors and $p$-values. This is where the Jeffreys-Lindley paradox shows up. For a numerical example, consider a $p_0$ value of $2.87 \times 10^{-7}$ (5$\sigma$); the corresponding Bayes factor in favor of $H_0$ will then be 1, 3, 20, or 150, if the ratio $\tau/\sigma$ is approximately $6.7 \times 10^5$, $2.0 \times 10^6$, $1.3 \times 10^7$, or $1.0 \times 10^8$, respectively. Note the extremely large values of $\tau/\sigma$ required for producing the paradox. This is a consequence of the stringent 5$\sigma$ convention applied to discovery claims in particle physics.

**Upper Right:** This is a region with relatively large values of $p_0$ and $p_1$, and where the Bayes factor hovers around 1. Regardless of how one looks at it, there is not enough evidence to decide between $H_0$ and $H_1$.

**Lower Right:** Here $p_1$ is small and $B_{01}$ large. Both support the rejection of $H_1$ in favor of $H_0$.

Curves of constant $\tau/\sigma$ represent fixed experimental conditions, such that repeated observations would fall randomly (but not necessarily uniformly) along one such curve. On a
given curve there is agreement between \( p \)-values and Bayes factors at high and low \( p_0 \), but somewhere in between there is a region of either no-decision (low \( \tau/\sigma \)) or paradox (high \( \tau/\sigma \)).

### 6.2.3 Possible solutions to the paradox

Over the years many solutions have been proposed to the Jeffreys-Lindley paradox. Here we briefly illustrate two arguments.

The first argument essentially blames the \( p \)-value method for the paradox and argues that with increasing values of \( \tau/\sigma \) the \( p \)-value discovery threshold \( \alpha_0 \) should be lowered. This argument is usually applied to the situation where \( \sigma \) depends on a sample size \( n \), so that \( \tau/\sigma \) is proportional to \( \sqrt{n} \). In figure 16 for example, one could think of the contours \( \tau/\sigma = 1, 10, 100, \ldots \) as corresponding to \( n = 1, 100, 10^4, \ldots \), respectively. If one chooses a discovery threshold of 1\% on the \( \tau/\sigma = 1 \) contour, 0.1\% on the \( \tau/\sigma = 10 \) contour, and so on, the dot-dashed curve labeled \( p_0 = \alpha_0/\sqrt{n} \) (where \( \alpha_0 \) is the discovery threshold on the \( \tau/\sigma = 1 \) contour) is obtained. At large \( \tau/\sigma \) this curve follows pretty closely the shape of the constant Bayes factor contours. Thus, cutting on \( p_0 < \alpha_0/\sqrt{n} \) instead of \( p_0 < \alpha_0 \) avoids the Jeffreys-Lindley paradox. Interestingly, this is the same solution that was proposed to avoid sampling to a foregone conclusion in section 6.1.

In a similar vein, it has been argued [17] that for experiments that collect more and more data, the realistic values of \( \mu \) to be considered under \( H_1 \) (assuming that no evidence for \( \mu > \mu_0 \) has been obtained, and we still believe that a small difference is possible) should be those that are closer and closer to \( \mu_0 \). Thus the prior \( \pi(\mu | \tau) \) in equation (37) should become narrower (smaller \( \tau \)), and this prevents \( B_{01} \) favoring \( H_0 \) (as shown in figure 16).

For the second argument, note that in the region of disagreement between \( p_0 \) and Bayes factors, both \( p_0 \) and \( p_{1pp} \) tend to be small: one is in the double-rejection region of the test for most values of \( \mu \) under \( H_1 \). This should alert the experimenter to the possibility that a third hypothesis may be true, or that there may be a mismodeling error. One such error could be that \( H_0 \), rather than a point null hypothesis, is in fact an interval hypothesis with width \( \epsilon \). Thus, instead of (33), one should really be testing

\[
H_0 : \mu_0 - \epsilon < \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0.
\]

We consider two different regimes for \( \epsilon \). The first has \( \epsilon/\sigma = 0.01 \) or 1, corresponding to a small or moderate widening of the original \( H_0 \). The second regime uses \( \epsilon/\sigma = 100 \) or \( 10^4 \), which almost changes \( H_0 \) to \( \mu \leq \mu_0 \), the complement of \( H_1 : \mu > \mu_0 \). For both regimes one will need to introduce a prior \( \pi_0(\mu) \) for \( \mu \) under \( H_0 \), and the Bayes factor becomes:

\[
B_{01} \equiv \int L_0(\mu) \pi_0(\mu) \, d\mu \int L_1(\mu) \pi_1(\mu) \, d\mu.
\]

For the \( p \)-value under \( H_0 \) one could again consider a prior-predictive version:

\[
p_{0pp} \equiv \int \pi_0(\mu) \int_{\mu_0}^{+\infty} f(t | \mu) \, dt \, d\mu,
\]
or choose a frequentist approach, such as the supremum $p$-value \[18\]:

$$p_{0\text{sup}} \equiv \sup_{\mu \in [\mu_0 - \epsilon, \mu_0]} \int_{-\infty}^{+\infty} f(t | \mu) \, dt. \tag{41}$$

Figures 17 and 18 illustrate the effect of these two definitions on the Jeffreys-Lindley paradox. Note first that in both cases one recovers figure 16 when $\epsilon/\sigma$ is small. When the prior-predictive $p_{0\text{pp}}$ of equation (40) is used (figure 17), a given observation above $\mu_0$ becomes more significant since its $p_0$-value is averaged over $\mu$ values below $\mu_0$. This causes the fixed-hypothesis contours to be compressed towards low $p_0$. At the same time, the Bayes factor of such an observation tends to decrease due to the numerator being replaced by an average; this causes the constant Bayes factor contours to move down. The net effect of these contour changes is to leave the paradox in place. This can be seen, for example, by considering the point with $p_0 = 10^{-4}$ on the $\tau/\sigma = 10 000$ contour. In all four plots of figure 17 this point hardly moves, having a Bayes factor $B_{01}$ close to 3.

On the other hand, when the supremum $p$-value of equation (41) is used for $p_0$ (figure 18), only the constant Bayes factor contours change. The fixed-hypothesis contours stay the same, because the supremum of $p_0$ over the interval $[\mu_0 - \epsilon, \mu_0]$ is attained at $\mu = \mu_0$. For fixed $\tau/\sigma$, increasing $\epsilon/\sigma$ causes the paradoxical region to be pushed toward larger values of $p_0$ and smaller values of $p_1$. Eventually the $p$-values agree with the Bayes factor and the paradox disappears. In principle one could even tune the value of $\epsilon/\sigma$ to obtain $B_{01} = 1$ at a specified value of $p_0$ (keeping $\tau/\sigma$ constant). Smaller values of $p_0$ would then correspond to $B_{01}$ disfavoring $H_0$, and larger $p_0$ to $B_{01}$ favoring $H_0$.

We conclude from this discussion of the second argument that introduction of a scale $\epsilon$ under $H_0$ is by itself not sufficient to suppress the paradox. One also needs to specify how to handle $\epsilon$ in the computation of $p_0$. Furthermore, as shown in figure 18 the paradox is not fully suppressed unless $\epsilon/\sigma$ is substantially larger than 1, of the same order as $\tau/\sigma$. Thus, the hierarchy $\epsilon \ll \sigma \ll \tau$, presented in ref. [15], is sufficient to produce the paradox, but not necessary. When using the supremum $p$-value, the condition $[\epsilon \ll \tau$ and $\sigma \ll \tau]$ is both necessary and sufficient.

### 6.2.4 Simple versus simple version of the Jeffreys-Lindley paradox

Figures 15 and 16 do not look very different from figures 10 and 11 discussed in the sections on likelihood ratios, in spite of the use of a different $p_1$ definition. This is a consequence of the fact that the Jeffreys-Lindley paradox can be reformulated in the context of a simple versus simple test.\(^{10}\) As noted at the beginning of section 6.2, the paradox occurs for tests of the form:

$$\text{Test 1: } H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu > \mu_0, \tag{42}$$

using a test statistic $T \sim f(t | \mu)$ and assuming a prior $\pi_1(\mu | \tau)$ for $\mu$ under $H_1$, where $\tau$ characterizes the scale of $\pi_1$.

---

\(^9\)Note that the supremum of $p_1$ over the interval $[\mu_0, \mu_0 + \tau]$ is also attained at $\mu = \mu_0$, so that $p_{0\text{sup}} + p_{1\text{sup}} = 1$ (for continuous pdf’s). Therefore there is nothing to be learned from a plot of $p_{0\text{sup}}$ versus $p_{1\text{sup}}$.

\(^{10}\)The simple versus simple scenario outlined in this section is unrelated to the basic insight described in section 6.2.1.
To proceed with the reformulation, introduce a variate $X$ whose randomness is the result of a two-step generating process: $X \sim f(x \mid \mu)$, where $\mu \sim \pi_1(\mu \mid \theta)$. Thus, for fixed $\theta$ the distribution of $X$ is:

$$
\tilde{f}(x \mid \theta) = \int \pi_1(\mu \mid \theta) f(x \mid \mu) \, d\mu.
$$

If for example $f(x \mid \mu)$ is Gaussian with mean $\mu$ and standard deviation $\sigma$, and $\pi_1(\mu \mid \theta)$ is the indicator function of the interval $[\mu_0, \mu_0 + \theta]$, this will yield:

$$
\tilde{f}(x \mid \theta) = \frac{1}{2\theta} \left[ \text{erf} \left( \frac{\mu_0 + \theta - x}{\sqrt{2}\sigma} \right) - \text{erf} \left( \frac{\mu_0 - x}{\sqrt{2}\sigma} \right) \right].
$$

As $\theta \to 0$ this pdf approaches $f(x \mid \mu_0)$, and we will assume that this remains true for any choice of prior $\pi_1$ (i.e., that for $\theta = 0$, $\pi_1(\mu \mid \theta)$ is a delta function at $\mu = \mu_0$).

Consider now the simple versus simple test:

$$
\text{Test 2: } H_0 : \theta = 0 \text{ versus } H_1 : \theta = \tau,
$$

using the test statistic $X \sim \tilde{f}(x \mid \theta)$. Test 2 is designed to determine whether or not the additional source of randomization $\pi_1$ is present in the process that generates $X$. If $\theta = 0$, there is no additional randomization and $\mu = \mu_0$. On the other hand, if $\theta = \tau$, additional randomization is present, its magnitude agrees with the prediction under $H_1$ in Test 1, and we must have $\mu > \mu_0$. Tests 1 and 2 yield the same information about $\mu$. However, since $H_1$ is composite in Test 1 but simple in Test 2, this has some interesting consequences. The $p$-value $p_1$ is prior-predictive in Test 1 but standard frequentist in Test 2 (as can be seen by interchanging the order of integration in equation (36)). The Bayes factor in Test 1 is a likelihood ratio for $H_0$ versus $H_1$. Tests 1 and 2 yield the same $p_0$ versus $p_1$ plots. The Jeffreys-Lindley paradox, which is a disagreement between $p$-values and Bayes factors in Test 1, is a disagreement between $p$-values and likelihood ratios in Test 2. This purely frequentist version of the Jeffreys-Lindley paradox is illustrated in figure 19 using the pdf $\tilde{f}(x \mid \theta)$ of equation (44). It shows that when testing a narrow distribution against a very broad one, it is possible to observe data with small $p$-value under the narrow-distribution hypothesis and yet large likelihood ratio in favor of that hypothesis.

Even though Test 1 is not of the simple versus simple type, it is possible to define a likelihood ratio statistic for it, as the ratio of the likelihood under $H_0$ to the maximized likelihood under $H_1$, where the maximum is taken over $\mu > \mu_0$. An interesting quantity is the Ockham factor, defined as the ratio of the Bayes factor to this likelihood ratio. For Test 1 the Ockham factor is approximately $\tau/\sqrt{2\pi\sigma}$. This is approximately proportional to the ratio of the widths of the distributions under $H_1$ and $H_0$ in Test 2. More interestingly, at large $\tau/\sigma$ the Type-II error rate of the simple versus simple test equals $N_0\sigma/\tau$, where $N_0$ is the number of standard deviations corresponding to the cutoff $\alpha_0$ used to reject $H_0$. Hence the Type-II error rate is inversely proportional to the Ockham factor: if the alternative hypothesis is true, the probability of rejecting the null with $p_0$ increases with $\tau/\sigma$, but so does the disagreement between $p_0$ and $B_{01}$! Referring again to figure 19, we see that for small $x$ values (say below $x = 2$), Bayes factors and $p$-values both favor $H_0$. At high $x$ they both disfavor $H_0$. In between there is a region where agreement between $p$-values and Bayes factors depends on the Ockham factor.
7 NUISANCE PARAMETERS

There are many methods for eliminating nuisance parameters from \( p \)-value calculations (see for example [18]), and the choice of method will generally have an effect on the construction of \( p_0 \) versus \( p_1 \) plots. We start with a couple of examples.

First consider the situation where one makes \( n \) measurements \( x_i \) from a Gaussian population with unknown mean \( \mu \) and unknown width \( \sigma \). A sufficient statistic consists of the pair \((\bar{x}, s)\), where \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \) is the sample mean and \( s = \left[ \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2} \) is the sample standard deviation. To test the hypotheses \( H_i : \mu = \mu_i \) (\( i = 0, 1 \)), the classical approach uses the test statistics \( t_i = \sqrt{n}(\bar{x} - \mu_i)/s \), which have Student’s \( t \) distribution under the respective \( H_i \). Thus one can calculate the \( p \)-values \( p_0 \) and \( p_1 \). Unfortunately the relation between \( p_0 \) and \( p_1 \) is not one-to-one: from \( p_0 \) one can obtain \( t_0 \), but from \( t_0 \) one cannot extract both \( \bar{x} \) and \( s \), which are needed to compute \( t_1 \) and then \( p_1 \). Hence it is not possible to make a plot of \( p_0 \) versus \( p_1 \). This problem is related to the fact that the power of the \( t \) test depends on the unknown value of \( \sigma \) and not just on the significance threshold \( \alpha \).

For the second example we consider the observation of a Poisson variate \( N \), whose mean is the product of a parameter of interest \( \mu \) and a nuisance parameter \( \kappa \). Again we wish to test \( H_0 : \mu = \mu_0 \) versus \( H_1 : \mu = \mu_1 \). Information about the nuisance parameter comes from a second Poisson measurement \( K \), with mean \( \kappa \). A well-known approach for this case is to condition on the sum \( N + K \). The distribution of \( N \), given a fixed value of \( N + K \), is binomial with parameter \( \mu/(1 + \mu) \). Here one can calculate conditional \( p \)-values \( p_0 \) and \( p_1 \), and plot one against the other.

The above examples rely on a special structure of the problem under study to eliminate nuisance parameters. Unfortunately such a special structure is not always available, and even when it is, it does not guarantee that a \((p_0, p_1)\) plot can be constructed. Here we offer a couple of suggestions for handling the general case. The first one is to use parametric bootstrap techniques to eliminate the nuisance parameters. To first order these techniques consist in substituting an estimate for the unknown nuisance parameter values. The resulting \( p \)-values are generally no longer uniform under their respective null hypothesis, but there exist higher-order refinements that restore some of that uniformity [19]. Bootstrap computations can quickly become rather intensive, but they have the advantage of being frequentist and therefore preserving the error structure of the tests discussed in section 4. As for the likelihood ratios, they can be replaced by profile likelihood ratios. Although the latter are not genuine likelihood ratios, with some caveats they can still be treated as representing statistical evidence in large samples [11].

Our second suggestion is to apply Bayesian methods on the nuisance parameters. Effectively, this amounts to replacing composite hypotheses by simple ones, by integrating out the nuisance parameters over an appropriate proper prior. Suppose for example that the probability density of the data \( x \) under \( H_i \) is given by \( f(x \mid \mu_i, \nu) \), with \( \nu \) a vector of nuisance parameters with prior \( \pi(\nu) \). Then we simply replace \( f \) by

\[
 f^*(x \mid \mu_i) = \int f(x \mid \mu_i, \nu) \pi(\nu) d\nu \quad (46)
\]
in the formulation of the hypotheses. The $p$-values become *prior-predictive* $p$-values:

$$p_i^* = \int_{x_0}^{\infty} f^*(x | \mu_i) \, dx = \int_{x_0}^{\infty} \int f(x | \mu_i, \nu) \pi(\nu) \, d\nu \, dx$$

$$= \int \left[ \int_{x_0}^{\infty} f(x | \mu_i, \nu) \, dx \right] \pi(\nu) \, d\nu = \int p_i(\nu) \pi(\nu) \, d\nu \tag{47}$$

(compare equation (36)), and the likelihood ratios become Bayes factors:

$$\lambda_{01} = \frac{f^*(x_0 | \mu_0)}{f^*(x_0 | \mu_1)} = \frac{\int f(x_0 | \mu_0, \nu) \pi(\nu) \, d\nu}{\int f(x_0 | \mu_1, \nu) \pi(\nu) \, d\nu}. \tag{48}$$

Although this approach lacks the frequentist error interpretation of the tests, it still enjoys the evidential interpretation of the $p$-values and Bayes factors. It is also conceptually simpler and more elegant, as well as computationally much easier, than the bootstrap.

8 Conclusion

We find that ($p_0, p_1$) plots such as figs. 5, 10 and 11 provide useful insights into several diverse statistical aspects of searches for new physics:

- The $CL_s$ criterion for excluding $H_1$;
- The Punzi definition of sensitivity;
- The relationship between $p$-values and likelihoods;
- The difference between the exclusion regions using $p$-values and likelihoods;
- The probabilities of misleading evidence;
- The Jeffreys-Lindley paradox.

In addition, we believe that these plots could be helpful in summarizing the results of such searches. When these involve many channels, with possibly different sensitivities, one could plot the results as points on a ($p_0, p_1$) plot, together with Gaussian likelihood-ratio contours (since the latter are large-sample limits of the actual data pdf’s). This would provide a convenient graphical overview of both the $p$-value and the likelihood-ratio evidence contained in the ensemble of channels investigated.

9 Acknowledgments

We are very grateful to Sir David Cox for insights into the Jeffreys-Lindley paradox, and to Bob Cousins for interesting discussions about it.
References

[1] L. Lyons, “Raster scan or 2-D approach?,” arXiv:1404.7395 [hep-ex] (2014).

[2] The use of two significances (one under the null and the other under the alternative hypothesis) as a system for assessing evidence is discussed in: Bill Thompson, “The nature of statistical evidence,” Lecture notes in statistics 189, Springer Science+Business Media, LLC, 2007, 152pp.

[3] R. B. D’Agostino and M. A. Stephens (editors), “Goodness-of-fit techniques,” Marcel Dekker, Inc., 1986, 563pp.

[4] M. Williams, “How good are your fits? Unbinned, multivariate goodness-of-fit tests in high energy physics,” JINST 5:P09004 (2010); arXiv:1006.3019 [hep-ex] (2010).

[5] In the statistics literature, testing problems that allow a ‘no-decision region’ were considered, possibly for the first time, in: E. L. Lehmann, “A theory of some multiple decision problems. II,” Ann. Math. Statist. 28, 547 (1957).

[6] A. L. Read, “Presentation of search results: the $CL_s$ technique,” J. Phys. G 28, 2693 (2002).

[7] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, “Asymptotic formulae for likelihood-based tests of new physics,” Eur. Phys. J. C 71, 1554 (2011).

[8] G. Punzi, “Sensitivity of searches for new signals and its optimisation”, arXiv:physics/0308063 [physics.data-an] (2003).

[9] T. Sellke, M. J. Bayarri and J. O. Berger, “Calibration of $p$-values for testing precise null hypotheses,” Amer. Statist. 55, 62 (2001); http://www.stat.duke.edu/~berger/papers/99-13.html.

[10] J. O. Berger, “A comparison of testing methodologies,” in Proceedings of the PHYSTAT-LHC workshop on statistical issues for LHC physics, CERN, Geneva, Switzerland, 27-29 June 2007, edited by H. B. Prosper, L. Lyons, and A. De Roeck, CERN Yellow Report CERN-2008-001 (2008), pg. 8-19.

[11] R. Royall, “On the probability of observing misleading statistical evidence,” with discussion, J. Amer. Statist. Assoc. 95, 760 (2000).

[12] I. J. Good, “Comment on ‘Bayesian interpretation of standard inference statements’ by J. W. Pratt,” J. R. Statist. Soc. B 27, 169 (1965).

[13] S. Berry and K. Viele, “A note on hypothesis testing with random sample sizes and its relationship to Bayes factors,” J. Data Science 6, 75 (2008).

[14] G. E. P. Box, “Sampling and Bayes’ inference in scientific modelling and robustness [with discussion],” J. R. Statist. Soc. A 143, 383 (1980).
[15] R. D. Cousins, “The Jeffreys-Lindley paradox and discovery criteria in high energy physics,” arXiv:1310.3791v4 [stat.ME], 28 Jun 2014. 18, 20

[16] R. E. Kass and A. E. Raftery, “Bayes factors,” J. Amer. Statist. Assoc. 90, 773 (1995). 18

[17] D. Cox, private communication (2014). 19

[18] L. Demortier, “P Values and Nuisance Parameters,” in Proceedings of the PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics, CERN, Geneva, Switzerland, 27-29 June 2007, edited by H. B. Prosper, L. Lyons, and A. De Roeck, CERN Yellow Report CERN-2008-001 (2008), pg. 23-33. 20, 22

[19] See for example D.A.S. Fraser and J. Rousseau, “Studentization and deriving accurate $p$-values,” Biometrika 95, 1 (2008), and C.J. Lloyd, “Some non-asymptotic properties of parametric bootstrap $p$-values in discrete models,” Electronic J. Statist. 6, 2449 (2012). 22
A The Bayes-CL$_s$ connection

This appendix describes a sufficient condition for CL$_s$ upper limits to agree with Bayesian upper limits.

Let $f(x \mid \mu)$ be a family of probability densities for the random variable $X$, indexed by
the parameter $\mu$, and consider the family of tests:

$$H_0[\mu^*] : \mu = \mu^* \quad \text{versus} \quad H_1[\mu^*] : \mu > \mu^*.$$  \hspace{1cm} (49)

Suppose we observe $X = x_0$. If we have a prior $\pi(\mu)$ for $\mu$ under $H_1[\mu^*]$, the Bayesian evidence in favor of $H_1[\mu^*]$ is simply the marginal probability of $x_0$ under $H_1[\mu^*]$:

$$p(x_0 \mid H_1[\mu^*]) = \int_{\mu^*}^{+\infty} f(x_0 \mid \mu) \pi(\mu) d\mu.$$  \hspace{1cm} (50)

Note that this probability is only correctly normalized if $\pi(\mu)$ is a proper prior. However, the argument that follows remains valid if $\pi(\mu)$ is improper. The $p$-value evidence against $H_0[\mu^*]$, when the alternative is $H_1[\mu^*]$, is:

$$p_0(\mu^*) = \int_{x_0}^{+\infty} f(x \mid \mu^*) dx = 1 - F(x_0 \mid \mu^*),$$  \hspace{1cm} (51)

where $F(x \mid \mu)$ is the cumulative probability distribution of $x$. This $p$-value evidence against $H_0[\mu^*]$ increases as $1 - p_0(\mu^*)$ increases. Assume now that the Bayesian and frequentist evidences are equal for all $\mu^*$ values larger than some prespecified $\mu_0$:

$$p(x_0 \mid H_1[\mu^*]) = 1 - p_0(\mu^*), \quad \text{for all } \mu^* \geq \mu_0,$$  \hspace{1cm} (52)

or:

$$\int_{\mu^*}^{+\infty} f(x_0 \mid \mu) \pi(\mu) d\mu = \int_{-\infty}^{x_0} f(x \mid \mu^*) dx, \quad \text{for all } \mu^* \geq \mu_0.$$  \hspace{1cm} (53)

This condition is sufficient to obtain equality of CL$_s$ and Bayesian upper limits on $\mu$ under $H_1[\mu_0]$. Indeed, the $\gamma$-credibility level upper limit $\mu_U$ on $\mu$ is the solution of:

$$\int_{\mu_0}^{\mu_U} p(\mu \mid x_0, H_1[\mu_0]) d\mu = \gamma,$$  \hspace{1cm} (54)

where the integrand is the posterior density of $\mu$ under $H_1[\mu_0]$:

$$p(\mu \mid x_0, H_1[\mu_0]) = \frac{f(x_0 \mid \mu) \pi(\mu)}{\int_{\mu_0}^{+\infty} f(x_0 \mid \mu') \pi(\mu') d\mu'} = \frac{f(x_0 \mid \mu) \pi(\mu)}{p(x_0 \mid H_1[\mu_0])}. $$  \hspace{1cm} (55)

Substituting equation (55) in (54) leads to:

$$\frac{p(x_0 \mid H_1[\mu_0]) - p(x_0 \mid H_1[\mu_U])}{p(x_0 \mid H_1[\mu_0])} = \gamma,$$  \hspace{1cm} (56)
and using condition (52) yields:

\[ 1 - \frac{1 - p_0(\mu_U)}{1 - p_0(\mu_0)} = \gamma. \] (57)

The quantity \( 1 - p_0(\mu_U) \) is in fact \( p_1(\mu_U) \), the \( p \)-value for testing \( \mu = \mu_U \) when the alternative is \( \mu = \mu_0 \). Hence, equation (57) is equivalent to

\[ \frac{p_1(\mu_U)}{1 - p_0(\mu_0)} = 1 - \gamma, \] (58)

which corresponds to the \( CL_s \) construction of upper limits.

We illustrate this result with two examples of families of distributions that satisfy condition (52). The first one is any family of continuous distributions parametrized by a location parameter:

\[ f(x | \mu) = f(x - \mu). \] (59)

It is straightforward to verify, by integration by substitution, that

\[ \int_{-\infty}^{x} f(x' - \mu) \, dx' = \int_{\mu}^{+\infty} f(x - \mu') \, d\mu', \] (60)

so that condition (52) is indeed satisfied for a flat prior, \( \pi(\mu) = 1 \). The second example is the Poisson family:

\[ f(n | \mu) = \frac{\mu^n}{n!} e^{-\mu}, \] (61)

for which we have:

\[ \sum_{i=0}^{n} \frac{\mu^i}{i!} e^{-\mu} = \int_{\mu}^{+\infty} \frac{t^n}{n!} e^{-t} \, dt, \] (62)

as can be checked by repeated integration by parts of the right-hand side. Although this is a discrete version of condition (52), also with a flat prior, nothing essential changes in the argument leading from (54) to (58).
Figure 1: Probability density functions (pdf’s) of a test statistic $T$ under two hypotheses $H_0$ (solid) and $H_1$ (dot-dashed). In (a) the separation of the pdf’s is small, making the two hypotheses almost indistinguishable. The separation in (b) is such that the data can exclude $H_1$, reject neither $H_0$ nor $H_1$, or reject $H_0$. Given an observed value $t$ of the test statistic $T$, diagram (b) illustrates the definitions of $p_0$ and $p_1$: $p_0$ is the tail area under $H_0$ in the direction of $H_1$, whereas $p_1$ is the tail area under $H_1$ in the direction of $H_0$. The quantities $t_{crit0}$ and $t_{crit1}$ are the critical values of $T$, beyond which the data are considered incompatible with $H_0$ and $H_1$ respectively; the value of $p_0$ at $T = t_{crit0}$ is $\alpha_0$. The $p_1$ value at this value of $T$ is denoted by $\beta_0$. In (c) the pdf’s are far apart and the observed value of $T$ will always reject at least one hypothesis.
Figure 2: Plot of $p_0$ versus $p_1$ with three lines: The vertical line at $p_0 = \alpha_0$ is an example of a cut used for rejecting $H_0$ when $p_0$ is small. Correspondingly, the horizontal line at $p_1 = \alpha_1$ could be used for excluding $H_1$ when $p_1$ is small (less than 10% in the figure). If the observed value of the test statistic $T$ was such that $(p_0, p_1)$ was in the large rectangle, the data would be consistent with both hypotheses, while for the small rectangle near the origin, it would be inconsistent with both. An alternative procedure for excluding $H_1$ is based on $CL_s = p_1 / (1 - p_0)$, and requires $(p_0, p_1)$ to be below the sloping line.
Figure 3: Probability densities (plots a, b, and c) and probabilities (plot d) used to draw fixed-hypothesis contours on figure 4.
Figure 4: Fixed-hypothesis contours in the $p_0$ versus $p_1$ plane for four different choices of pdf: (a) Gauss, (b) Cauchy, (c) Gamma, and (d) Poisson. In each case one is testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ where $\mu$ is the mean of the distribution (or mode, in the case of Cauchy). In the Gauss, Cauchy, and Gamma cases, the contours depend on a simple combination of $\mu_0$ and $\mu_1$, but not on their individual values. The line $p_1 = 1 - p_0$ is an upper boundary for observable $p$-values in plots (a), (b), and (c), but not in the discrete Poisson case of plot (d).
Figure 5: Plot of $p_0$ versus $p_1$ with fixed-hypothesis contour lines, as in figure 4(a). For fixed pdf’s under $H_0$ and $H_1$, the possible values of the test statistic $T$ correspond to a curve in the ($p_0, p_1$) plane. The examples shown are for Gaussian pdf’s where $\Delta \mu/\sigma$ is zero (i.e., identical pdf’s under $H_0$ and $H_1$), 1.67 and 3.33. When $\Delta \mu/\sigma = 3.33$, the separation of the pdf’s is large enough that the data cannot fall in the large no-decision region, defined by $p_0 > \alpha_0$ & $p_1 > \alpha_1$. We refer to the black dot at the intersection of the horizontal dashed and vertical dot-dashed lines as the ‘Punzi point’ (see text). For a given choice of $\alpha_0$, the intersection of the line $p_0 = \alpha_0$ with the relevant contour has ordinate $\beta_0$, the probability of failing to reject $H_0$ when $H_1$ is true (the plot shows this for the $\Delta \mu/\sigma = 1.67$ contour). The relation between $\alpha_1$ and $\beta_1$ is similar.
Figure 6: Solid line: Cauchy probability density function with mode $\mu_c$ and scale parameter $\gamma$ both equal to 1. Dashed line: asymmetric pdf obtained by a one-to-one transformation of the Cauchy density (see equation (14) in the text, with $\mu_c = \gamma = \Delta \mu / \sigma = 1$).
Figure 7: Contours of constant likelihood ratio $\lambda_{01} \equiv L_0/L_1$ in the $p_0$ versus $p_1$ plane for four different choices of pdf: (a) Gauss, (b) Cauchy, (c) Gamma, and (d) Poisson, where the lines merely join up the discrete $(p_0, p_1)$ points as "contours". In each case one is testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ where $\mu$ is the pdf mean for Gauss and Poisson, mode for Cauchy, and rate parameter for Gamma. To facilitate comparison, the same contours are drawn in all four cases. In plot (d), points line up vertically across contours, since by construction $\mu_0$ is the same everywhere.
Figure 8: Likelihood-ratio contours for testing the value of an exponential decay rate for four different values of the number $n$ of decay time measurements included in the test (solid lines), compared with the corresponding contours for a Gaussian test (dashed lines). At large $n$, the exponential contours converge to the Gaussian ones. From left to right in each plot, the contours correspond to $\lambda_{01} = 1/32, 1/8, 8, \text{and } 32$. 
Figure 9: Likelihood-ratio contours for testing the value of a Poisson mean for four different values of the number $n$ of measurements included in the test (dots), compared with the corresponding contours for a Gaussian test (dashed lines). At large $n$, the Poisson contours converge to the Gaussian ones. From left to right in each plot, the contours correspond to $\lambda_0 = 1/32, 1/8, 8, \text{and } 32.$
Figure 10: Plot of $p_0$ versus $p_1$ with likelihood ratio contours (colored, solid lines), when the pdf’s are Gaussians of equal width. The likelihood ratio is unity along the $p_1 = 1 - p_0$ diagonal, where $H_1$ is identical to $H_0$, and along the $p_1 = p_0$ diagonal, where the observed value of the test statistic favors each hypothesis equally. Going down a line of constant $p_0$ from $p_1 = 0.5$ to $p_1 = 0$ corresponds to increasing separation of the pdf’s (e.g., more and more data), and also to increasing $L_0/L_1$. This gives rise to the possibility of $p_0$ being small while $L_0/L_1$ is large. In fact one could exclude $H_0$ or $H_1$ based on the observed likelihood ratio instead of $p_0$ or $p_1$; the corresponding exclusion regions have a different shape from the $p$-value exclusion regions. The dashed line is a contour of constant $\Delta \mu/\sigma$. Its intersections with likelihood-ratio contours provides various probabilities of misleading evidence (see text).
Figure 11: Log-log version of the plot of $p_0$ versus $p_1$, assuming that the pdf’s of the test statistic under $H_0$ and $H_1$ are Gaussian with the same width. Solid lines show the likelihood-ratio contours for $L_0/L_1 = 1/32$, 1/8, 1, 8 and 32, and dashed lines show the fixed-hypothesis contours for $\Delta \mu/\sigma = 1.67$ and 3.33. The $p_0$ coordinates of points $a$, $b$, $c$ and $d$, and the $p_1$ coordinates of points $A$, $B$, $C$ and $D$ yield the probabilities of misleading evidence listed in the table below.

| $\Delta \mu/\sigma$ | 1.67 | 3.33 |
|----------------------|------|------|
| $\mathbb{P}(L_0/L_1 < 1/32 \mid H_0)$ | 0.18% | 0.34% |
| $\mathbb{P}(L_0/L_1 < 1/8 \mid H_0)$ | 1.9% | 1.1% |
| $\mathbb{P}(L_0/L_1 > 8 \mid H_1)$ | 1.9% | 1.1% |
| $\mathbb{P}(L_0/L_1 > 32 \mid H_1)$ | 0.18% | 0.34% |
Figure 12: Four examples of sequential testing on the mean of a Gaussian distribution with unit width. Plots (a) and (c) assume that $H_0$ is true, whereas plots (b) and (d) assume the truth of $H_1$. A sequential testing procedure (see text) describes a random walk in the $(p_0, p_1)$ plane (shown by the black broken lines). The blue curves represent the boundary defined by the law of the iterated logarithm (LIL). The red likelihood-ratio contours (for $\lambda_{01} = 1/8$ in plots (a) and (c), and for $\lambda_{01} = 8$ in plots (b) and (d)) are examples of decision boundaries that avoid the possibility of testing to a foregone conclusion implied by the LIL. The green line in plots (b) and (d) represents the $CL_s = 5\%$ decision boundary, which does not avoid this possibility.
Figure 13: Plot showing fifty fixed-hypothesis contours (green curves) crossed by a random walk (black broken line) associated with a sequential test procedure of the Gauss($\mu_0, \sigma$) versus Gauss($\mu_1, \sigma$) type. At each step the sample size $n$ increases by one, and the walk moves to a contour with improved resolution. Contours are labeled by the value of $n$. The blue curve shows the relationship between $p_0$ and $n$ described by the LIL boundary. The red dotted line represents a fixed discovery threshold $\alpha_0$. Since this line crosses to the large-$p_0$ side of the LIL boundary, it is guaranteed to have the pathology of sampling to a foregone conclusion. In contrast, with the $p_0$ cutoff set as $\alpha_0/\sqrt{n}$ (solid red line), repeated sampling does not necessarily lead to exclusion of a true $H_0$. 
Figure 14: Insight for the Jeffreys–Lindley paradox. The likelihood ratio contours in (a) are those of figure 7(a) for comparing hypotheses whose pdf’s are equal-width Gaussians. The line $bcde$ is at fixed $p_0$, with the points $b$ to $e$ corresponding to increasing separation of the $H_0$ and $H_1$ pdf’s, as shown in diagrams (b) to (e) respectively.
Figure 15: Plot of $p_0$ versus prior-predictive $p_{1pp}$ for testing $H_0: \mu = \mu_0$ versus $H_1: \mu > \mu_0$. The test statistic has a Gaussian distribution with mean $\mu$ and standard deviation $\sigma$. The prior for $\mu$ under $H_1$ equals $1/\tau$ for $\mu_0 < \mu \leq \mu_0 + \tau$ and is zero otherwise. Fixed-hypothesis contours (dashed lines) are labeled by the value of $\tau/\sigma$. Constant Bayes factor contours (colored solid lines) are also shown.
Figure 16: Log-log version of figure 15, with a larger range of fixed-hypothesis contours and more realistic values for the Bayes factor contours. Also shown are the constant $p_0$ threshold (dotted line) at $\alpha_0 = 1\%$ and the corresponding $n$-dependent threshold $\alpha_0/\sqrt{n}$ (dot-dashed line), where $n$ is the sample size.
Figure 17: Plots illustrating what happens to the Jeffreys-Lindley paradox when $H_0$ is an interval hypothesis with width $\epsilon$ instead of a point-null hypothesis. Both $p_{opp}$ and $p_{1pp}$ are prior-predictive $p$-values. The contours on these plots are the same as in figure 16, although for clarity only the contours $\tau/\sigma = 10000$ and $B_{01} = 1$ are labeled. Compared with figure 16, the fixed-hypothesis contours and the constant Bayes factor contours are both changed in such a way that the paradox remains present regardless of the value of $\epsilon/\sigma$. 
Figure 18: Plots illustrating what happens to the Jeffreys-Lindley paradox when $H_0$ is an interval hypothesis with width $\epsilon$ instead of a point-null hypothesis. The $p$-value $p_{1\text{pp}}$ is a prior-predictive $p$-value, whereas $p_{0\text{sup}}$ is a supremum $p$-value. The contours on these plots are the same as in figure 16, although for clarity only the contours $\tau/\sigma = 10000$ and $B_{01} = 1$ are labeled. Compared with figure 16, only the constant Bayes factor contours are changed; the fixed-hypothesis contours are the same. The result is that the paradox disappears for a suitably high value of $\epsilon/\sigma$. 
Figure 19: Plot of the integrated pdf \((44)\) for several values of the parameter \(\theta\) (the pdf for \(\theta = 100\) has been truncated at the upper end). If for example \(x = 2\) is observed, the \(p\)-value under \(H_0: \theta = 0\) is 2.3% (shaded area), but the likelihood ratio of \(\theta = 0\) to \(\theta = 100\) is 5.5. It is clear that for very large \(\theta\) values, significantly small \(p\)-values that disfavor \(H_0\) will be associated with likelihood ratios that favor \(H_0\). This is a simple versus simple version of the Jeffreys-Lindley paradox.