Notes on Statistics for HEP

Raghav Kansal

January 7, 2025

Abstract

This is a series of walk-throughs of some concepts in statistics relevant for high energy physics (HEP). It is primarily based on (me trying to understand) Refs. [1] and [2], and the CMS combine tool [3]. Both references are useful and highly recommended for reading; the emphasis here is heavily on **implementing and visualizing what they discuss in Python** — as well as providing some heuristic derivations of some of their formulae — as that is what personally helped me build an intuition for these concepts.

These notes are primarily intended to be followed interactively online as Jupyter notebooks, but are also available as a PDF. Feedback and corrections are very welcome through e-mail or as issues on Github. Finally, for those interested, some more notes and tutorials I have written up, such as on the standard model and machine learning in HEP, are available here.

Contents

Chapter 1

Introduction

Once data is collected by high energy physics (HEP) experiments and reconstructed offline, it is analyzed to search and measure processes of interest. Typically, the raw data is entirely dominated by irrelevant background processes which we want to filter out in favor of the signal. The first step towards this is through appropriate online triggers, followed by offline selections to isolate the signal.

Optimizing the event selection for all but a handful of data-driven searches requires simulations of the signal and background processes. Additionally, once the selections and phase space in which to perform the measurement have been finalized, the expected signal and background yields have to be carefully estimated, which often again necessitates simulations, as well as data-driven methods via unbiased control regions.

Once we have our observations, and signal and background estimates, the final critical step is to interpret the results in a robust statistical framework. At the LHC,

this is typically done using a frequentist, likelihood-based approach. In these notes, this approach is introduced by way of simple experimental examples.

These notes are organized as follows. Chapter 2 introduces the concepts of the likelihood functions and test statistics, with Chapter 3 discussing the framework for hypothesis testing, including p-values, significances, and the statistical definition of a "discovery". Chapters 4 and 5 then describe frequentist confidence intervals and upper limits, and the important concepts of expected significances and limits, respectively. Finally, asymptotic approximations to simplify these computations are discussed in Part II.

Part I

Frequentist statistics at the LHC

Chapter 2

The likelihood function and test statistics

2.1 The data model

Let us take the simplest possible case of a (one bin) counting experiment, where in our "signal region" we expect s signal events and b background events. The probability to observe n events in our signal region is distributed as a Poisson with mean $s + b$:

$$P(n; s, b) = \mathrm{Pois}(n; s + b) = \frac{(s + b)^n e^{-(s+b)}}{n!} \tag{2.1}$$

Since we have only one observation but two free parameters, this experiment is underconstrained. So, let's also add a "control region" where we expect no signal and b background events. The probability of observing m events in our control region is

therefore:

$$\text{Pois}(m; b) = \frac{b^m e^{-b}}{m!} \tag{2.2}$$

Combining the two, the joint probability distribution for n and m is:

$$P(n, m; s, b) = \text{Pois}(n; s + b) \cdot \text{Pois}(m; b) = \frac{(s + b)^n e^{-(s+b)}}{n!} \cdot \frac{b^m e^{-b}}{m!} \tag{2.3}$$

This is also called the model for the data and is plotted for sample s, b values in Figure 2.1.

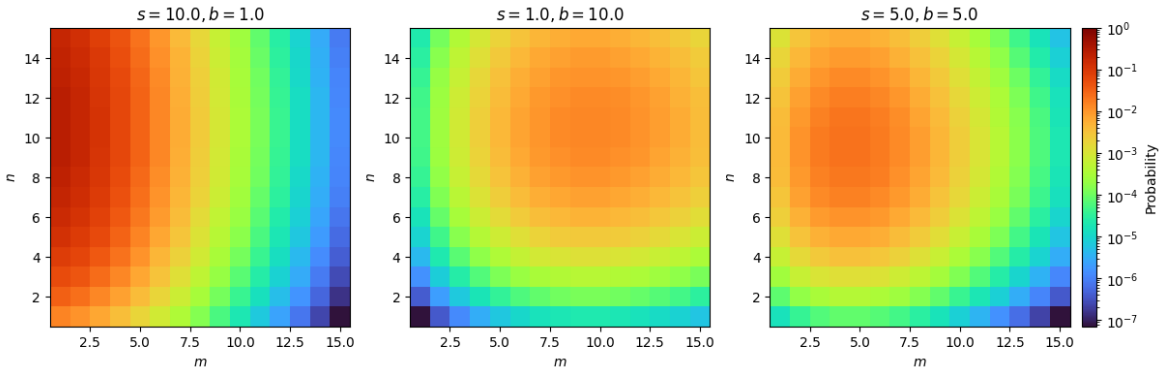

Figure 2.1: Sample 2D Poisson distributions.

2.2 The likelihood function

In the *frequentist* philosophy, however, all our parameters n, m etc. are simply fixed values of nature and, hence, don't have a probability distribution. Instead, we work with the *likelihood function*, which is a function only of our parameters of interest (POIs), s in our example, and "nuisance parameters" (b), given fixed values for n and m:

$$L(s, b) = P(n, m; s, b) = \frac{(s + b)^n e^{-(s+b)}}{n!} \cdot \frac{b^m e^{-b}}{m!}. \tag{2.4}$$

Importantly, this is not a probability distribution on s and b! To derive that, we would have to use Bayes' rule to go from $P(n, m; s, b) \rightarrow P(s, b; n, m)$; however, such probability distributions don't make sense in our frequentist world view, so we're stuck with this likelihood formulation. Often, it's more convenient to consider the negative log-likelihood:

$$-\ln L = \ln n! + \ln m! + s + 2b - n\ln(s+b) - m\ln b \tag{2.5}$$

2.3 The profile likelihood ratio

Fundamentally, the goal of any experiment is to test the compatibility of the observed data (n, m here) with a certain hypothesis H. We do this by mapping the data to a "test statistic" t, which is just a number, and comparing it against its distribution under H, $P(t|H)$. Our problem, thus, boils down to 1) choosing the most effective t for testing H, and 2) obtaining $P(t|H)$.

In the case of testing a particular signal strength, we use the "profile likelihood ratio":

$$\lambda(s) = \frac{L(s, \hat{\hat{b}}(s))}{L(\hat{s}, \hat{b})}, \tag{2.6}$$

where $\hat{s}, \hat{b}$ are the maximum-likelihood estimates (MLEs) for s and b, given the observations n, m, and $\hat{\hat{b}}(s)$ is the MLE for b given n, m, *and* s. The MLE for a parameter is simply the value of it for which the likelihood is maximized, and will be discussed in the next section. The numerator of $\lambda(s)$ can be thought of as a way to "marginalize" over the nuisance parameters by simply values that maximize the likelihood for any given s, while the denominator is effectively a normalization factor, such that $\lambda(s) \leq 1$.

Again, it's often more convenient to use the (negative) logarithm:

$$t_s = -2 \ln \lambda(s) \tag{2.7}$$

Note that $\mathrm{Max}[\lambda(s)] = 1 \Rightarrow \mathrm{Min}[t_s] = 0$. $\lambda(s)$ and t_s are plotted for sample n, m values with $n - m = 10$ in Figure 2.2. The maximum (minimum) of the profile likelihood ratio (t_s) is at $s = n - m = 10$, as we expect; however, as the ratio between n and m decreases — i.e., the experiment becomes more noisy — the distributions broaden, representing the reduction in sensitivity, or the higher uncertainty on the true value of s.

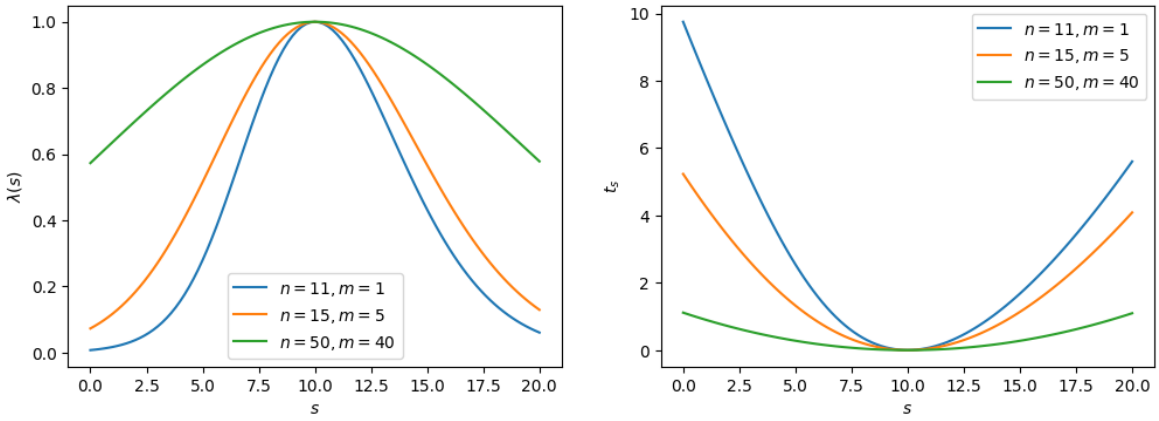

Figure 2.2: The profile likelihood ratio $\lambda(s)$ (left) and the t_s test statistic (right) for our one-bin Poisson model.

Note that the likelihood ratio and t_s are also broadened due to the nuisance parameter; i.e., because we are missing information about b. This can be demonstrated by plotting them with $b = m$, emulating perfect information of b (Figure 2.3), and indeed, we see the functions are narrower than in Figure 2.2. More generally, increasing (decreasing) the uncertainties on the nuisance parameters will broaden (narrow) the test statistic distribution. This is which is why experimentally we want to constrain them through auxiliary measurements as much as possible.

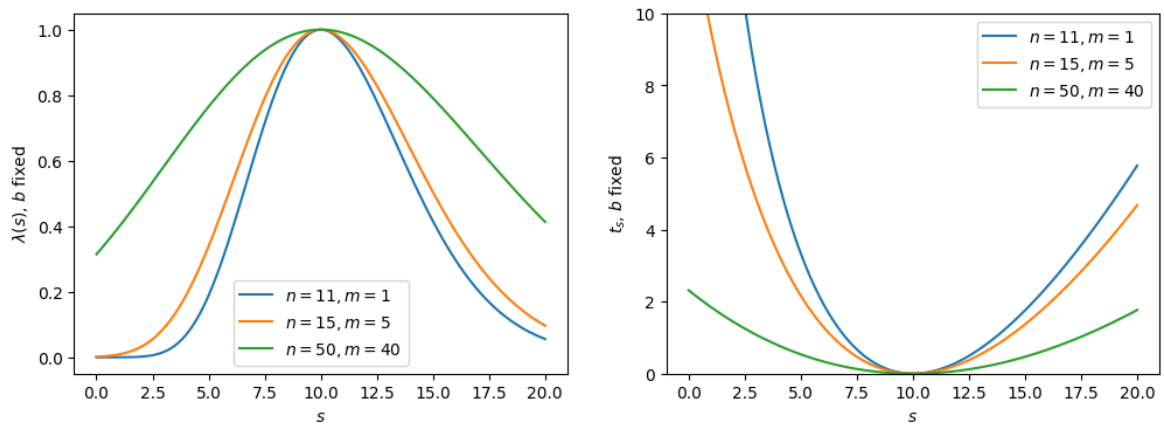

Figure 2.3: The profile likelihood ratio $\lambda(s)$ (left) and the t_s test statistic (right) with $b = m$, demonstrating the effect of decreasing uncertainties on our nuisance parameters.

2.4 Maximum-likelihood estimates

MLEs for s and b can be found for this example by setting the derivative of the negative log-likelihood to 0 (more generally, this would require numerical minimization):

$$\frac{\partial(-\ln L)}{\partial s} = 1 - \frac{n}{s+b} = 0 \tag{2.8}$$

$$\frac{\partial(-\ln L)}{\partial b} = 2 - \frac{n}{s+b} - \frac{m}{b} = 0 \tag{2.9}$$

Solving simultaneously yields, as you might expect:

$$\hat{b} = m, \hat{s} = n - m, \tag{2.10}$$

Or just for $\hat{\hat{b}}(s)$ from Eq. 2.8:

$$2b^2 + (2s - n - m)b - ms = 0 \tag{2.11}$$

Plugging this back in, we can get $\lambda(s)$ and t_s for any given s.

2.5 Alternative test statistic

So far, our construction allows for $s < 0$; however, physically the number of signal events can't be negative. Rather than incorporating this constraint in the model, it's more convenient to impose this in the test statistic, by defining:

$$\tilde{\lambda}(s) = \begin{cases} \frac{L(s,\hat{\hat{b}}(s))}{L(\hat{s},\hat{b})}, & \hat{s} \geq 0. \\ \frac{L(s,\hat{\hat{b}}(s))}{L(\hat{0},\hat{\hat{b}}(0))}, & \hat{s} < 0. \end{cases}, \tag{2.12}$$

and

$$\tilde{t}_s = -2 \ln \tilde{\lambda}(s) \tag{2.13}$$

The difference between the nominal and alternative test statistics is highlighted in Figure 2.4. For $n < m$, the $\tilde{\lambda}(s) = 1$ and $\tilde{t}_s = 0$ values are at $s = 0$, since physically that is what fits best with our data (even though the math says otherwise).

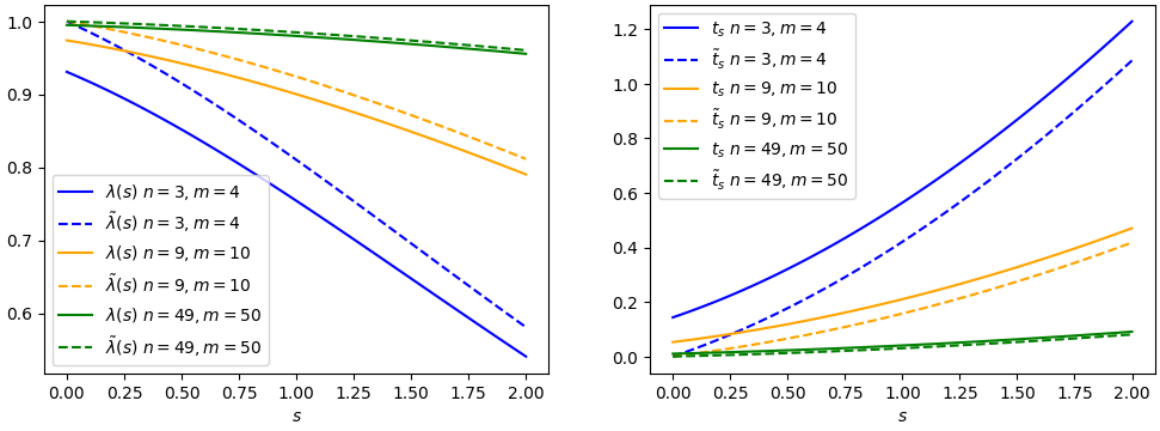

Figure 2.4: Comparing the nominal vs alternative test statistic.

Next, we want to translate this to a probability distribution of $\tilde{t}_s$ under a particular signal hypothesis (H_s) (i.e., an assumed value of s): $p(\tilde{t}_s|H_s)$, or just $p(\tilde{t}_s|s)$ for

simplicity.

Chapter 3

Hypothesis testing

The goal of any experiment is to test whether our data support or exclude a particular hypothesis H, and quantify the (dis)agreement. For example, to what degree did our search for the Higgs boson agree or disagree with the standard model hypothesis?

We have already discussed the process of mapping data to a scalar test statistic t that we can use to test H. However, we need to know the probability distribution of t under H to quantify the (in)consistency of the observed data with H and decide whether or not to exclude H.

We must also recognize that there's always a chance that we will exclude H even if it's true (called a Type I error, or a false positive), or not exclude H when it's false (Type II error, or false negative). The probability of each is referred to as α and β, respectively. This is summarized handily in Table 3.1.

Before the test, we should decide on a probability of making a Type I error, α, that we are comfortable with, called a "significance level". Typical values are 5% and 1%,

Table 3.1: Table of error types, reproduced from Ref. [4].

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Fail to reject	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

although if we're claiming something crazy like a new particle, we better be very sure this isn't a false positive; hence, we set a much lower significance level for these tests of 3×10^{-7}. (The *significance* of this value will be explained in Section 3.2 below.)

3.1 Deriving $p(\tilde{t}_s|s)$

We can approximate $p(\tilde{t}_s|s)$ by generating several pseudo- or "toy" datasets assuming s expected signal events. In this case, this means sampling possible values for n, m from our probability model. We will continue with our simple counting experiment (Section 2.1), for which such toy datasets are generated and then used to create histograms for $p(\tilde{t}_s|s)$ in Figure 3.1. Note that one complication in generating these toys is that the n, m distributions from which we want to sample (Eq. 2.3) also depend on the nuisance parameter b. However, we see from the figure that this does not matter as much as we might expect.

We make two important observations:

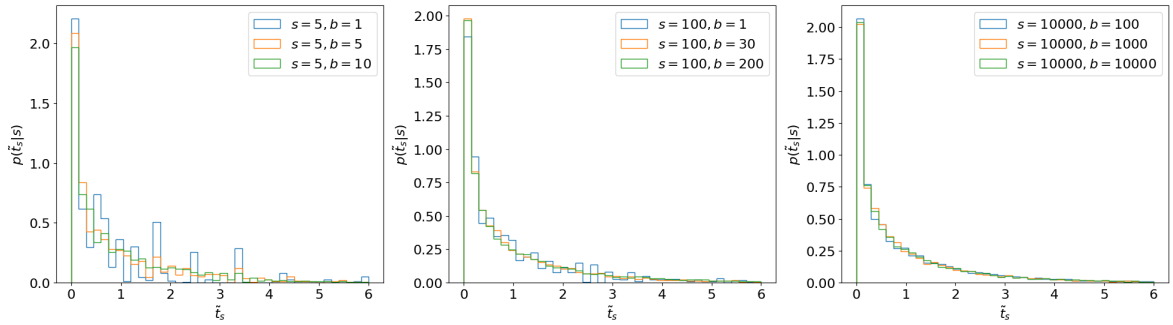

Figure 3.1: Estimating $p(\tilde{t}_s|s)$ through toys.

1. $p(\tilde{t}_s|s)$ does not depend on nuisance parameters as long as we have sufficiently large statistics (in this case, when b is sufficiently large). This is a key reason for basing our test statistic on the profile likelihood.
2. In fact, $p(\tilde{t}_s|s)$ doesn't even depend on the POI s! (Again, as long as s is large.)

Reference [1] shows that, asymptotically, this distribution follows a χ^2 distribution with degrees of freedom equal to the number of POIs, as illustrated in Figure 3.2.[1] We can see that the asymptotic form looks accurate even for s, b as low as ~ 5. Note that for cases where we can't use the asymptotic form, Ref. [2] recommends using $b = \hat{\hat{b}}(s)$ when generating toys, so that we (approximately) maximize the agreement with the hypothesis.

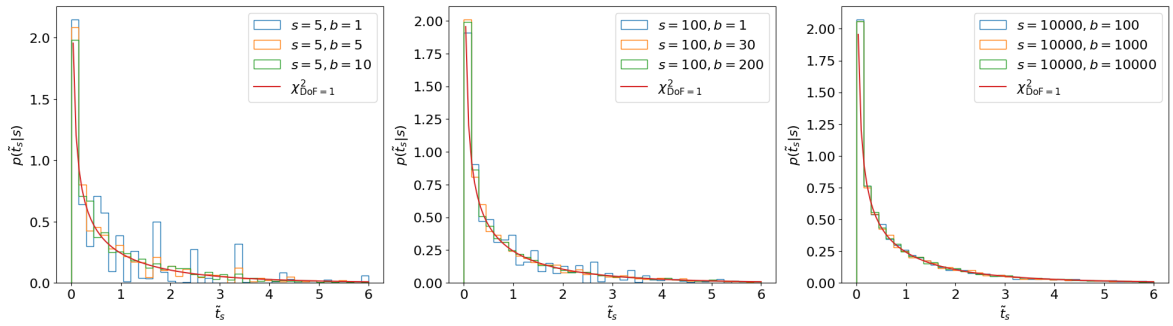

Figure 3.2: Asymptotic form of $p(\tilde{t}_s|s)$.

[1] One can find the derivation in the reference therein; essentially, like with most things in physics, this follows from Taylor expanding around the minimum...

3.2 p-values and significance

Now that we know the distribution of the test statistic $p(\tilde{t}_s|H_s) \equiv p(\tilde{t}_s|s)$, we can finally test H_s with our experiment. We just need to calculate the "observed" test statistic $\tilde{t}_s^{\text{obs}}$ from our observations, and compare it to the $p(\tilde{t}_s|s)$.

Example 3.1. Let's say we're testing the hypothesis of $s = 10$ signal events in our model and we observe $n = 20, m = 5$ events. We can map this observation to our test statistic $\tilde{t}_s^{\text{obs}}(s = 10, n_{\text{obs}} = 20, m_{\text{obs}} = 5) = 1.07$, and see where this falls in our $p(\tilde{t}_s|s)$ distribution (Figure 3.3).

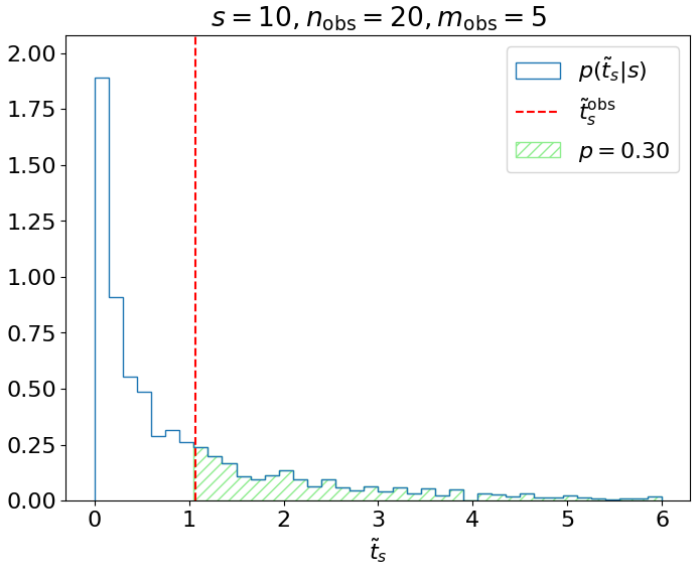

Figure 3.3: Testing H_s in Example 3.1.

Ultimately, we care about, given $p(\tilde{t}_s|s)$, the probability of obtaining $\tilde{t}_s^{\text{obs}}$ *or a value more inconsistent* with H_s; i.e., the green shaded region above. This is referred to as the *p-value* of the observation:

$$p_s = \int_{\tilde{t}_{\text{obs}}}^{\infty} p(\tilde{t}_s|s)\mathrm{d}\tilde{t}_s = 1 - F(\tilde{t}_{\text{obs}}|s), \tag{3.1}$$

which is 0.30 for this example, where

$$F(\tilde{t}_s|s) = \int_{-\infty}^{\tilde{t}_s} p(\tilde{t}'_s|s)\mathrm{d}\tilde{t}'_s \tag{3.2}$$

is the cumulative distribution function (CDF) of $\tilde{t}_s$. We reject the hypothesis if this p-value is less than our chosen significance level α; the idea being that if H_s were true and we repeated this measurement many times, then the probability of a false-positive (p-value $\leq \alpha$) is exactly α, as we intended.

The p-value is typically converted into a *significance* (Z), which is the corresponding number of standard deviations away from the mean in a Gaussian distribution:

$$Z = \Phi^{-1}(1 - p), \tag{3.3}$$

where Φ is the CDF of the standard Gaussian. This is more easily illustrated in Figure 3.4, where φ is the standard Gaussian distribution:

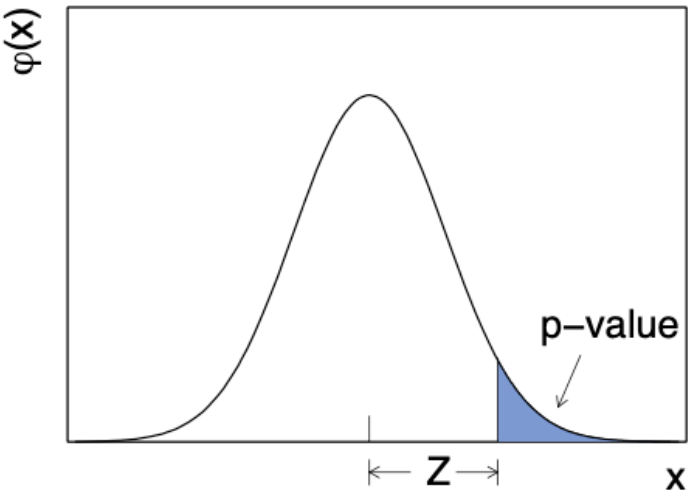

Figure 3.4: Relationship between significance Z and the p-value, reproduced from Ref. [2].

The significance in Example 3.1 is, therefore, $\Phi^{-1}(1 - 0.30) = 0.53$. We sometimes

say that our measurement is (in)consistent or (in)compatible with H at the 0.53σ level, or within 1σ, etc.

3.3 Signal discovery

So far, we have been testing the signal hypothesis, but usually when searching for a particle, we instead test the "background-only" hypothesis H_0 and decide whether or not to reject it. This means we want $\tilde{t}_0^{\text{obs}}$ and $p(\tilde{t}_0|0)$ (Figure 3.5).[2]

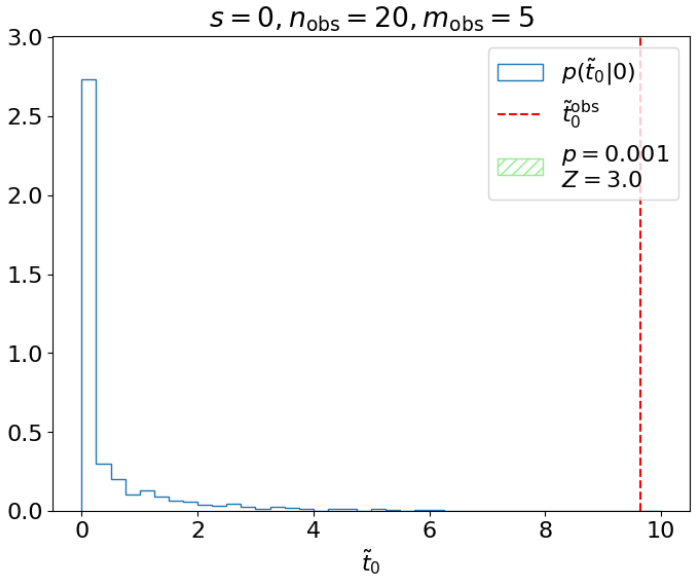

Figure 3.5: Testing the background-only hypothesis in Example 3.1.

We could say for this experiment, therefore, that we exclude the background-only hypothesis at the "3 sigma" level. However, for an actual search for a new particle at the LHC, this is insufficient to claim a discovery, as the probability of a false positive at 3σ, $1/1000$, is too high. The standard is instead set at 5σ for discovering new signals, corresponding to the 3×10^{-7} significance level quoted earlier, as we really don't want

[2]Ref. [1] refers to the special case of the test statistic $\tilde{t}_s$ for $s = 0$ as q_0.

to be making a mistake if we're claiming to have discovered a new particle! 3σ, 4σ, and 5σ are commonly referred to as "evidence", "observation", and "discovery", respectively, of the signals we're searching for.

In summary, the framework for hypothesis testing comprises:

1. Defining a test statistic t to map data $\mathbf{x}$ (in our example, $\mathbf{x} = (n, m)$) to a single number.
2. Deriving the distribution of t under the hypothesis being tested $p(t|H)$ by sampling from "toy" datasets assuming H.
3. Quantifying the compatibility of the observed data $\mathbf{x}_{\text{obs}}$ with H with the p-value or significance Z of t_{obs} relative to $p(t|H)$.

This p-value / significance is what we then use to decide whether or not to exclude H. A particularly important special case of this, as discussed above, is testing the background-only hypothesis when trying to discover a signal.

Chapter 4

Confidence intervals and limits

4.1 Confidence intervals using the Neyman construction

Next, we discuss going beyond hypothesis testing to setting intervals and limits for parameters of interest. The machinery from Section 3 can be extended straightforwardly to extracting "confidence intervals" for our parameters of interest (POIs): a range of values of the POIs that are allowed, based on the experiment, at a certain "confidence level" (CL), e.g. 68% or 95%. Very similar to the idea of the significance level, the CL is defined such that if we were to repeat the experiment many times, a 95%-confidence-interval must contain, or *cover*, the true value of the parameter 95% of the time.

This can be ensured for any given CL by solving Eq. 3.1 for a p-value of $1 - \mathrm{CL}$:

$$p = 1 - \mathrm{CL} = \int_{\tilde{t}_s^{\mathrm{obs}}}^{\infty} p(\tilde{t}_s|s_{\pm})\mathrm{d}\tilde{t}_s, \tag{4.1}$$

where s_- and s_+ are the lower and upper limits on s, respectively.

This can be solved by scanning s and finding the values of s for which the RHS $= 1 - \mathrm{CL}$, as demonstrated in Figure 4.1 for the experiment in Example 3.1 ($n_{\mathrm{obs}} = 20, m_{\mathrm{obs}} = 5$). This procedure of inverting the hypothesis test by scanning along the values of the POIs is called the "Neyman construction".

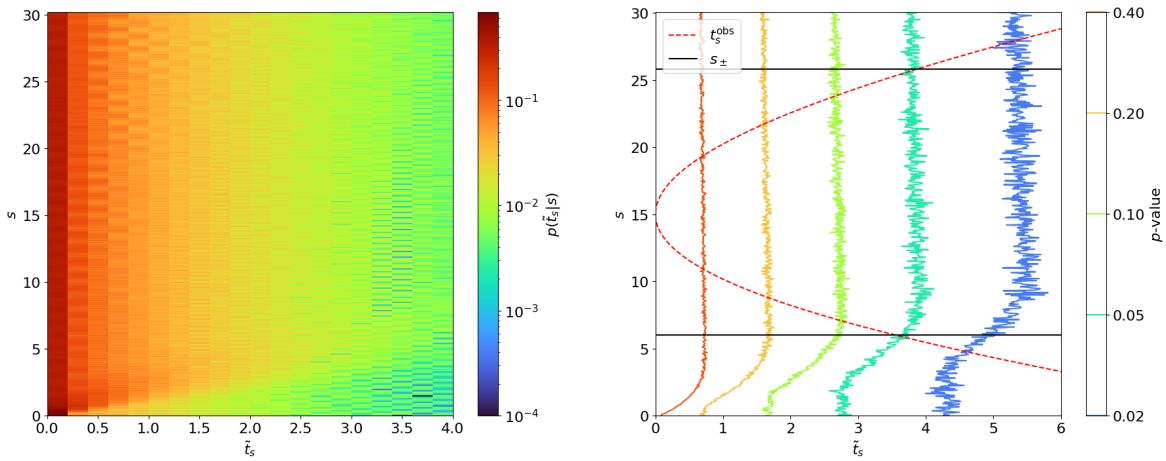

Figure 4.1: Demonstration of the Neyman construction for a 95% confidence interval for the experiment in Example 3.1 ($n_{\mathrm{obs}} = 20, m_{\mathrm{obs}} = 5$). Left: Scanning $p(\tilde{t}_s|s)$ using 10,000 toys each for different values of s. Right: Converting this to a contour plot of the p-values for different $\tilde{t}_s$'s as a function of s, with the observed t_s^{obs} in red. The points at which t_s^{obs} intersects with the p-value = 0.05 contour are marked in black and signify the limits of the 95% confidence interval for s - in this case, $[6.0, 25.8]$.

One subtlety to remember is that, in principle, we should also be scanning over the nuisance parameters (b) when estimating the p-values. However, this would be very computationally expensive so in practice, we continue to use $b = \hat{\hat{b}}(s)$, to always (approximately) maximize the agreement with the hypothesis. Ref. [2] calls this trick "profile construction".

4.2 Upper limits

Typically if a search does not have enough sensitivity to directly observe a new signal, we instead quote an upper limit on the signal strength. This is similar in practice to the Neyman construction for confidence intervals, solving Eq. 4.1 only for the upper boundary. However, an important difference is that when setting upper limits, we have to modify the test statistic so that a best-fit signal strength *greater* than the expected signal ($\hat{s} > s$) does not lower the compatibility with H_s:

$$\tilde{q}(s) = \begin{cases} \tilde{t}(s), & \hat{s} < s. \\ 0, & \hat{s} \geq s. \end{cases} = \begin{cases} -2\ln\tilde{\lambda}(s), & \hat{s} < s. \\ 0, & \hat{s} \geq s. \end{cases} = \begin{cases} -2\ln\frac{L(s,\hat{\hat{b}}(s))}{L(0,\hat{\hat{b}}(0))}, & \hat{s} < 0. \\ -2\ln\frac{L(s,\hat{\hat{b}}(s))}{L(\hat{s},\hat{b})}, & 0 \leq \hat{s} < s. \\ 0, & \hat{s} \geq s. \end{cases} \quad (4.2)$$

The upper limit test statistic $\tilde{q}(s)$ is set to 0 for $\hat{s} > s$ so that this situation does not contribute to the p-value integral in Eq. 3.1. Figure 4.2 demonstrates this, and the difference between $\tilde{t}_s$ and $\tilde{q}_s$, for different sample observations.

Note that (as one may expect from Figure 4.2) the distribution $p(\tilde{q}_s|s)$ no longer behaves like a standard χ^2 but, instead, as a "half-χ^2". This is essentially a χ^2 plus a delta function at 0 (since, under the signal hypothesis, on average there will be an over-fluctuation half the time, for which $\tilde{q}_s = 0$), as shown in Figure 4.3.

We can now revisit Example 3.1 to set an upper limit on s rather than a confidence interval (Figure 4.4). $p(\tilde{q}_s|s)$ is shifted to the left with respect to $p(\tilde{t}_s|s)$; hence, the upper limit of 24 is slightly lower than the upper bound of the 95% confidence interval we derived using $\tilde{t}_s$.

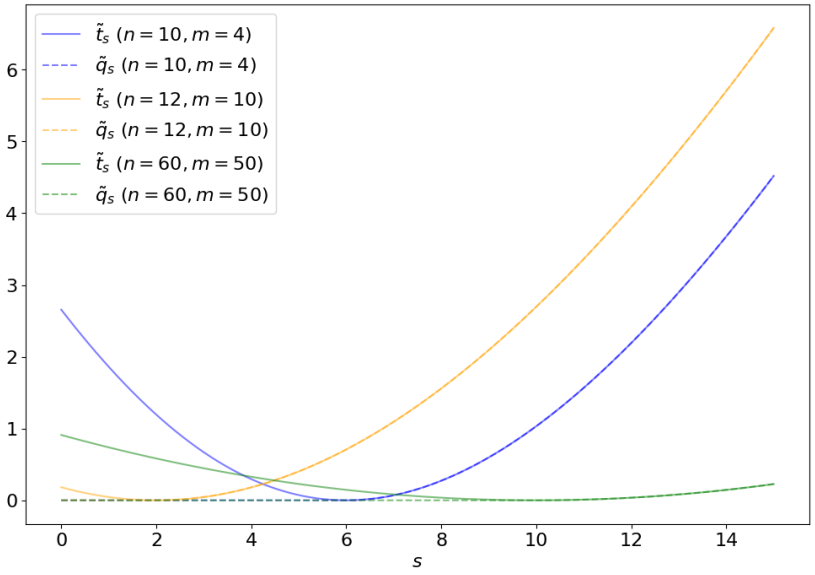

Figure 4.2: Comparing $\tilde{t}_s$ and $\tilde{q}_s$.

4.3 The CL_s criterion

We now introduce two conventions related to hypothesis testing and searches in particle physics. Firstly (the simple one), the POI s is usually re-parametrized as $s \to \mu \cdot s$, where μ is now considered the POI, referred to as the "signal strength", and s is a fixed value representing the number of signal events we expect to see for

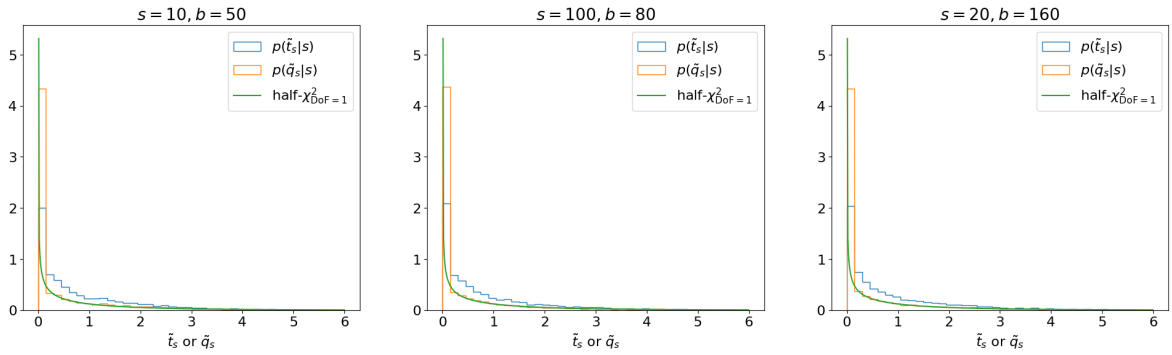

Figure 4.3: Comparing $p(\tilde{t}_s|s)$ and $p(\tilde{q}_s|s)$. $p(\tilde{q}_s|s)$ asymptotically follows a half-χ^2 distribution (green).

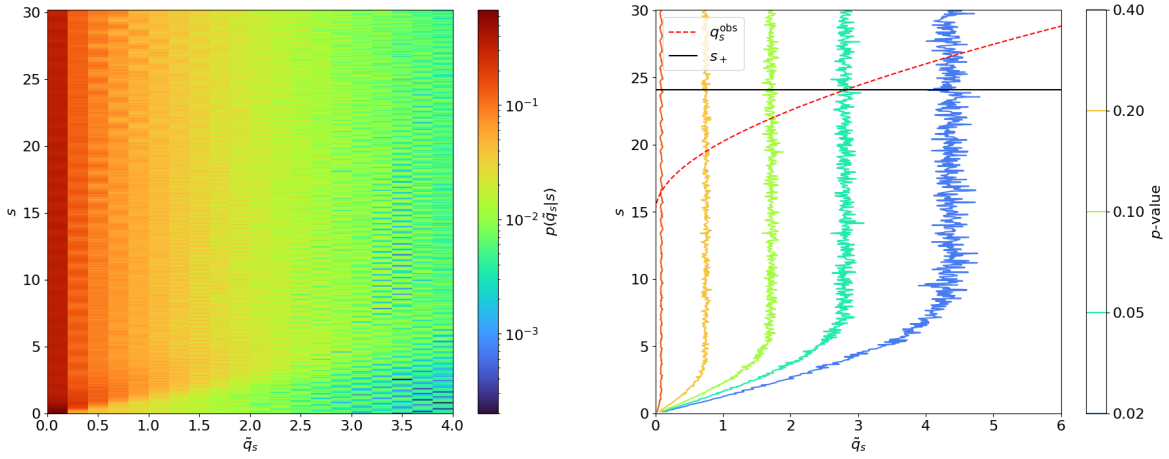

Figure 4.4: Extending the Neyman construction to an upper limit on s. Left: Scanning the upper limit test statistic distribution $p(\tilde{q}_s|s)$ using 10,000 toys each for different values of s. Right: Converting this to a contour plot of the p-values for different $\tilde{q}_s$'s as a function of s, with the observed q_s^{obs} in red. The point at which q_s^{obs} intersects with the p-value = 0.05 contour is marked in black and signifies the upper limit at 95% CL.

the nominal signal strength μ of 1. For the example in Figure 4.4, if we expect $s = 10$ signal events, then we would quote the upper limit as $24/s = 2.4$ on μ at 95% CL.

The second, important, convention is that we use a slightly different criterion for confidence intervals, called "CL_s". This is motivated by situations where we have little sensitivity to the signal we're searching for, as in the below example.

Example 4.1. Let's say we expect $s = 10$ and observe $n = 70$, $m = 100$. Really, what this should indicate is that our search is not at all sensitive, since our search region is completely dominated by background and, hence, we should not draw strong conclusions about the signal strength. However, if we follow the above procedure for calculating the upper limit, we get $\mu \leq 0.001$ at 95% CL.

This is an extremely aggressive limit on μ, where we're excluding the nominal $\mu = 1$ signal at a high confidence level. Given the complete lack of sensitivity to the

signal, this is not a sensible result. The CL_s method solves this problem by considering both the p-value of the signal + background hypothesis H_s (referred to as p_{s+b} or just p_μ for short), *and* the p-value of the background-only hypothesis H_0 (p_b), to define a new criterion:

$$p'_\mu = \frac{p_\mu}{1 - p_b} \tag{4.3}$$

In cases where the signal region is completely background-dominated, the compatibility with the background-only hypothesis should be high, so $p_b \sim 1$ and, hence, p'_μ will be increased. On the other hand, for more sensitive regions, compatibility should be lower $\Rightarrow p_b \sim 0$ and $p'_\mu \sim p_\mu$.

To be explicit, here

$$p_b = \int_{-\infty}^{\tilde{t}_{\mathrm{obs}}} p(\tilde{t}_s|0)\mathrm{d}\tilde{t}_s, \tag{4.4}$$

where we should note that:

1. We're looking at the distribution of $\tilde{t}_s$ — *not* $\tilde{t}_0$ — under the background-only hypothesis, since the underlying test is of H_s, not H_0; and
2. We're integrating *up to* $\tilde{t}_{\mathrm{obs}}$, unlike for p_s, because lower $\tilde{t}$ means greater compatibility with the background-only hypothesis.

The effect of the CL_s criterion is demonstrated in Figure 4.5 for Examples 3.1 and 4.1. In the former, the background-only distribution is shifted to the right of the $s + b$ distribution. This indicates that the experiment is sensitive to μ and, indeed, we find $p'_\mu \sim p_\mu$. In Example 4.1, however, the search is not sensitive and, hence, the background-only and $s + b$ distributions almost completely overlap, meaning $p_b \sim 1$

and $p'_\mu >> p_\mu$.[1]

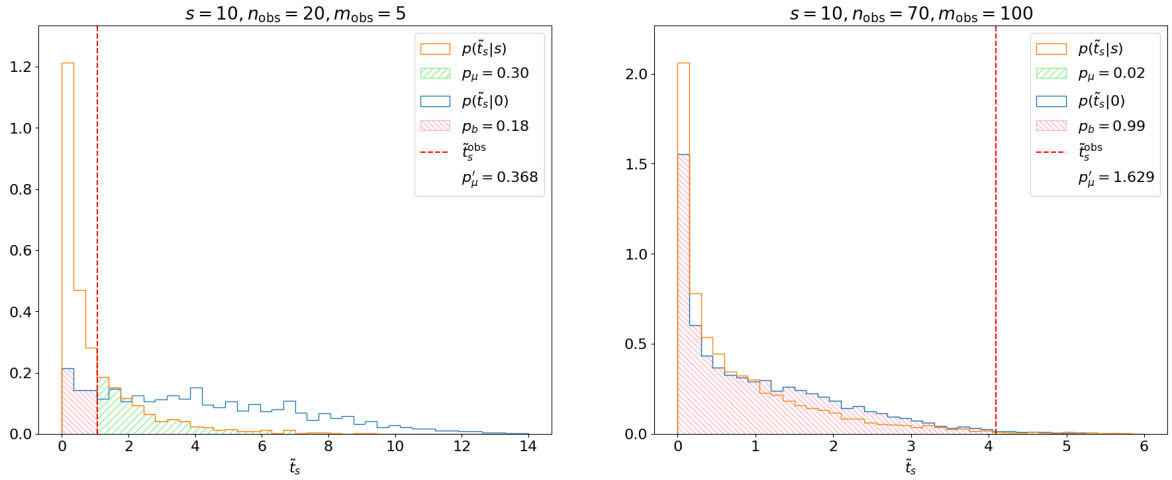

Figure 4.5: Demonstration of CL_s criterion for Examples 3.1 (left) and 4.1 (right).

Finally, if we repeat the Neyman construction using the CL_s criterion p'_μ instead of p_μ for Example 4.1, we can find an upper limit of $\mu \leq 1.2$ at 95% CL, which is indeed a looser, more conservative, upper limit. The upper limit for Example 3.1 remains unchanged at $\mu \leq 2.4$, as we would expect.

[1]Note that, unlike $p(\tilde{t}_s|s)$, $p(\tilde{t}_s|0)$ *doesn't* follow a simple χ^2; asymptotically, it is closer to a *noncentral* χ^2, as will be discussed in Section 7.

Chapter 5

Expected significances and limits

5.1 Expected significance

The focus so far has been only on evaluating the *results* of experiments. However, it is equally important to characterize the expected sensitivity of the experiment *before* running it (or before looking at the data).

Example 5.1. Concretely, we continue with the simple one-bin counting experiment (Section 2.1). Let's say we expect $b = 10$ background events and — at the nominal signal strength $\mu = 1$ — $s = 10$ signal events. How do we tell if this experiment is at all useful for discovering this signal, i.e., does it have any sensitivity to the signal?

One way is to calculate the significance with which we expect to exclude the background-only hypothesis if the signal were, in fact, to exist. Practically, this means we are testing H_0 and, hence, need $p(\tilde{t}_0|\mu = 0)$ as before. However, now we also need the distribution of the test statistic $\tilde{t}_0$ under the *background + signal* hypothesis

$p(\tilde{t}_0|\mu = 1)$. Then, by calculating the significance for each sampled $\tilde{t}_0$ under $H_{\mu=1}$, we can estimate the distribution of expected significances. This is illustrated for Example 5.1 in Figure 5.1.

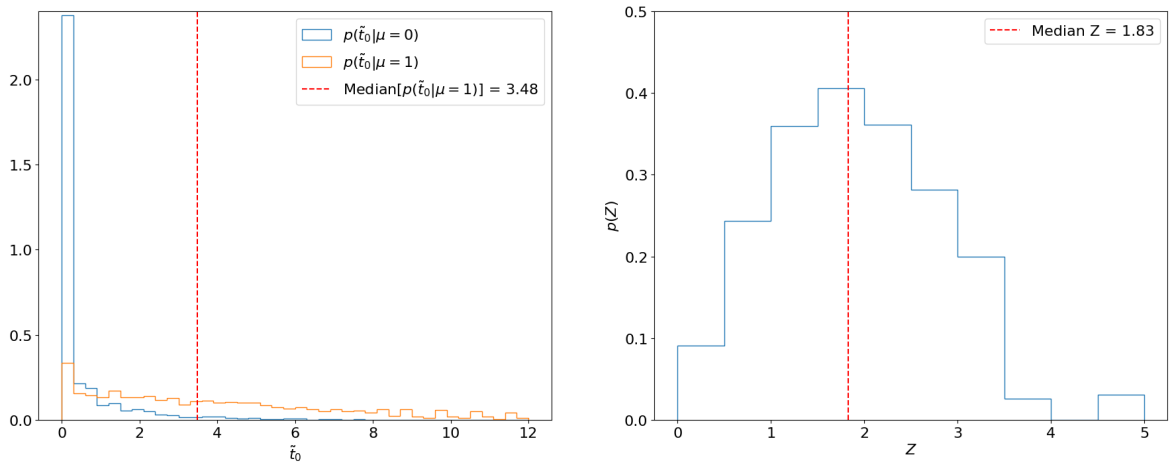

Figure 5.1: Left: Distributions of $\tilde{t}_0$ under the background-only and background + signal hypotheses using 30,000 toys each. The median of the latter is marked in red. Right: Distribution of the significances (with respect to the background-only hypothesis) of each sampled $\tilde{t}_0$ under the signal hypothesis.

Importantly, we usually quote the **median** of this distribution as the expected significance, since the median is "invariant" to monotonic transformations (i.e., the median p-value will always correspond to the median Z as well, whereas the mean p-value will not correspond to the mean Z). Similarly, we quote the $16\%/84\%$ and $2\%/98\%$ quantiles as the $\pm 1\sigma$ and $\pm 2\sigma$, respectively, expected significances. These quantiles correspond to the cumulative probabilities for a standard Gaussian (Figure 5.2). For Example 5.1, we thus find the median expected significance to be 1.83.

Note that instead of converting each sampled $\tilde{t}$ under $H_{\mu=1}$ into a significance and finding the median of that distribution, as in Figure 5.1 (right), we can take advantage of the invariance of the median and directly use the significance of the median $\tilde{t}$ under $H_{\mu=1}$ (Figure 5.1, left). We will do this below for the expected limit.

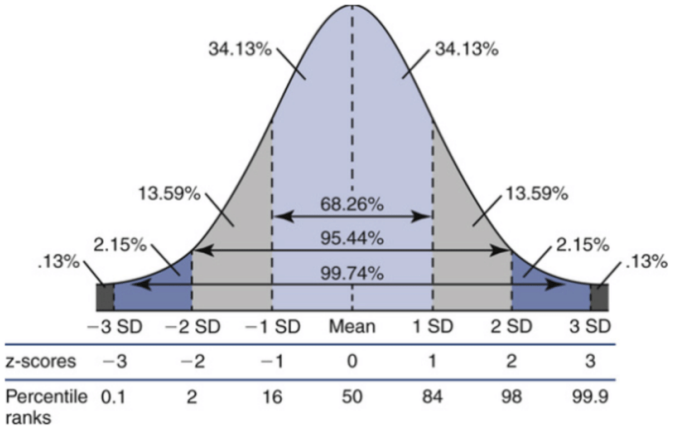

Figure 5.2: Gaussian quantiles, reproduced from Ref. [5].

5.2 Expected limits

The other figure of merit we care about in searches is the upper exclusion limit set on the signal strength. To derive the *expected* limit, we do the opposite of the above and ask, if the signal were not to exist, what value of μ would we expect to exclude at the 95% CL.[1]

This means we need:

1. The distribution $p(\tilde{q}_\mu|\mu)$ to solve for μ^+ in Eq. 4.1 and be able to do the upper limit calculation (as in Section 4.2);
2. $p(\tilde{q}_\mu|0)$ to get the median (and other quantiles') expected $\tilde{q}_\mu^{\text{obs}}$ for different signal strengths under the background-only hypothesis; and, furthermore,
3. To scan over the different signal strengths to find the μ that results in a median p-value of 0.05 — or, rather, p'_μ-value (Eq. 4.3), since we're using the CL_s method for upper limits (Section 4.3).

[1] 95% is the standard CL for upper limits in HEP.

First, let's look at the first two steps for just the $\mu = 1$ signal strength in Example 5.1. These steps are similar to, and essentially an inversion of, the procedure for the expected significance: we're now finding the $p'_{\mu=1}$-value with respect to the signal + background hypothesis, for the median $\tilde{q}_\mu$ sampled under the background-only hypothesis. This is demonstrated in Figure 5.3.

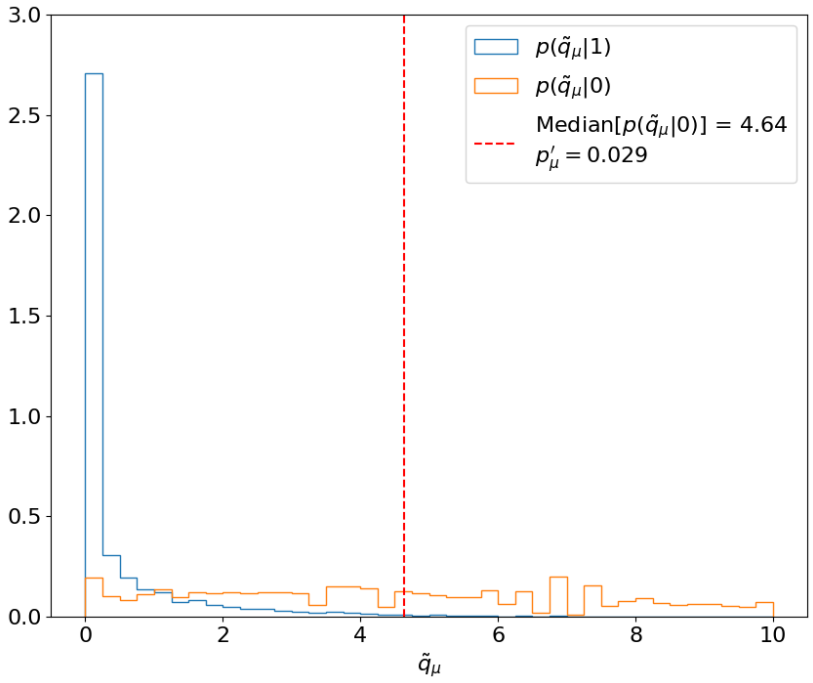

Figure 5.3: Calculating the median expected $p'_{\mu=1}$-value with respect to the signal + background hypothesis, for test statistics $\tilde{q}_\mu$ sampled under the background-only hypothesis. $p(\tilde{q}_\mu|1)$ and $p(\tilde{q}_\mu|0)$ are estimated using 30,000 toys each. Then, the median $p(\tilde{q}_\mu|0)$ (red) is used to calculate the p'_μ-value following the CL_s criterion.

The key difference with respect to calculating the expected significance is step 3, in which this procedure has to be repeated for a range of signal strengths to find the value that gives a median (and $\pm 1\sigma, \pm 2\sigma$ quantile-) p'_μ of 0.05. This is thus the minimum value of μ that we expect to be able to exclude at 95% CL, as shown in Figure 5.4.

Thus, we have our expected limits. The right plot of Figure 5.4 is colloquially

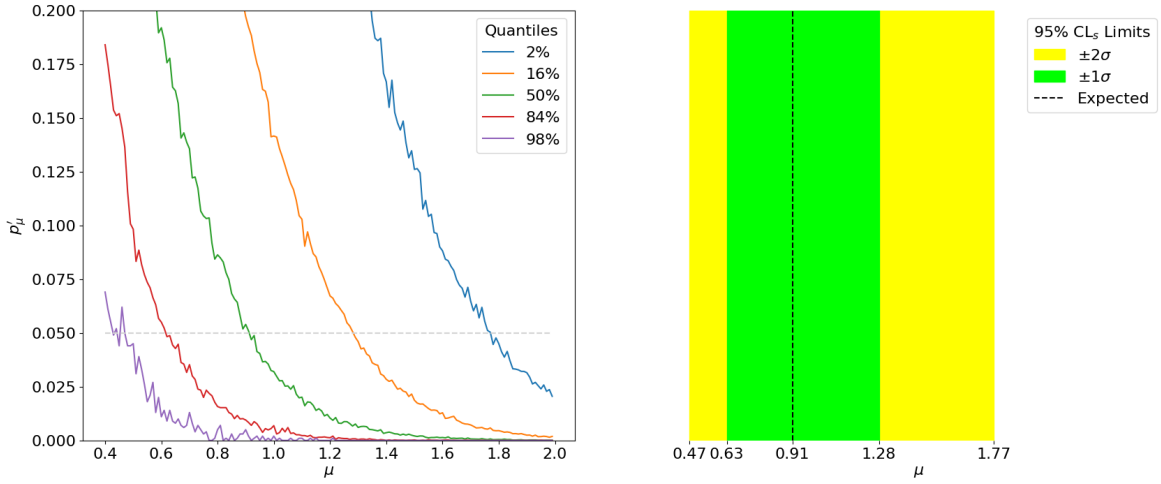

Figure 5.4: Left: The expected median and $\pm 1\sigma, \pm 2\sigma$ quantiles of p'_μ for different μ's. The intersection of these with $p'_\mu = 0.05$ (gray) corresponds to the expected exclusion limits. Right: The median and $\pm 1\sigma, \pm 2\sigma$ expected limits at 95% CL$_s$ on μ.

known as a "Brazil-band plot", and is the standard way of representing limits. For example, Figure 5.5 is the corresponding plot by ATLAS for the Higgs discovery (scanning over the Higgs mass).

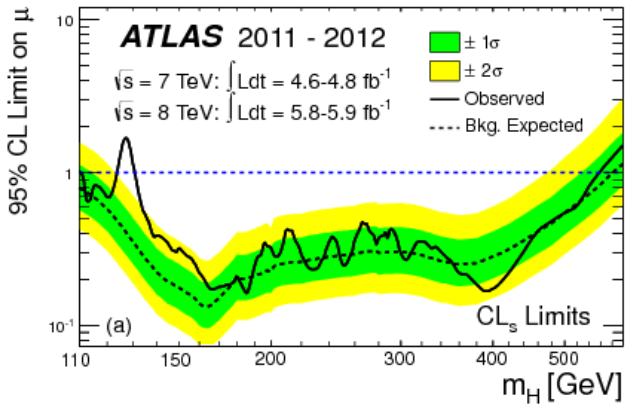

Figure 5.5: Expected and observed 95% CL$_s$ upper limits for the SM Higgs by ATLAS in 2012, for different hypothetical Higgs masses [6].

Part II

Asymptotic formulae

Chapter 6

Asymptotic form of the MLE

So far, we have discussed how to extract meaningful statistical results from HEP experiments by making extensive use of pseudodata / toy experiments to estimate the sampling distributions of profile-likelihood-ratio-based test statistics. While this worked nicely for our simple counting experiment, generating a sufficiently large number of toys can quickly become computationally intractable for the more complex searches (and statistical combinations of searches) that are increasingly prevalent at the LHC, containing at times up to thousands of bins and nuisance parameters. This and the following section discuss a way to approximate these sampling distributions without the need for pseudodata. This was introduced in the famous "CCGV" paper [1] in 2011 and has since become the de-facto procedure at the LHC.

As hinted at previously, such as in Figures 3.2 and 4.3, the distributions $p(\tilde{t}_\mu|\mu')$ and $p(\tilde{q}_\mu|\mu')$ (where, in general, $\mu' \neq \mu$) have similar forms regardless of the nuisance parameters (or sometimes even the POIs). This is not a coincidence: we will now

derive their "asymptotic" — i.e., in the large sample limit — forms, starting first with the asymptotic form of the maximum likelihood estimator (MLE).

It is important to remember that the MLE $\hat{\mu}$ of μ is a random variable with its own probability distribution. We can estimate it as always by sampling toys, shown in Figure 6.1 for our counting experiment (Eq. 2.3). One can observe that $p(\hat{\mu})$ follows a Gaussian distribution as the number of events N increases, and indeed this becomes clear if we try to fit one to the histograms (Figure 6.2). We will now show this to be true generally, deriving the analytic distribution in Sections 6.1—6.3, and discussing the results and the important concept of the *Asimov* dataset for numerical estimation in Sections 6.4 and 6.5, respectively.

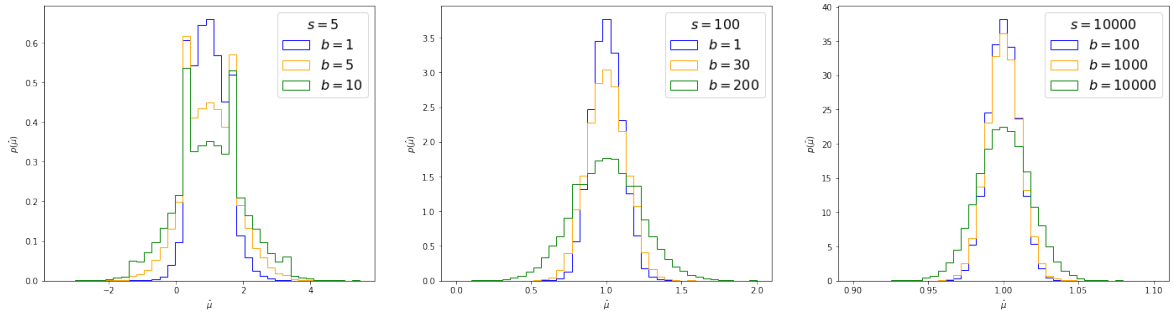

Figure 6.1: Distribution of the MLE of μ for different s and b produced using 30,000 toy experiments each. (Note the x-axis range is becoming narrower from the left-most to the right-most plot.)

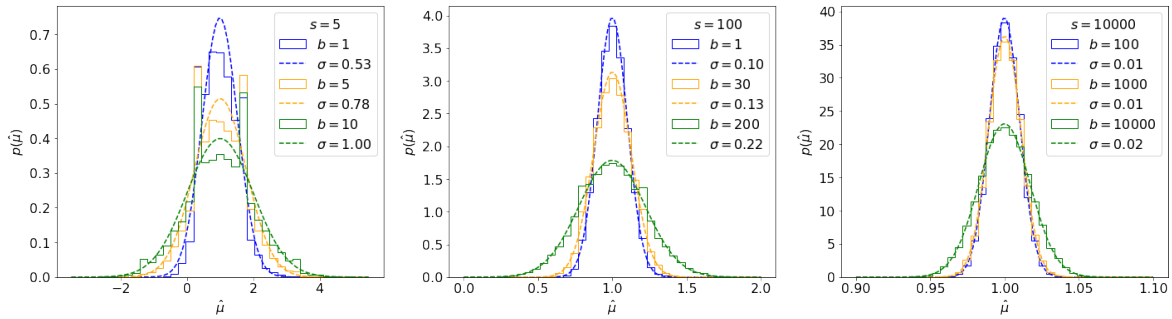

Figure 6.2: Gaussian fits to distributions of $\hat{\mu}$ for different s and b from Figure 6.1.

6.1 Statistics background

We first provide a lightning review of some necessary statistics concepts and results.

Definition 6.1. Let the negative log-likelihood (NLL) $-\ln L(\mu) \equiv -l(\mu)$. The derivative of the NLL $-l'(\mu)$ is called the **score** $s(\mu)$. It has a number of useful properties: [1]

1. Its expectation value at μ' $\mathbb{E}_{\mu=\mu'}[s(\mu')] = 0$.
2. Its variance $\mathrm{Var}[s(\mu)] = -\mathbb{E}[l''(\mu)]$.

Note that the expectation value here means an average over observations which are distributed according to a particular μ, which here we're calling the "true" μ: μ'.

Definition 6.2. $-\mathbb{E}[l''(\mu)] \equiv \mathcal{I}(\mu)$ is called the **Fisher information**. It quantifies the information our data contains about μ and importantly, as we'll see, it (approximately) represents the inverse of the variance of $\hat{\mu}$. More generally, for multiple parameters,

$$\mathcal{I}_{ij}(\mu) = -\mathbb{E}[\frac{\partial^2 l}{\partial \mu_i \partial \mu_j}] \tag{6.1}$$

is the Fisher information matrix. It is also commonly called the **covariance matrix**.

Theorem 6.1. Putting this together, by the central limit theorem [8], this means $p(s(\mu'))$ follows a normal distribution with mean 0 and variance $\mathcal{I}(\mu')$, up to terms of order $O(\frac{1}{\sqrt{N}})$:

$$s(\mu') \xrightarrow{\sqrt{N}>>1} \mathcal{N}(0, \sqrt{\mathcal{I}(\mu')}), \tag{6.2}$$

where N represents the data sample size.

[1] See derivations in e.g. Ref. [7].

6.2 The Fisher information

For our simple counting experiment, the Fisher information matrix $\mathcal{I}(\mu, b)$ can be found by taking second derivatives of the NLL (Eq. 2.5). The $\mathcal{I}_{\mu\mu}$ term, for example, is:

$$\mathcal{I}_{\mu\mu}(\mu, b) = -\mathbb{E}[\partial^\mu \partial^\mu l(\mu, b)] = \mathbb{E}\left[n \cdot \frac{s^2}{(\mu s + b)^2}\right] = \mathbb{E}[n] \cdot \frac{s^2}{(\mu s + b)^2} = \frac{(\mu' s + b')s^2}{(\mu s + b)^2}. \quad (6.3)$$

In the last step we use the fact that $\mathbb{E}[n]$ under true $\mu = \mu'$, $b = b'$, is $\mu' s + b'$. For the remainder of this section, $\mathcal{I}(\mu, b)$ will always be evaluated at the true values of the parameters,[2] so this can be simplified to $\mathcal{I}_{\mu\mu}(\mu', b') = \frac{s^2}{\mu' s + b'}$. This is plotted in Figure 6.3, where we can the Fisher information captures the fact that as b increases, we lose sensitivity to — or *information* about — μ.

For completeness (and since we'll need it below), the full Fisher information matrix for our problem, repeating the steps in Eq. 6.3, is:

$$\mathcal{I}(\mu', b') = \begin{pmatrix} \mathcal{I}_{\mu\mu} & \mathcal{I}_{\mu b} \\ \mathcal{I}_{b\mu} & \mathcal{I}_{bb} \end{pmatrix} (\mu', b') = \begin{pmatrix} \frac{s^2}{\mu' s + b'} & \frac{s}{\mu' s + b'} \\ \frac{s}{\mu' s + b'} & \frac{1}{\mu' s + b'} + \frac{1}{b'} \end{pmatrix} \quad (6.4)$$

6.3 Derivation

We now have enough background to derive the asymptotic form of the MLE. We do this for the 1D case by Taylor-expanding the score of $\hat{\mu}$, $l'(\hat{\mu})$ - which we know to be

[2] The reason for this is discussed shortly in Section 6.3.

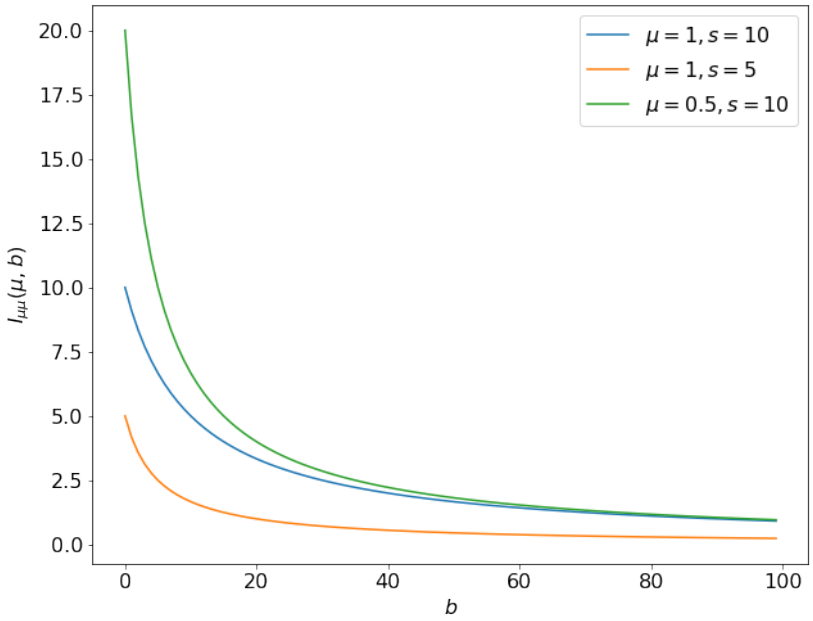

Figure 6.3: The Fisher information $\mathcal{I}_{\mu\mu}(\mu, b)$ for different μ and s, as a function of the expected background b.

$= 0$ - around μ':

$$l'(\hat{\mu}) = l'(\mu') + l''(\mu')(\hat{\mu} - \mu') + O((\hat{\mu} - \mu')^2) = 0 \tag{6.5}$$

$$\Rightarrow \hat{\mu} - \mu' \simeq -\frac{l'(\mu')}{l''(\mu')} \xrightarrow{\sqrt{N}>>1} \frac{1}{\mathcal{I}(\mu')} N(0, \sqrt{\mathcal{I}(\mu')}) = N\left(0, \frac{1}{\sqrt{\mathcal{I}(\mu')}}\right), \tag{6.6}$$

where we plugged in the distribution of $l'(\mu')$ from Eq. 6.2, claimed $l''(\mu')$ asymptotically equals its expectation value $\mathbb{E}[l''(\mu')] = \mathcal{I}(\mu')$ by the law of large numbers [9], and are ignoring the $O((\hat{\mu} - \mu')^2)$ term.[3]

For multiple parameters, $\mathcal{I}$ is a matrix so the variance generalized to the matrix inverse:

$$\hat{\mu} - \mu' \simeq N(0, \sqrt{\mathcal{I}^{-1}_{\mu\mu}(\mu', b')}), \tag{6.7}$$

[3]For a more rigorous derivation, see e.g. Ref. [10].

6.4 Result

Thus, we see that $\hat{\mu}$ asymptotically follows a normal distribution around the true μ value, μ', with a variance $\sigma^2_{\hat{\mu}} = \mathcal{I}^{-1}_{\mu\mu}(\mu', b')$, up to $\mathcal{O}(1/\sqrt{N})$ terms. Intuitively, from the definition of the Fisher information $\mathcal{I}$, we can interpret this as saying that the more information we have about μ from the data, the lower the variance should be on $\hat{\mu}$.

Continuing with our counting experiment from Section 2.1, inverting $\mathcal{I}$ from Eq. 6.4 gives us

$$\sigma_{\hat{\mu}} = \sqrt{\mathcal{I}^{-1}_{\mu\mu}(\mu', b')} = \frac{\sqrt{\mu' s + 2b'}}{s}. \tag{6.8}$$

Note that, as we might expect, this scales as $\sim \sqrt{b}$, which is the uncertainty of our Poisson nuisance parameter b — showing mathematically why we want to keep uncertainties on nuisance parameters as low as possible. This is compared to the toy-based distributions from Section 6 in Figure 6.4 this time varying the true signal strength μ' as well, where we can observe that this matches very well for large s, b, while for small values there are some discrete differences.

We can also check the total per-bin errors between the asymptotic form and the toy-based distributions directly, as shown in Figure 6.5 (for $\mu' = 1$ only). Indeed, this confirms that the error scales as $\sim \frac{1}{\sqrt{s}}$ and $\sim \frac{1}{\sqrt{b}}$, as claimed above.

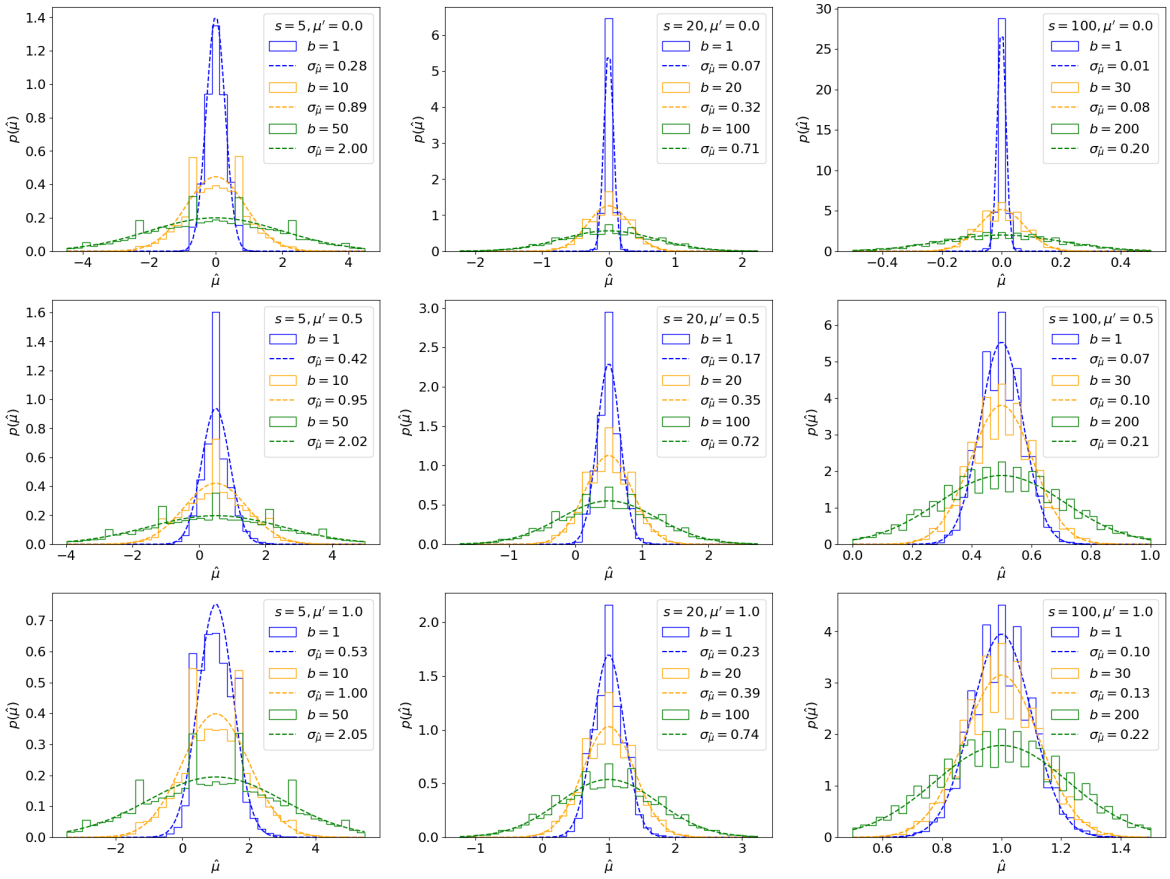

Figure 6.4: Asymptotic (dotted lines) and toy-based (solid lines) distributions, using 30,000 toys each, of the MLE of μ for different s, b, and true signal strengths μ'.

6.5 Numerical estimation and the Asimov dataset

In this section, because of the simplicity of our data model, we were able to derive the Fisher information $\mathcal{I}$ and, hence, the asymptotic form of $\hat{\mu}$ analytically. In general, this is not possible and we typically have to minimize l, find its second derivatives, and solve Eq. 6.3 etc. *numerically* instead.

However, when calculating the Fisher information, how do we deal with the expectation value over the observed data (n, m in our case)? Naively, this would

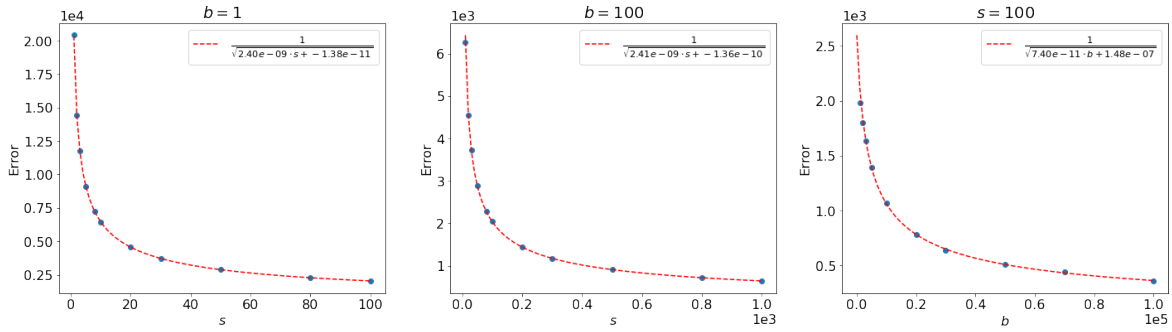

Figure 6.5: Error between the sampled toy distributions, using 50,000 toys each, and the asymptotic distributions of the MLE of μ for different s and b (blue), with $1/\sqrt{N}$ fits in red.

require averaging over a bunch of generated toy n, m values again, which defeats the purpose of using the asymptotic form of $\hat{\mu}$!

Instead, we can switch the order of operations in Eq. 6.3,[4] rewriting it as:

$$\mathcal{I}_{ij}(\mu, b) = -\mathbb{E}[\partial^i \partial^j l(\mu, b; n, m)] = -\partial^i \partial^j \mathbb{E}[l(\mu, b; n, m)] = -\partial^i \partial^j l(\mu, b; \mathbb{E}[n], \mathbb{E}[m]). \tag{6.9}$$

Importantly, this says we can find $\mathcal{I}$ by simply evaluating the likelihood for a dataset of observations equal to their expectation values under μ' instead of averaging over the distribution of observations and *then* getting its second derivatives.

Definition 6.3. Such a dataset is called the **Asimov** dataset, and $L(\mu; \mathbb{E}[n], \mathbb{E}[m]) \equiv L_A$ is referred to as the "Asimov likelihood".[5]

[4] We are able to do this because, as we saw above, the score is linear in n for Poisson likelihoods.

[5] The *Asimov* dataset is named after Isaac Asimov, the popular science fiction author, whose book *Franchise* is about a supercomputer choosing a single person as the sole voter in the U.S. elections, because they can represent the entire population.

Chapter 7

Asymptotic form of the profile likelihood ratio

We can now proceed to derive the asymptotic form of the sampling distribution $p(t_\mu|\mu')$ of the profile likelihood ratio test statistic t_μ, under a "true" signal strength of μ'. This asymptotic form is extremely useful for simplifying the computation of (expected) significances, limits, and intervals; indeed, standard procedure at the LHC is to use it in lieu of toy-based, empirical distributions for $p(t_\mu|\mu')$.

7.1 Asymptotic form of the profile likelihood ratio

We start with deriving the asymptotic form of the profile likelihood ratio test statistic t_μ (Eq. 2.7) by following a similar procedure to Section 6.3 — and using the results

therein — of Taylor expanding around its minimum at $\hat{\mu}$:[1]

$$t_\mu = -2\ln\lambda(\mu) \tag{7.1}$$

$$= -2l(\mu, \hat{\hat{b}}(\mu)) + 2l(\hat{\mu}, \hat{b}) \tag{7.2}$$

$$\simeq \underbrace{-2l(\hat{\mu}, \hat{\hat{b}}(\hat{\mu})) + 2l(\hat{\mu}, \hat{b})}_{\hat{\hat{b}}(\hat{\mu})=\hat{b} \text{ so this is 0}} \underbrace{- 2l'(\hat{\mu}, \hat{\hat{b}}(\hat{\mu}))(\mu - \hat{\mu})}_{l'(\hat{\mu},\hat{b})=0} - 2l''(\hat{\mu}, \hat{\hat{b}}(\hat{\mu})) \cdot \frac{(\mu - \hat{\mu})^2}{2} \tag{7.3}$$

$$= -l''(\hat{\mu}, \hat{b}) \cdot (\mu - \hat{\mu})^2 \tag{7.4}$$

$$= \underbrace{-\mathbb{E}[l''(\hat{\mu}, \hat{b})]}_{\text{By law of large numbers}} \cdot(\mu - \hat{\mu})^2 \tag{7.5}$$

$$= \underbrace{-\mathbb{E}[l''(\mu', b')]}_{\text{Since bias of MLEs} \sim 0} \cdot(\mu - \hat{\mu})^2 \tag{7.6}$$

$$= \underbrace{\mathcal{I}_{\mu\mu}(\mu', b')}_{\text{From definition of Fisher information}} \cdot(\mu - \hat{\mu})^2 \tag{7.7}$$

$$\Rightarrow \underbrace{\boxed{t_\mu \simeq \frac{(\mu - \hat{\mu})^2}{\sigma^2_{\hat{\mu}}}}}_{\text{Using } \sigma_{\hat{\mu}} \simeq \sqrt{\mathcal{I}^{-1}_{\mu\mu}(\mu', b')}} + \mathcal{O}((\mu - \hat{\mu})^3) + \mathcal{O}(\frac{1}{\sqrt{N}}). \tag{7.8}$$

Here, just like in Eq. 6.6, we use the law of large numbers in Line 7.5 and take $l''(\hat{\mu}, \hat{b})$ to asymptotically equal its expectation value under the true parameter values μ', b': $l''(\hat{\mu}, \hat{b}) \xrightarrow{\sqrt{N}>>1} \mathbb{E}[l''(\hat{\mu}, \hat{b})]$. We then in Line 7.6 also use the fact that MLEs are generally unbiased estimators of the true parameter values in the large sample limit to say $\mathbb{E}[l''(\hat{\mu}, \hat{b})] \xrightarrow{\sqrt{N}>>1} \mathbb{E}[l''(\mu', b')]$. Finally, in the last step, we use the asymptotic form of the MLE (Eq. 6.7).

[1] Note: this is not a rigorous derivation; it's just a way to motivate the final result, which is taken from Ref. [1]. (If you know of a better way, let me know!)

7.2 Asymptotic form of $p(t_\mu|\mu')$

Now that we have an expression for t_μ, we can consider its sampling distribution. With a simple change of variables, the form of $p(t_\mu|\mu')$ should hopefully be evident: recognizing that μ and $\sigma^2_{\hat{\mu}}$ are simply constants, while $\hat{\mu}$ we know is distributed as a Gaussian centered around μ' with variance $\sigma^2_{\hat{\mu}}$, let's define $\gamma \equiv \frac{\mu-\hat{\mu}}{\sigma_{\hat{\mu}}}$, so that

$$t_\mu \simeq \frac{(\mu-\hat{\mu})^2}{\sigma^2_{\hat{\mu}}} = \gamma^2, \tag{7.9}$$

$$\gamma \sim \mathcal{N}\left(\frac{\mu-\mu'}{\sigma_{\hat{\mu}}}, 1\right). \tag{7.10}$$

For the special case of $\mu = \mu'$, we can see that t_μ is simply the square of a standard normal random variable, which is the definition of the well-known χ^2_k distribution with $k = 1$ degrees of freedom (DoF):

$$p(t_\mu|\mu) \sim \chi^2_1. \tag{7.11}$$

In the general case where μ may not $= \mu'$, t_μ is the square of random variable with unit variance but *non-zero mean*. This is distributed as the similar, but perhaps less well-known, **non-central chi-squared** $\chi'^2_k(\Lambda)$, again with 1 DoF, and with a "non-centrality parameter"

$$\Lambda = \bar{\gamma}^2 = \left(\frac{\mu-\mu'}{\sigma_{\hat{\mu}}}\right)^2, \tag{7.12}$$

$$p(t_\mu|\mu') \sim \chi'^2_1(\Lambda). \tag{7.13}$$

The "central" vs. non-central chi-squared distributions are visualized in Figure 7.1 for $k = 1$. We can see that $\chi'^2_k(\Lambda)$ simply shifts towards the right as Λ increases (at $\Lambda = 0$ it is a regular central χ^2). As $\Lambda \to \infty$, $\chi'^2_k(\Lambda)$ becomes more and more like a normal distribution with mean Λ.[2]

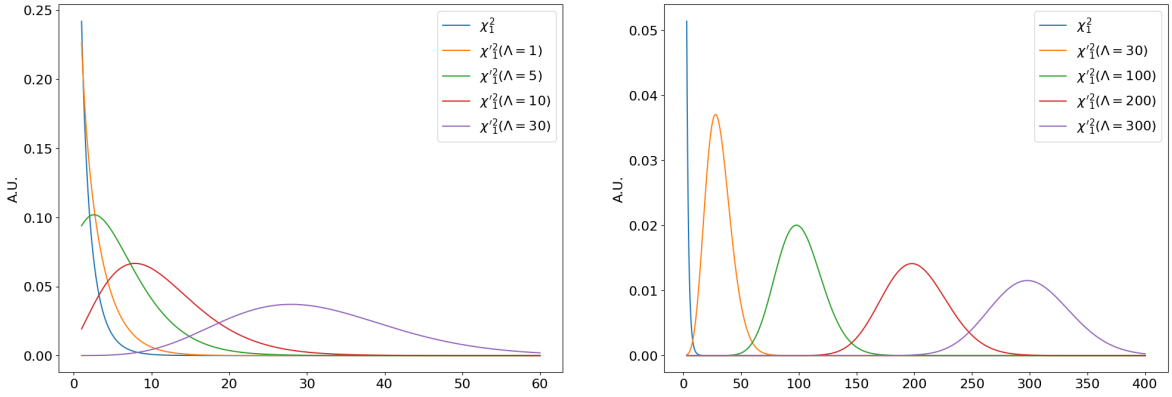

Figure 7.1: Central χ^2_k and non-central $\chi'^2_k(\Lambda)$ distributions for Λ between $1 - 30$ (left) and $30 - 300$ (right).

By extending the derivation in Eq. 7.8 to multiple POIs, one can find the simple generalization to multiple POIs μ:

$$p(t_\mu|\mu') \sim \chi'^2_k(\Lambda), \tag{7.14}$$

where the DoF k are equal the number of POIs $\dim \mu$, and

$$\Lambda = (\mu - \mu')^T \cdot \tilde{\mathcal{I}}^{-1}(\mu') \cdot (\mu - \mu'), \tag{7.15}$$

where $\tilde{\mathcal{I}}^{-1}$ is $\mathcal{I}^{-1}$ restricted only to the components corresponding to the POIs.

[2]More information can be found in e.g. Ref. [11].

7.3 Estimating $\sigma^2_{\hat{\mu}}$

The critical remaining step to understanding the asymptotic distribution of t_μ is estimating $\sigma^2_{\hat{\mu}}$ to find the non-centrality parameter Λ in Eq. 7.12. We now discuss two methods to do this.

7.3.1 Method 1: Inverting the Fisher information / covariance matrix

The first method is simply using $\sigma_{\hat{\mu}} \simeq \sqrt{\mathcal{I}^{-1}_{\mu\mu}(\mu', b')}$ as in Section 6.[3] This is shown in Figure 7.2 for our counting experiment, using the analytic form for $\sigma_{\hat{\mu}}$ from Eq. 6.8. We can see that this asymptotic approximation agrees well with the true distribution for some range of parameters, but can deviate significantly for others, as highlighted especially in the right plot.

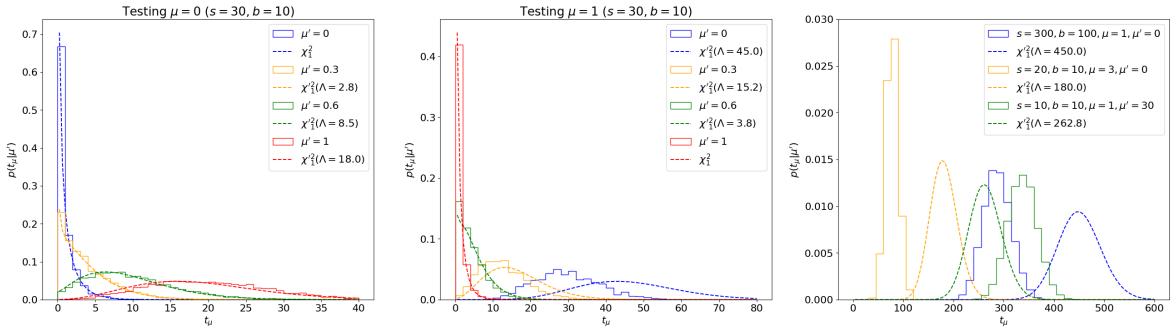

Figure 7.2: Comparing the distribution $p(t_\mu|\mu')$ (solid) with non-central $\chi'^2_1(\Lambda)$ distributions (dotted) for a range of s, b, μ, μ' values, with $\sigma^2_{\hat{\mu}}$ estimated using the inverse of the Fisher information matrix.

[3] More generally, we'd need $\tilde{\mathcal{I}}^{-1}$ for Eq. 7.15.

7.3.2 Interlude on Asimov dataset

While we are able to find the analytic form for $\sqrt{\mathcal{I}^{-1}_{\mu\mu}(\mu', b')}$ easily for our simple counting experiment, in general it has to be calculated numerically. As introduced in Section 6.5, to handle the expectation value under μ', b' in Eq. 6.1, we can make use of the **Asimov dataset**, where the observations n_A, m_A are taken to be their expectation values under μ', b', simplifying the calculation of $\mathcal{I}$ to Eq. 6.9.

Explicitly, for our counting experiment (Eq. 2.3), the Asimov observations are simply

$$n_A = \mathbb{E}[n] = \mu' s + b', \tag{7.16}$$

$$m_A = \mathbb{E}[m] = b'. \tag{7.17}$$

We'll now consider a second powerful use of the Asimov dataset to estimate $\sigma^2_{\hat{\mu}}$.

7.3.3 Method 2: The "Asimov sigma" estimate

Putting together Eqs. 2.10 and 7.17, we can derive a nice property of the Asimov dataset: the MLEs $\hat{\mu}, \hat{b}$ equal the true values μ', b':

$$\hat{b} = m_A = b' \tag{7.18}$$

$$\hat{\mu} = \frac{n_A - m_A}{s} = \frac{\mu' s + b' - b'}{s} = \mu'. \tag{7.19}$$

Thus, t_μ evaluated for the Asimov dataset is exactly the non-centrality parameter Λ that we are after!

$$t_{\mu,A} \simeq \left(\frac{\mu - \hat{\mu}}{\sigma_{\hat{\mu}}}\right)^2 = \left(\frac{\mu - \mu'}{\sigma_{\hat{\mu}}}\right)^2 = \Lambda. \tag{7.20}$$

While, not strictly necessary to obtain the asymptotic form for $p(t_\mu|\mu')$, we can also invert this to estimate $\sigma_{\hat{\mu}}$, as

$$\sigma_A \simeq \frac{(\mu - \mu')^2}{t_{\mu,A}}, \tag{7.21}$$

where σ_A is known as the "Asimov sigma".

The asymptotic distributions using $\Lambda = t_{\mu,A}$ are plotted in Figure 7.3. We see that this estimate matches the sampling distributions very well, even for cases where the covariance-matrix-estimate failed! Indeed, this is why estimating $\sigma_{\hat{\mu}} \simeq \sigma_A$ is the standard in LHC analyses, and that is the method we'll employ going forward.

Reference [1] conjectures that this is because the Fisher-information-approach is restricted only to estimating the second-order term of Eq. 7.8, while with $t_{\mu,A}$ we're matching the shape of the likelihood at the minimum which may be able capture some of the higher order terms as well.

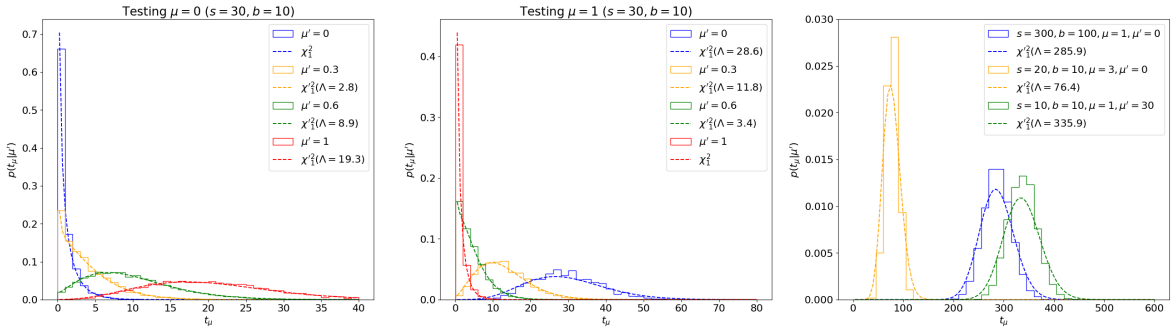

Figure 7.3: Comparing the sampling distribution $p(t_\mu|\mu')$ with non-central $\chi'^2_1(\Lambda)$ distributions for a range of s, b, μ, μ' values, with the Asimov sigma estimation for $\sigma^2_{\hat{\mu}}$.

Despite the pervasive use of the asymptotic formula at the LHC, it's important to remember that it's an *approximation*, only valid for large statistics. Figure 7.4 shows it breaking down for $s, b \lesssim 10$ below.

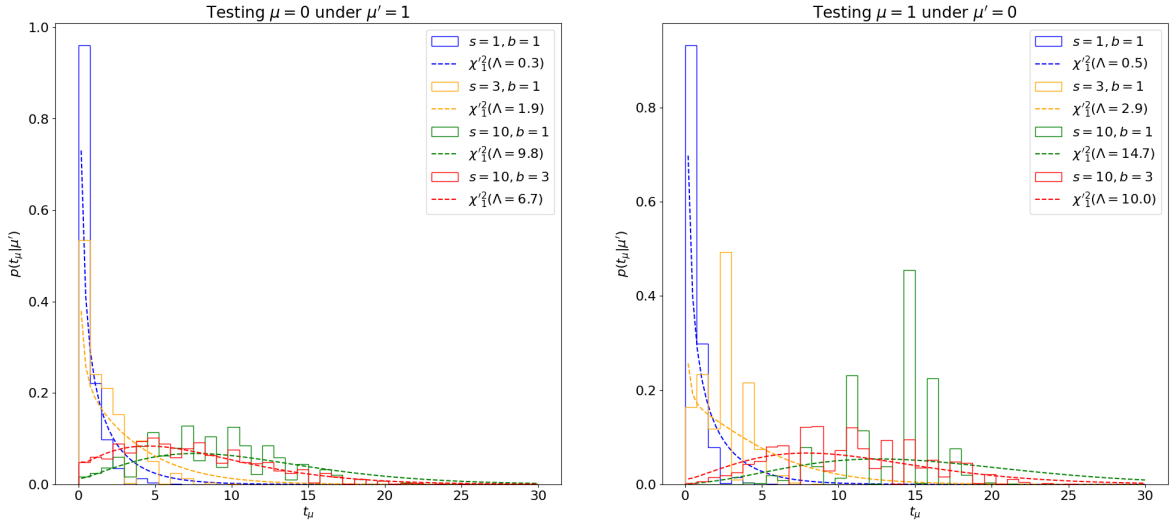

Figure 7.4: Comparing the sampling distribution $p(t_\mu|\mu')$ with non-central $\chi'^2_1(\Lambda)$ distributions for different $s, b \leq 10$, showing the break-down of the σ_A approximation for $\sigma^2_{\hat{\mu}}$ at low statistics.

7.4 The PDF and CDF

The probability distribution function (PDF) for a $\chi'^2_k(\Lambda)$ distribution can be found in e.g. Ref. [11] for $k = 1$:

$$p(t_\mu|\mu') \simeq \chi'^2_1(\Lambda) = \frac{1}{2\sqrt{t_\mu}}\left(\varphi(\sqrt{t_\mu} - \sqrt{\Lambda}) + \varphi(\sqrt{t_\mu} + \sqrt{\Lambda})\right), \tag{7.22}$$

where φ is the PDF of a standard normal distribution. For $\mu = \mu' \Rightarrow \Lambda = 0$, this simplifies to:

$$p(t_\mu|\mu) \simeq \chi^2 = \frac{1}{\sqrt{t_\mu}}\varphi(\sqrt{t_\mu}). \tag{7.23}$$

The cumulative distribution function (CDF) for $k = 1$ is:

$$F(t_\mu|\mu') \simeq \Phi(\sqrt{t_\mu} - \sqrt{\Lambda}) + \Phi(\sqrt{t_\mu} + \sqrt{\Lambda}) - 1, \tag{7.24}$$

where Φ is the CDF of the standard normal distribution. For $\mu = \mu' \Rightarrow \Lambda = 0$, again this simplifies to:

$$F(t_\mu|\mu) \simeq 2\Phi(\sqrt{t_\mu}) - 1. \tag{7.25}$$

From Eq. 3.1, we know the p-value p_μ of the observed t_μ^{obs} under a signal hypothesis of H_μ is

$$p_\mu = 1 - F(t_\mu^{\text{obs}}|\mu) = 2(1 - \Phi(\sqrt{t_\mu^{\text{obs}}})), \tag{7.26}$$

with an associated significance

$$Z = \Phi^{-1}(1 - p_\mu) = \Phi^{-1}(2\Phi(\sqrt{t_\mu^{\text{obs}}} - 1) \tag{7.27}$$

7.5 Application to hypothesis testing

Let's see how well this approximation agrees with the toy-based p-value we found in Example 3.1. For the same counting experiment example, where we expect $s = 10$ and observe $n_{\text{obs}} = 20$, $m_{\text{obs}} = 5$, we found the p-value for testing the $\mu = 1$ hypothesis $p_{\mu=1} = 0.3$ (and the associated significance $Z = 0.52$). Calculating t_μ^{obs} for this example

and plugging it into the asymptotic approximation from Eq. 7.26 gives:[4]

$$t_\mu^{\text{obs}} = 1.08 \tag{7.28}$$

$$\Rightarrow p_{\mu=1} = 2(1 - \Phi(\sqrt{1.08})) = 0.3 \tag{7.29}$$

$$\Rightarrow Z = 0.52. \tag{7.30}$$

We see that it agrees exactly!

The agreement more generally, with varying $s, \mu, n_{\text{obs}}, m_{\text{obs}}$, is plotted in Figure 7.5. We observe generally strong agreement, except for low n, m where, as expected, the asymptotic approximation breaks down.

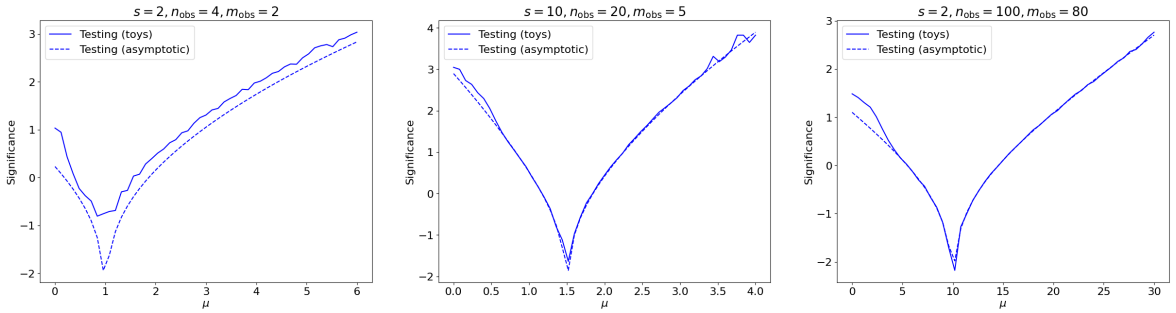

Figure 7.5: Comparing the significances, as a function of the signal strength μ of the hypothesis being tested, for simple counting experiments (Eq. 2.3) with different $s, n_{\text{obs}}, m_{\text{obs}}$'s, derived using $30,000$ toys each (solid) to estimate the $p(t_\mu|\mu)$ distribution vs. the asymptotic approximation (dashed).

7.6 Summary

We have been able to find the asymptotic form for the profile-likelihood-ratio test statistic $t_\mu \simeq \frac{(\mu-\hat{\mu})^2}{\sigma^2_{\hat{\mu}}}$, which is distributed as a *non-central chi-squared* ($\chi'^2_k(\Lambda)$) distribu-

[4]Note that we're using t_μ here, not the alternative test statistic $\tilde{t_\mu}$; however, in this case since $\hat{\mu} > 0$, they are equivalent.

tion. We discussed two methods for finding the non-centrality parameter Λ, out of which the Asimov sigma σ_A estimation generally performed better. Finally, the asymptotic formulae were applied to simple examples of hypothesis testing to check the agreement with toy-based significances. These asymptotic formulae can be extended to the alternative test statistics for positive signals $\tilde{t}_\mu$ and upper-limit-setting $\tilde{q}_\mu$, as in Ref. [1], to simplify the calculation of both observed and expected significances, limits, and intervals.

Bibliography

[1] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells, Asymptotic formulae for likelihood-based tests of new physics, Eur. Phys. J. C **71**, 1554 (2011), [Erratum: Eur.Phys.J.C 73, 2501 (2013)], arXiv:1007.1727 .

[2] Kyle Cranmer, Practical Statistics for the LHC, in *2011 European School of High-Energy Physics* (2014) arXiv:1503.07622 .

[3] Aram Hayrapetyan *et al.* (CMS), The CMS Statistical Analysis and Combination Tool: Combine, Comput. Softw. Big Sci. **8**, 19 (2024), arXiv:2404.06614 [physics.data-an] .

[4] Wikipedia contributors, Type i and type ii errors — Wikipedia, the free encyclopedia (2024), [Online; accessed 7-December-2024].

[5] Caroline Stiller, *Baseline assessment and effect of a supplementary community-based nutrition intervention study on the prevention/treatment of anemia among young Adivasi children in West Bengal, India*, Ph.D. thesis, University of Hohenheim (2021).

[6] ATLAS Collaboration, Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Phys. Lett. B **716**, 1–29 (2012), arXiv:1207.7214 .

[7] Wikipedia contributors, Informant (statistics) — Wikipedia, the free encyclopedia (2024), [Online; accessed 7-December-2024].

[8] Wikipedia contributors, Central limit theorem — Wikipedia, the free encyclopedia, https://en.wikipedia.org/w/index.php?title=Central_limit_theorem&oldid=1257009135 (2024), [Online; accessed 7-December-2024].

[9] Wikipedia contributors, Law of large numbers — Wikipedia, the free encyclopedia, https://en.wikipedia.org/w/index.php?title=Law_of_large_numbers&oldid=1256071749 (2024), [Online; accessed 7-December-2024].

[10] Gregory Gundersen, Asymptotic Normality of MLE, https://gregorygundersen.com/blog/2019/11/28/asymptotic-normality-mle/ (2019), [Online; accessed 7-December-2024].

[11] Wikipedia contributors, Noncentral chi-squared distribution — Wikipedia, the free encyclopedia (2024), [Online; accessed 7-December-2024].