Statistical Analysis Walk-Through: Toy Search for the 125 GeV Higgs in the 4-Muon Final State

Håkon D. Fossheim

https://github.com/fosheimdet/FYSSP100

1 Introduction

The goal of this report is to go through a typical statistical analysis used in HEP with fake data corresponding to a hypothetical search for a 125 GeV Higgs in the 4-muon final state. For the most part, we will try to motivate and explain the statistical methods used along with the results they produce, however, section 2 is devoted to important material used in the analysis, but which are better left on their own to preserve a more structured walk-through.

The analysis is done via frequentist inference. We will use the term null hypothesis, H_0, to refer to the background hypothesis while the alternative hypothesis, H_1, refers to the signal+background hypothesis for a 125 GeV Higgs.

The appendix contain our derivations of the approximations upon which the whole analysis rests, namely that the bin counts are Poisson distributed and that the likelihood is the product of the Poissons of each bin.

2 Theoretical background

Let us consider the total number of particle interactions during our experiment. It is given by[2]

$$N = \sigma \int \mathcal{L}(t) dt \tag{1}$$

Where $\mathcal{L}$ is the instantaneous luminosity and σ is the cross section for the interaction. If we assume that the instantaneous luminosity is constant over the runtime of the experiment, we get

$$N = \sigma \mathcal{L} \Delta t = \sigma L \tag{2}$$

Where L is the luminosity.

Furthermore, the probability to observe a certain number of events in a given bin, viz. $p(n)$, is given by the binomial distribution. The reason for this is that each particle interaction will either produce an event which passes our selection and finally ends up in our bin (success) or not (failure), thereby making these particle interactions Bernoulli trials. The probability of observing n successes out of the N particle interactions in our run will therefore be given by the Binomial [1]

[1] If we sort all the interactions according to when they occurred, the probability of observing n successes with a specific specific time ordering is $p^n(1-p)^{N-n}$ and the binomial coefficient simply counts how many ways there are to re-order the successes without labeling them.

$$P(n) = \binom{N}{n} p^n (1-p)^{N-n} \tag{3}$$

We show in the appendix that as $N \rightarrow \infty$ and $p \rightarrow 0$ with the expected number of events, $\lambda = Np$ fixed, the binomial distribution approaches

$$P(n; \lambda) = \frac{\lambda^n}{n!} e^{-\lambda} \tag{4}$$

, namely the Poisson distribution. Considering the vast number of particle interactions in a run and the extremely low probability of producing a success, the Poisson approximation will be a very good one for our application. The benefit of the Poisson is that we no longer need to consider N and p, which are practically impossible to determine. Rather, we estimate the expected number of events λ as b under the background hypothesis and $s + b$ under the signal+background hypothesis:

$$P_{H_0}(n; b) = \frac{b^n}{n!} e^{-b} \tag{5}$$

$$P_{H_1}(n; s+b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \tag{6}$$

The goal of our analysis is to be able to say something quantitative about how confident we are in the background/null hypothesis *in relation* to the alternative hypothesis. Note that if we end up rejecting the background hypothesis in favor of the alternative hypothesis, this does not necessarily allow us to conclude that H_1 is true, since H_1 is potentially just one of many possible alternative hypotheses one could pose for the data, and there might be others that fit the data even better.

Likewise, rejecting the alternative hypothesis does not allow us to conclude that the background hypothesis is true, since there might be other alternative hypotheses that would not lead to a rejection or the rejection might just have come from a statistically rare data set. To be able to say something about the probability of either hypothesis being true given our data, $\mathbf{n}$, one must also define $P(H_0)$ and $P(H_1)$, since according to Baye's theorem

$$P(H_0|\mathbf{n}) = \frac{P(\mathbf{n}|H_0)P(H_0)}{\sum_{H_0} P(\mathbf{n}|H_0)P(H_0)} \tag{7}$$

$$P(H_1|\mathbf{n}) = \frac{P(\mathbf{n}|H_1)P(H_1)}{\sum_{H_1} P(\mathbf{n}|H_1)P(H_1)} \tag{8}$$

, which is the realm of Bayesian inference and is not dealt with in the frequentist approach, which is what we will apply in our analysis.

We will from now on assume that the alternative hypothesis predicts an excess in events compared to the null hypothesis. In the case of the opposite being true, p-values would be calculated by integrating the left tail of P_{H_0} up to n_{obs} or $s + b$ (if using a single-bin count as the test-statistic).

The observed p-value is given by the probability of observing n_{obs} or more events under the background only hypothesis, i.e.

$$p_{obs} = \sum_{n=n_{obs}}^{\infty} P_{H_0}(n; b) \tag{9}$$

While the expected p-value is given by

$$p_{s+b} = \sum_{n=n_{s+b}}^{\infty} P_{H_0}(n; s+b) \tag{10}$$

and is the probability of observing $s+b$ or more events under the background only hypothesis, i.e. how likely the it is, given the background hypothesis is true, to observe as many or more events as what is expected under the alternative/signal+background hypothesis. In effect, it is a measure of the overlap/disagreement of the null and alternative hypothesis. It is common practice to specify the p-values in terms of significance. The significance, Z, is simply how many standard deviations you have to go from the mean of the standard normal distribution in order for its right tail to have the same area as the p-value. It's given by

$$Z(p) = \phi^{-1}(1-p) \tag{11}$$

Where ϕ^{-1} is the inverse cumulative distribution function.

2.1 Why significance increases with luminosity

The expected significance increases with luminosity. To understand why, consider Eq.(2), namely $N = \sigma L$. An increase in luminosity by a given factor, C, will increase the total number of interactions in our experiment, N, by the same factor. Assuming the probability of an arbitrary interaction producing an event which ends up in our bin, viz. p, remains constant, the probability of observing N events in our single-bin count is now given by a new binomial distribution with this new value of CN inserted in place of N in Eq.(3). We can again derive the Poisson, but the expectation value will now be $\lambda = CNp$.

Furthermore, the variance of the Poisson distribution can be shown to be equal to its expectation value, namely $\sigma^2 = \lambda$. A measure of how localized the Poisson is for a given λ w.r.t. the n-axis in units of the expectation value, is then

$$\frac{(\mu+\sigma)-\mu}{\mu} = \frac{\sigma}{\mu} = \frac{\sqrt{\lambda}}{\lambda} = \frac{1}{\sqrt{\lambda}} \tag{12}$$

$$\Rightarrow \lim_{\mu\to\infty} \frac{\sigma}{\mu} = 0 \tag{13}$$

Which shows that the Poisson becomes more localized as the expectation value increases, if we measure its spread in units of the expectation value itself.

Now, if we assume that the ratio of expected signal to background events remains the same as we increase the luminosity[2], we get

$$\frac{s}{b} = k \tag{14}$$

[2] Which is the same assumption we used in deriving the Poisson

Where k is a constant. From this it follows that

$$(s+b) = b(1+k) \propto N \propto L \tag{15}$$

and thus

$$\begin{aligned} E[\mathrm{P}_{\mathrm{H}_0}(n)] &= b \propto L \\ E[\mathrm{P}_{\mathrm{H}_1}(n)] &= b(1+k) \propto L \end{aligned}$$

Which shows that the expected value of the probability distribution of H_0 and the H_1 remains separated by a fixed distance of k (in units of b).

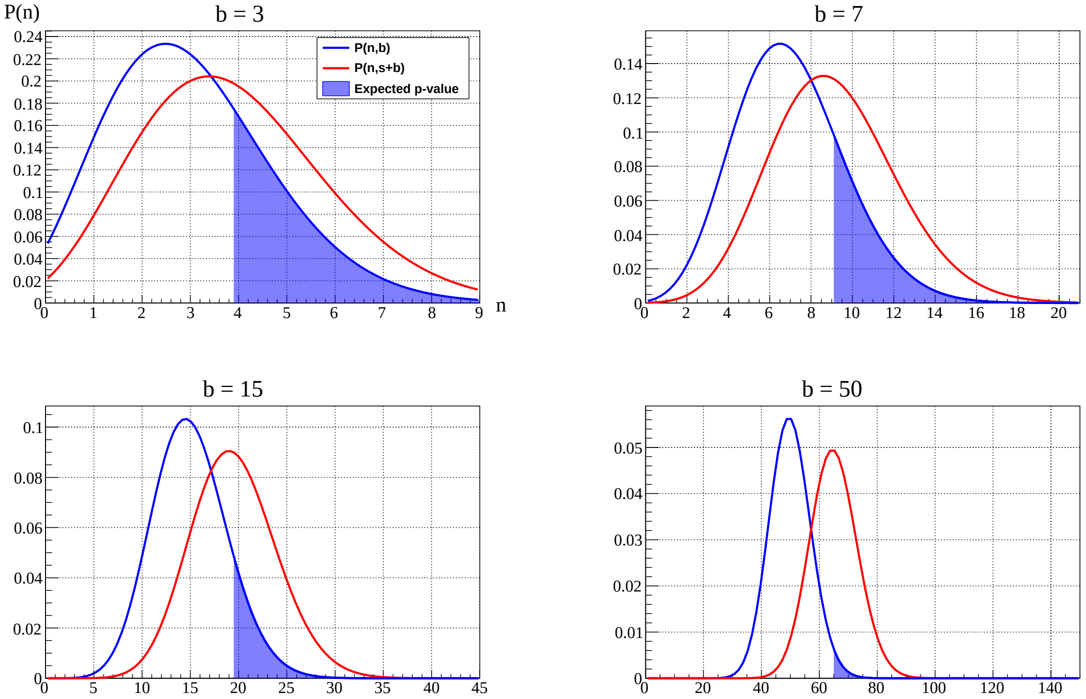

Figure (1) Demonstration of how an increase in the luminosity/expectation value affects the overlap of the PMF's of H_0 and H_1. The red area is the expected p-value, from which the expected significance can be calculated. A smaller p-value leads to a higher significance. Note that the Poisson is a discrete distribution, and the continuous curves shown in the figure are the results of replacing the factorial $n!$ by the gamma function.

Therefore, as the luminosity increases, $\mathrm{P}_{\mathrm{H}_0}(n)$ and $\mathrm{P}_{\mathrm{H}_1}(n)$ become more and more localized, while their separation remains fixed. This results in smaller p-values at higher luminosities, which in turn results in a higher significance. The significance therefore increases monotonically with luminosity, as can be seen in Fig.(4).

3 Analysis Walk-Through

3.1 Finding the optimal mass window

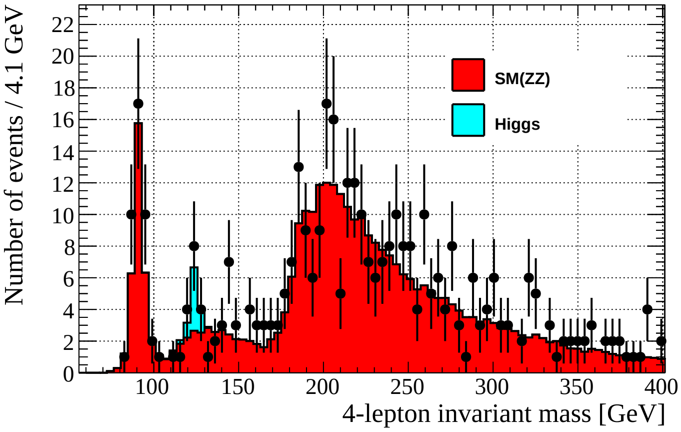

Figure (2) 4-Lepton mass distribution of our search. Plotted using a bin size of 4.1 GeV.

Figure (2) shows the data sample for our analysis. In an ideal world, we would be able to filter out all events which didn't have the properties of a Higgs boson with a 4-muon final state. However, due to suboptimal detector efficiencies, imperfect triggers etc., there will inevitably be some events that pass all our selection criteria and end up in our cut which is optimized to exclude as many such background events as possible.

One can however try to estimate this background, as well as the predicted signal yield via advanced Monte-Carlo simulations. The estimate for the standard model background and expected signal yield are shown in red and blue in Fig.(2), respectively. The actual observed bin counts are shown in black with their corresponding 68%CL uncertainties.

A final cut can be made to increase the signal to background ratio, namely the mass range which we wish to consider.

For a given mass window, the bin counts will be Poisson distributed. The expected background will be Poisson distributed with expectation value b and the signal+background will have the same distribution, but with expected value $s + b$. [3]

Now, the expected significance of a given mass window will be a measure of how well the two hypotheses are separated. The higher the significance, the more separated they are, which means that the hypothesis test will have a higher sensitivity.

We can then vary the mass window around the signal peak, where $\frac{s}{b}$ is assumed to have its maximum, and plot this significance as a function of the full width of the mass window. This is shown in Figure (3).

[3] b and $s + b$ are found by summing the bin counts of the expected background or signal+background distributions of all the bins in the mass window.

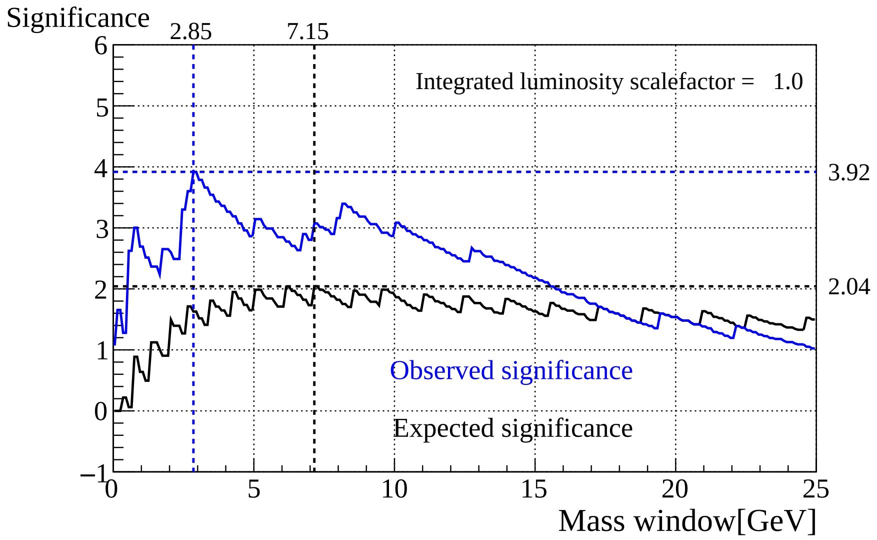

Figure (3) Observed and expected significance as a function of the size of our single-bin. The x-axis corresponds to various full widths of this single-bin, which is centered on the signal peak, viz. 125 GeV. Significance is plotted along the y-axis with observed- and expected significance plotted in blue and black respectively.

	Expected	Observed
Optimal width	7.15 GeV	2.85 σ
Corresponding Z	2.04 GeV	3.92 σ

Table (1) Optimal mass windows and their corresponding significance.

The window full width which resulted in the highest expected significance, was found to be 7.15 GeV. This is the window we will use throughout the analysis, as it yields the best sensitivity of our hypothesis test.

Now, we have also plotted and found the optimal observed significance as a function of the window width. The observed significance is a measure of how compatible our data are with $P_{H0}(n)$. It should be ignored when constructing the optimal mass window. The optimal mass window found via the maximum expected significance is that window which gives our test of the background- vs signal+background hypotheses the greatest sensitivity, and one would therefore expect to see the greatest discrepancy in data between these two hypotheses in this window. Picking a window bases on the observed significance would only reduce the sensitivity of the test or in best case leave it unaltered in the case that it happens to coincide with the maximum expected significance window.

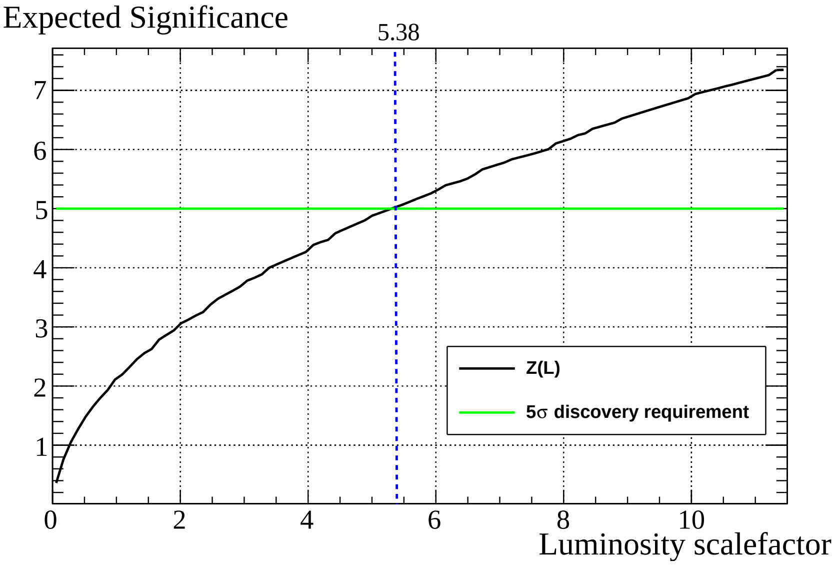

Figure (4) Expected significance as a function of the luminosity scalefactor. If H_1 is true, we expect to need a 5.38 times higher luminosity in order to claim discovery, which can be achieved by increasing the instantaneous luminosity or the data acquisition period.

The motivation given in section 2.1 for why significance increases with luminosity helps us understand the shape of the significance plot in Fig.(4). One might initially think that picking the mass window where the ratio $\frac{s}{b}$ is the greatest would yield the highest significance. This would likely corresponds to picking the bin at the mass peak, i.e. the 125 GeV bin. However, this would result in s and b both being small, and the overlap of the distributions $p(n, b)$ and $p(n, s + b)$ would be large, as illustrated in Fig.(1). As we start increasing the mass window, the ratio $\frac{s}{b}$ will go down, but this will be compensated for by the fact that s and b are both larger, i.e. the luminosity of our mass window is larger and the two Poisson distributions will be more localized as illustrated in Fig.(1). There is therefore a "tug war" between the ratio $\frac{s}{b}$ and the integrated luminosity of our mass window. The point at which this "tug war" results in the highest expected significance, can be seen from Fig.(3) to be at a mass window of full width 7.15 GeV around 125 GeV.

As we keep increasing the mass window from here, s will mostly remain constant while b keeps increasing. The distributions $P(n, b)$ and $P(n, s + b)$ will therefore start overlapping, and approach the exact same distribution, namely $\mathcal{N}(n; b, \sqrt{b})$ as $b \to \infty$ with s constant, resulting in a significance which approaches 0 in this limit.

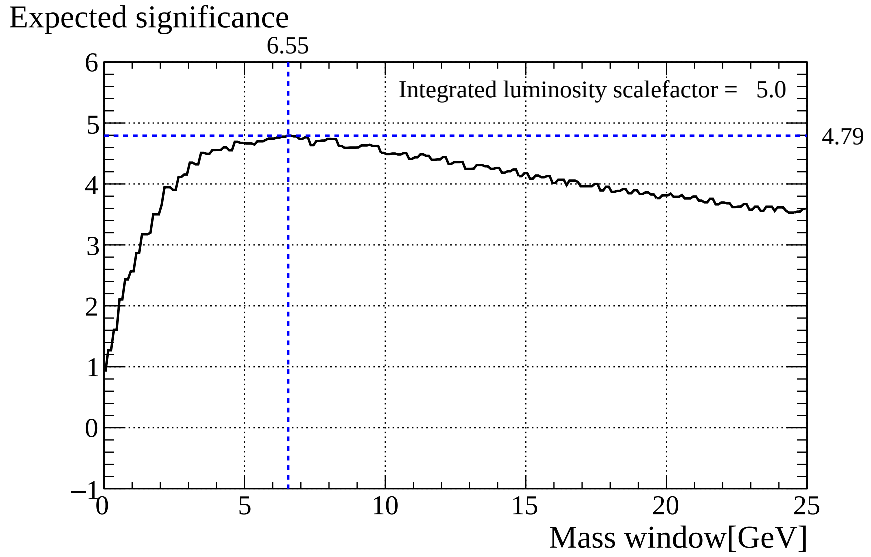

Figure (5)

As seen in Fig.(5), The optimal mass window which maximizes the expected significance for a luminosity scale factor of 5 was found to be 6.55 GeV, which yields an expected significance of $Z_{5lum} = 4.79\sigma$. Comparing this to the significance at a luminosity scale factor of one, namely $Z_{1lum} = 2.04\sigma$, we see that this luminosity provides a far more sensitive test. In fact, it is close to the 5 σ discovery limit. The exact amount we need to increase the luminosity in order to expect to be able to make a discovery was found in Fig.(5) to be at a factor of 5.38 times higher luminosity.

3.2 Estimating MC scalefactors via maximum likelihood

In order to determine the best fit of the simulated background and signal distribution, we turn to the likelihood function. For this purpose, it is given by

$$L(\alpha, \mu; \mathbf{n}) = \prod_{i=1}^{B} P(n_i; \alpha b + \mu s) \tag{16}$$

Where $P(n_i; \alpha b + \mu s)$ is the Poisson distribution with expectation value $\alpha b + \mu s$ evaluated at n_i and B are the number of bins.
α and μ are variables which globally[4] scale the background and signal+background estimates, respectively.
The likelihood function is a measure for how probable the given data, $\mathbf{n}$, are for various values of the shape parameters, which in our case are α and μ. It therefore makes sense to use the values of α and μ which maximizes Eq.(16) as our estimators for the scalefactors.

[4] By "globally" we mean that the estimate in every bin gets multiplied by the same scalefactor.

Because the logarithm is a strictly monotone function, this is equivalent to finding the maximum of $\ln L(\alpha, \mu; \mathbf{n})$.

Let's first consider the scalefactor for the background only, namely α. In this case, we let the product in Eq.(16) run over bins with a very low or non-existent signal yield. As Fig.(2) shows, this corresponds to the "sidebands" below and above the resonance peak. In our analysis, the upper sideband corresponding to $150 \leq m_H \leq 400\,\text{GeV}$ is used.

The estimate for the background scalefactor is then given by the value of α which satisfies

$$\frac{\partial \ln L(\alpha, \mu = 0; \mathbf{n})}{\partial \alpha} = 0 \tag{17}$$

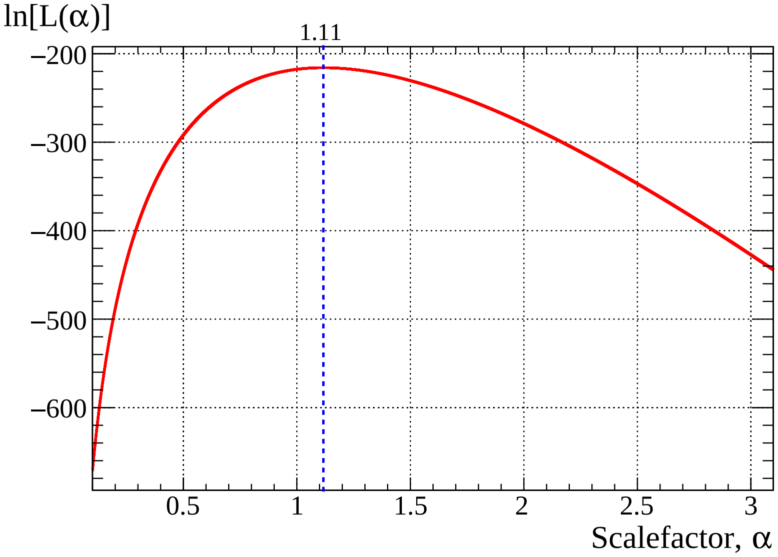

Figure (6) Log likelihood plotted against the background scalefactor, α.

Which we will denote as α_{max}. As seen in Fig.(6), this was found to be $\alpha_{max} = 1.11$.

Under the assumption that the likelihood is a Gaussian, which is the case in the limit of an infinite sample size[1], we have

$$\ln L(\alpha_{max}) - \ln L(\alpha) = \frac{1}{2} \frac{(\alpha - \alpha_{max})^2}{\sigma^2_{\alpha_{max}}} \tag{18}$$

This equality only holds approximately when the likelihood deviates from a Gaussian, but the approximation is surprisingly good even for small sample sizes.

From the above equation we see that when the RHS equals $\frac{1}{2}$, α will be located a distance of one standard deviation away from the maximum, namely

$$\ln L(\alpha_{max}) - \ln L(\alpha) = \frac{1}{2} \iff |\alpha - \alpha_{max}| = \sigma_{\alpha_{max}} \tag{19}$$

However, when the likelihood is not a Gaussian, the values of α for which Eq.(19) hold cannot be thought of as corresponding to being one standard deviation away from α_{max}, since the two values will be different due to an asymmetric distribution.
Instead, these two values, α_{low} and α_{up}, can be thought of as the values which bracket the 68% confidence interval, i.e. where we would expect α_{max} to land 68% of the time when repeating the experiment a large number of times.

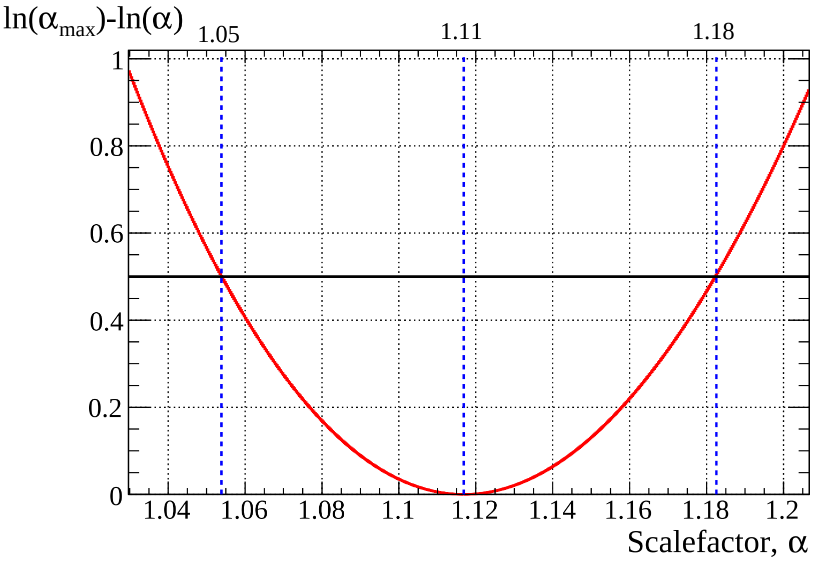

Figure (7)

As can be seen in Fig.7, this confidence interval was found to be

$$\alpha_{max}{}^{\alpha_{up}-\alpha_{max}}_{\alpha_{max}-\alpha_{low}} = \alpha_{max}{}^{+\Delta\alpha}_{-\Delta\alpha} = 1.11^{+0.07}_{-0.06} \tag{20}$$

Now that the optimal scale factor is found, we can re-count the number of background events predicted by the MC simulation in the optimal mass window around the Higgs resonance peak (namely 7.15 GeV) and multiply it by α_{max} to get our improved estimate of the background in this window. The uncertainties will simply be $-\Delta b = -\Delta\alpha \cdot b$ and $+\Delta b = +\Delta\alpha \cdot b$:

$$b_{scaled} = 1.11^{+0.07}_{-0.06} \cdot 4.83 = 5.39^{+0.32}_{-0.30} \tag{21}$$

3.2.1 Two-parameter case

Similarly to how $L(\alpha|\mathbf{n})$ approaches a 1D Gaussian in the limit of infinitely many measurements, $L(\alpha, \mu|\mathbf{n})$ approaches a 2D Gaussian in this limit.
We can proceed in a similar manner, and plot $\ln L(\alpha, \mu; \mathbf{n})$, as shown in Fig.(8).
From this, the optimal values were found to be

$$\alpha_{max} = 1.11, \;\; \mu_{max} = 1.28 \tag{22}$$

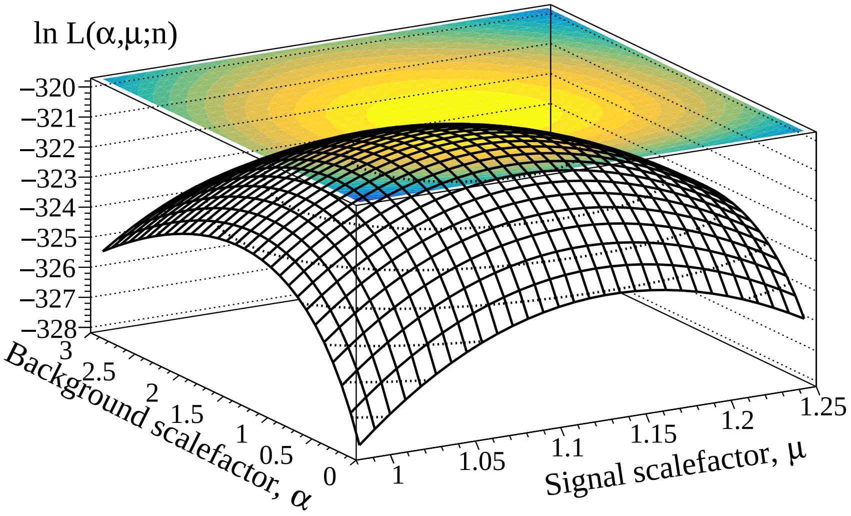

Figure (8)

To find the uncertainties, we consider the contour plot of $\ln L_{max} - \ln L$, as shown in Fig.(9). Similar to before, the uncertainties can be read from the first $\ln L_{max} - \ln L = \frac{1}{2}$ contour. They were found to be

$$\alpha_{max} = 1.11^{+0.058}_{-0.056}, \quad \mu_{max} = 1.28^{+0.64}_{-0.54} \tag{23}$$

We had to restricted the range of α to get a more readable plot, as the "standard deviation" of α is far smaller than that of μ, resulting in the Gaussian being elongated along the α axis. The skewness of the contours indicates how correlated α and μ are. When they are maximally correlated at $\rho^2_{\alpha,\mu} = 1$, they will be at 45 degree angles with the axes. In the case that they are independent, this angle will be zero. As seen, the correlation is quite small in our case.

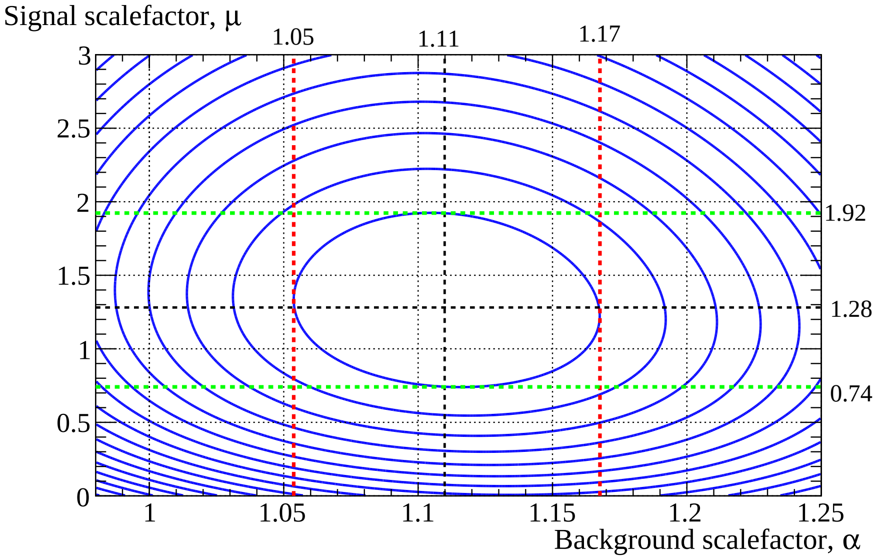

Figure (9) The center of the black dotted line indicates the minimum of $\ln L_{max} - \ln L$. The innermost ellipse corresponds to the "1σ" contour. The uncertainties on α and μ can be found from the innermost ellipse in the manner illustrated in the figure.

4 The optimal test-statistic

A test-statistic is a function which reduces our data, $\mathbf{n}$, to a scalar, $t(\mathbf{n})$. Through the right choice of t, we can achieve a better separation of the H_0 and H_1 distributions than by using their distributions under e.g. a single-bin count.
A commonly chosen test statistic is:

$$t(\mathbf{n}, \mu) = -2 \ln \lambda = -2 \ln \left\{ \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})} \right\} \tag{24}$$

Which is a measure of how compatible the data are with the most probable value of the signal, as illustrated in the figure below.

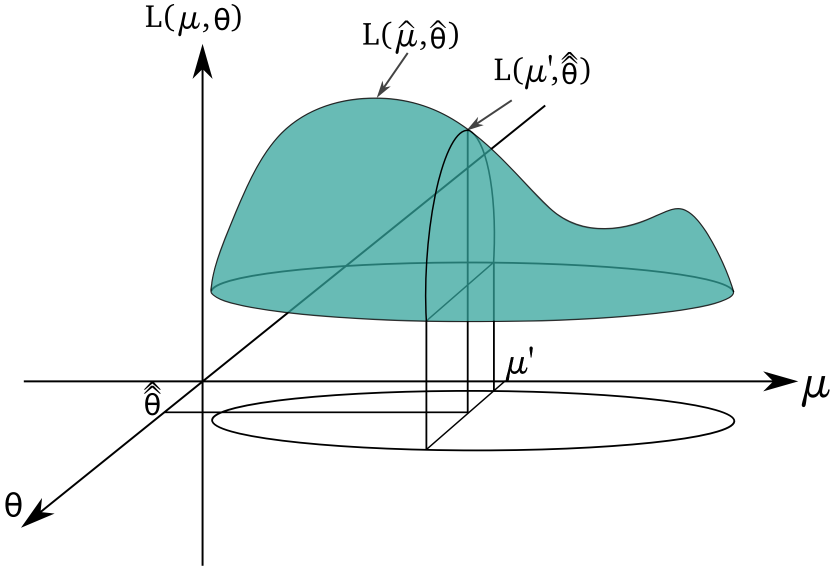

Figure (10) Illustration of the profile likelihood test-statistic. Here one would set $\mu = \mu'$ according to which of the two hypotheses one wishes to evaluate $t(\mathbf{n};\mu,\theta)$ under. Note that this depiction is not meant to be accurate, and typically one would only have one maximum (which our methods so far have relied on).

In our case however, we will follow the walk-through and use the following as our test statistic:

$$t(\mathbf{n}) = -2\ln\lambda = -2\ln\left\{\frac{L(\mathbf{n};\mu=1,\hat{\hat{\alpha}}_{\mu=1})}{L(\mathbf{n};\mu=0,\hat{\hat{\alpha}}_{\mu=0})}\right\} \tag{25}$$

According to the Neyman-Pearson lemma[1], when performing a hypothesis test between two hypotheses, this is the most *powerful* test statistic at a significance level α. In other words, it provides the best separation of the two distributions.

Okay, so now we have a test statistic under which the distributions H_0 and H_1 are more separated than under a single-bin count, n.

But in contrast to a single bin-count, for which we actually have the distributions of our hypotheses, namely the Poisson distributions in Eq.(5) and Eq.(6), we don't know the distributions under $t(\mathbf{n})$.

So what we do is we generate *pseudo-experiments*. For each "experiment", we loop through the bins and set number of observed events in each bin, n_i, to a random value. Depending on which hypothesis we're generating pseudo-data for, this random value is distributed according to a Poisson with expected value b_i or $b_i + s_i$, which are the expected number of simulated background or signal+background events in bin i, respectively.

After n_i has been set for each bin, we can now compute $t(\mathbf{n})$ given in Eq.(25) and increment the corresponding bin in Fig.11 by one. Repeatedly doing this many times generates the approximate unnormalized probability distributions of H_0 and H_1 under $t(\mathbf{n})$. One can then compute p-values etc., but must remember to normalize the integrals by the total bin counts for the distribution in question.

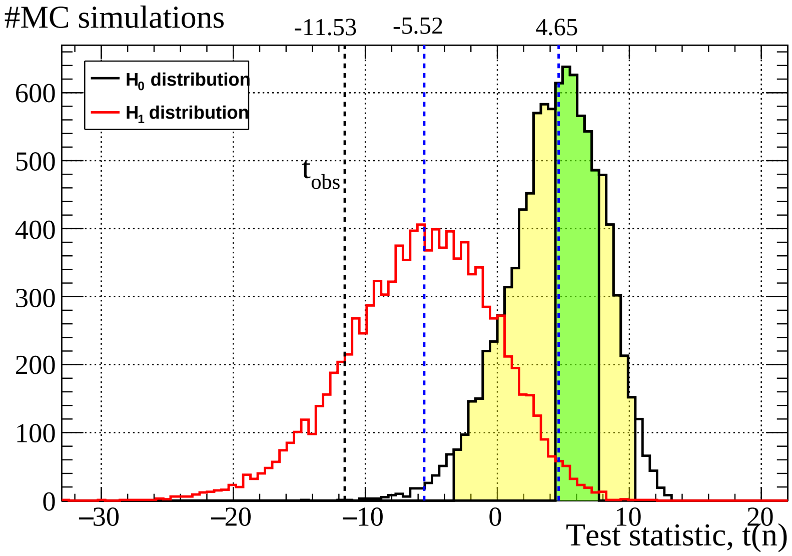

Figure (11) Distributions of the test statistic for the background (black) and signal plus background hypothesis (red) for 1e4 pseudo-experiments. The 68 and 95% confidence intervals for H_0 have been colored in green and yellow respectively. The observed value of the test statistic was $t(\mathbf{n}) = -11.53$. Furthermore, $median(H_0) = 4.65$ while for the background + signal hypothesis the median was $median(H_1) = -5.52$.

4.1 Discovery- and Exclusion aimed

Now that we have the background and signal+background distributions under the optimal test statistic, we can investigate whether we can claim discovery, and if not, what upper limits we can set on the Higgs mass, given our data.

4.1.1 Discovery

	median(H_0)	median(H_1)	Data
P-value	0.5314	0.0071	0.0003
Significance, Z	-0.079	2.4522	3.432

Table (2) P-values and significances under the background only hypothesis. The P-values are calculated by integrating the left tail of H_0 from the value specified in the first row. The significance, Z, is as always calculated via Eq.(11).

From Table (2), we see that observed significance is $Z = 3.43\ \sigma$, which is in fact higher than the expected significance of $Z = 2.45\ \sigma$, meaning that our data is more signal-like than we would expect for our luminosity. To claim a discovery, the observed data needs to have a significance of $Z = 5\sigma$ under the background hypothesis. This would correspond to a far lower P-value, namely

$p \leq 2.87 \cdot 10^{-7}$. Since the test-statistic in Eq. ((25)) already produces the lowest p-values possible for any data set, there is no gain to be made in altering it.
To meet the criteria of $Z = 5\sigma$ one must therefore gather more data, i.e. either increase the instantaneous luminosity of the accelerator, or increase the luminosity of the data set by simply gathering data over a longer period. This would yield a higher separation of the background only and background+signal hypotheses, as discussed in section 2.1, which would give a higher expected significance.
If the data now yield a significance of five or more, we can claim a discovery by rejecting the background hypothesis. The probability of falsely having rejected the background hypothesis (i.e. type I error) will be a mere $2.87 \cdot 10^{-7}$ or less.

4.1.2 Exclusion

	median(H_0)	median(H_1)	Data
CL|H_1	0.0213	0.528	0.8361

Table (3) Confidence levels calculated from the H_1 distribution. For a given value of the test statistic, the corresponding CL|H_1 is found by integrating the right tail of H_1 from this value. As seen, the observed confidence level is 0.84 which is far higher than 0.05, and thus we cannot exclude the H_1 hypothesis at a 95% confidence level.

Table (3) show the confidence levels under H_1 corresponding to t_{obs} and the two medians of our distributions, which can be seen in Fig.(11). What we're mainly interested in is the observed confidence level, namely CL(t_{obs})|H_1. If the data falls withing a predefined critical region of the $s + b$ distribution, one rejects the alternative hypothesis. This critical region is given by the portion of the tail pointing towards the expectation value of the background hypothesis that corresponds to an area of p_{crit}. Often this area is, as in our case, set to $p_{crit} = 0.05$.
This means that the chance of wrongly rejecting the alternative/s+b hypothesis will be 5% as this is how frequently t_{obs} would land in the critical region if the alternative hypothesis is true. When rejecting the background hypothesis, one typically says that it is rejected at a confidence level of $(1 - p)$%, i.e. 0.95% in our case.
We see from table(3) that $\text{CL}_{s+b}(t_{obs}) = 0.84 \geq 0.05$, and thus the signal+background hypothesis cannot be rejected.

4.1.3 Setting upper limits

As mentioned, when the probability of observing n_{obs} or less events under H_1 is less than 0.05, we can exclude H_1. This would, under $t(\mathbf{n})$, correspond to the case where t_{obs} lands in the 0.05 area of the right tail of H_1.
We found that t_{obs} did not land in this area, and thus at a signal cross-section scalefactor of $\mu = 1$, we cannot exclude H_1.
As mentioned in section 2.1, $\frac{s}{b}$ is assumed constant when testing the background against any specific alternative hypothesis, who's cross-section factor is determined this ratio. To determine upper limits, we can increase this ratio by increasing μ. This would correspond to shifting the

H_1 distribution if Fig.(11) to the left. One can then imagine that $\mathrm{CL}(t_{obs})|\mathrm{H}_1(\mu)$ would start decreasing, as the area of H_1 to the right of t_{obs} decreases as the distribution gets shifted to the left. In Fig.(12), we see how this area decreases as we shift the H_1 distribution left by increasing μ. We start at $\mu = 0$, which corresponds to the background-only hypothesis.

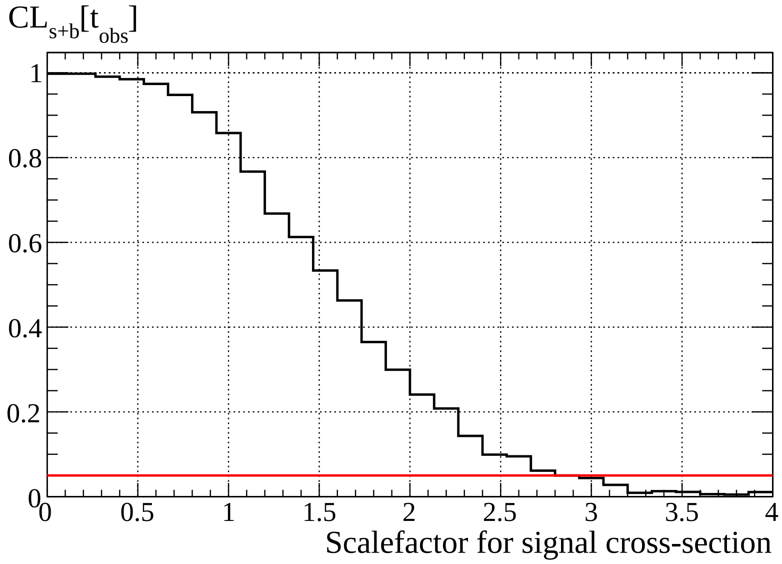

Figure (12) Each bin value has been calculated via 1e3 pseudo-experiments.

When the area above t_{obs} becomes 0.05 or less, we have find the value of value of μ, namely μ_{upper}, above which we can exclude any signal+background hypotheses which have a higher signal cross-section scalefactor. To determine more accurately discern this limit, we zoom in on the intersection of the 0.05 red line and our plot, as shown in Fig.(13).

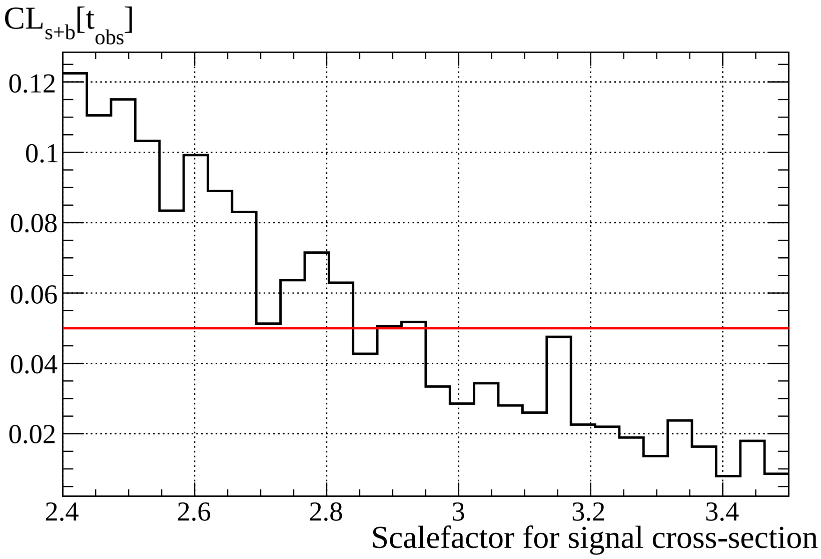

Figure (13) Each bin value has been calculated via 1e3 pseudo-experiments. $CL_{obs}|H_1$ is smaller than 0.05 above the bin corresponding to a cross-section scalefactor of 2.75. The fluctuations are quite large with such low numbers of MC simulations, so we choose to be conservative, and exclude only those hypotheses with a cross-section scalefactor of $\mu = 3$ or higher.

In conclusion, we set the upper limit on the 125 GeV Higgs cross section to be 3 times higher than what's given by the $\frac{s}{b}$ ratio of the optimal mass window of our given invariant 4-lepton mass distribution. Meaning that based on our data, given by Fig.(2), we can exclude any hypotheses predicting a 3 times or higher 125 GeV Higgs cross section than what was assumed when creating the simulated background and background+signal distributions.

References

[1] L. Lista. *Statistical Methods for Data Analysis in Particle Physics*. Vol. 941. Springer, 2016.

[2] M. Thomson. *Modern Particle Physics*. 3rd ed. Cambridge University Press, 2018.

Appendices

A Deriving the single-bin PMF

As motivated in section 2, the probability distribution $P(n_i)$ of the number of observed events in a specific bin is given by the binomial distribution;

$$P(n_i) = \binom{N}{n_i} p_i^{n_i} (1-p_i)^{N-n_i} \tag{26}$$

Where p_i is the probability for a given particle interaction to result in an event which ends up in bin i, and N is the total number of interactions in our experiment. Now, this binomial perfectly describes the probability distribution of n_i. However, it is very impractical to use due our inability to find N and p_i. What we can do however, is take the limit of Eq.(26) as $N \to \infty$ and with a fixed expectation value[5], which will yield an excellent approximation due to the enormous number of interactions and minuscule probability that a specific one ends up in our bin. The expectation value of the Binomial is given by :

$$\lambda = Np_i \Rightarrow p_i = \frac{\lambda}{N} \tag{27}$$

We thus have

$$\begin{aligned}
\lim_{N\to\infty} P(n_i) &= \lim_{N\to\infty} \left\{ \frac{N!}{(N-n_i)!n_i!} \left(\frac{\lambda}{N}\right)^{n_i} \left(1-\frac{\lambda}{N}\right)^{N-n_i} \right\} \\
&= \frac{\lambda^{n_i}}{n_i!} \lim_{N\to\infty} \left\{ \underbrace{\frac{N!}{(N-n_i)!N^{n_i}}}_{(i)} \underbrace{\left(1-\frac{\lambda}{N}\right)^{N}}_{(ii)} \underbrace{\left(1-\frac{\lambda}{N}\right)^{-n_i}}_{(iii)} \right\}
\end{aligned}$$

Now, using the property that

$$\lim_{x\to a} [f(x)g(x)] = \lim_{x\to a} f(x) \lim_{x\to a} g(x) \tag{28}$$

,which can shown by use of the epsilon-delta definition of a limit, we can find the limit of the individual terms. We recognize the limit of the second term as

$$(ii) = \lim_{N\to\infty} \left(1-\frac{\lambda}{N}\right)^{N} = e^{-\lambda} \tag{29}$$

The third term goes to 1, and so we're left with

$$\begin{aligned}
(i) = \lim_{N\to\infty} \frac{N!}{(N-n_i)!N^{n_i}} &= \lim_{N\to\infty} \frac{N(N-1)\cdot ... \cdot (N-n_i+1)}{N^{n_i}} \\
&= \lim_{N\to\infty} \frac{N}{N}\frac{N-1}{N} \cdot ... \cdot \frac{N-n_i+1}{N} \\
&= 1
\end{aligned} \tag{30}$$

[5] Which implies $p_i \to \infty$

Where in the last step we again used Eq.(28). Inserting for (i), (ii) and (iii) then gives $\text{app}_e q : dPsi_d a \lim_{N\to\infty} P(n_i) = \frac{\lambda^{n_i}}{n_i!}e^{-\lambda}$

B Deriving the likelihood function

We now wish to find the joint probability distribution of all the bins in our invariant mass histogram.

It is not immediately obvious why this in our case is given (to a very good approximation) by the product of the PMF's of the individual bins, as this is the case if and only if they are independent. This section is devoted to demonstrating why the approximation holds in the limit of infinitely many interactions and as the probabilities, p_i, go to zero.

Now, we can think of each particle interaction as having a fixed probability of producing an event which ends up in bin i, namely p_i. Let's therefore consider each bin as being a "category".

What we want to know is $P(n_1, ..., n_B)$, namely the probability of observing exactly n_1 events in bin 1, n_2 in bin 2, etc.

Let's also consider a last category, corresponding to the particle interactions which doesn't result in an event in any of our bins, with a corresponding probability

$$p_{outside} = 1 - \sum_{i=1}^{B} p_i \tag{31}$$

And

$$n_{outside} = N - \sum_{i=1}^{B} n_i \tag{32}$$

If there are N total interactions during our experiment, then the probability of getting $n_1, n_2, ..., n_B$ events in the corresponding bins and $N - \sum_{i=1}^{B} n_i$ events which doesn't land in any bin, is given by the number of ways to permute our outcomes[6] so as to give the same bin counts multiplied by the probability of any one of these unique permutations[7].

This probability is given by

$$\left(\prod_{i=1}^{B} p_i^{n_i}\right)(p_{outside})^{n_{outside}} = \left(\prod_{i=1}^{B} p_i^{n_i}\right)\left(1 - \sum_{i=1}^{B} p_i\right)^{N-\sum_{i=1}^{B} n_i} \tag{33}$$

And the number of unique permutations is given by

$$\binom{N}{n_1}\binom{N-n_1}{n_2}\cdot ... \cdot \binom{N-\sum_{i=1}^{B-2} n_i}{n_{B-1}}\binom{N-\sum_{i=1}^{B-1} n_i}{n_B} \tag{34}$$

Expanding this out in gory detail:

$$\frac{N!}{[N-n_1]!n_1!}\frac{[N-n_1]!}{[N-(n_1+n_2)]!n_2!}\cdot ... \cdot \frac{[N-\sum_{i=1}^{B-2} n_i]!}{[N-\sum_{i=1}^{B-1} n_i]!n_{B-1}!}\frac{[N-\sum_{i=1}^{B-1} n_i]!}{[N-\sum_{i=1}^{B} n_i]!n_B!} \tag{35}$$

[6] In time.

[7] Each permutation is assumed equally likely.

We see that we can cancel all the square brackets except the one in the last denominator, and we are therfore left with:

$$\text{number of unique permutations} = \frac{N!}{[N - \sum_{i=1}^{B} n_i]!} \prod_{i=1}^{B} \frac{1}{n_i!} \tag{36}$$

And so the exact joint PMF of our bins is given by

$$P(n_1, ..., n_B, n_{outside}) = \left\{ \frac{N!}{[N - \sum_{i=1}^{B} n_i]!} \prod_{i=1}^{B} \frac{1}{n_i!} \right\} \left\{ \left(\prod_{i=1}^{B} p_i^{n_i} \right) \left(1 - \sum_{i=1}^{B} p_i \right)^{N - \sum_{i=1}^{B} n_i} \right\} \tag{37}$$

Now using $\lambda_i = N p_i$ and inserting for $p_i = \frac{\lambda_i}{N}$:

$$\begin{aligned} P(n_1, ..., n_B, n_{outside}) &= \frac{N!}{[N - \sum_{i=1}^{B} n_i]!} \left(\prod_{i=1}^{B} \frac{1}{n_i!} \right) \left(\prod_{i=1}^{B} \left[\frac{\lambda_i}{N} \right]^{n_i} \right) \left(1 - \sum_{i=1}^{B} p_i \right)^{N - \sum_{i=1}^{B} n_i} \\ &= \left(\prod_{i=1}^{B} \frac{\lambda_i^{n_i}}{n_i!} \right) \frac{N!}{[N - \sum_{i=1}^{B} n_i]! N^B} \left(1 - \sum_{i=1}^{B} \frac{\lambda_i}{N} \right)^{N - \sum_{i=1}^{B} n_i} \\ &= \left(\prod_{i=1}^{B} \frac{\lambda_i^{n_i}}{n_i!} \right) \underbrace{\frac{N!}{[N - \sum_{i=1}^{B} n_i]! N^B}}_{(i)} \underbrace{\left(1 - \sum_{i=1}^{B} \frac{\lambda_i}{N} \right)^{N}}_{(ii)} \underbrace{\left(1 - \sum_{i=1}^{B} \frac{\lambda_i}{N} \right)^{- \sum_{i=1}^{B} n_i}}_{(iii)} \end{aligned} \tag{38}$$

By the same reasoning as in Eq.30, term (i) goes to 1 as $N \to \infty$. We also see that this is the case for (iii). However, the second term does not:

$$\begin{aligned} \lim_{N \to \infty} (ii) = \lim_{N \to \infty} \left(1 - \sum_{i=1}^{B} \frac{\lambda_i}{N} \right)^{N} &= \lim_{N \to \infty} \left(1 - \frac{\sum_{i=1}^{B} \lambda_i}{N} \right)^{N} \\ &= e^{- \sum_{i=1}^{B} \lambda_i} = \prod_{i=1}^{B} e^{-\lambda_i} \end{aligned} \tag{39}$$

We therefore get

$$\lim_{N \to \infty} P(n_1, ..., n_B, n_{outside}) = \left(\prod_{i=1}^{B} \frac{\lambda_i^{n_i}}{n_i!} \right) \prod_{i=1}^{B} e^{-\lambda_i} = \prod_{i=1}^{B} \frac{\lambda_i^{n_i}}{n_i!} e^{-\lambda_i} \tag{40}$$

Dropping the superfluous variable $n_{outside}$ gives us
$\text{app}_e q : dPsi_d a \lim_{N \to \infty} P(n_1, ..., n_B) = \prod_{i=1}^{B} \frac{\lambda_i^{n_i}}{n_i!} e^{-\lambda_i}$
Which shows that in the limit of infinitely many interactions, the joint PMF of the bins in our histogram is given by the product of the individual bin distributions.

This is what we will use as our "likelihood" function, namely

$$L(\mathbf{n};\lambda) = \prod_{i=1}^{B} \frac{\lambda_i^{n_i}}{n_i!} e^{-\lambda_i} \tag{41}$$

Note that it will only be approximately valid, as we don't actually have an infinite number of interactions. The approximation is however a very good one.