S1: Principals of Data Science - Coursework

William Knottenbelt, wdk24

December 9, 2023

Section A

Part (a)

The total probability density function is given by:

$$p(M; f, \lambda, \mu, \sigma) = fs(M; \mu, \sigma) + (1 - f)b(M; \lambda),$$

where s is the normal distribution and b is the exponential decay distribution, which is only non-zero for $M \geq 0$.

The condition for p to be properly normalised over $M \in [-\infty, +\infty]$ is:

$$\int_{-\infty}^{+\infty} p \, dM = 1,$$

To show that p is properly normalised we will use the identity:

$$\int_{-\infty}^{+\infty} e^{-ax^2} = \sqrt{\frac{\pi}{a}}, \tag{1}$$

We have:

$$\int_{-\infty}^{+\infty} p \, dM = f \int_{-\infty}^{+\infty} s \, dM + (1 - f) \int_{0}^{+\infty} b \, dM.$$

For the signal term, we have:

$$\int_{-\infty}^{+\infty} s \, dM = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M - \mu)^2}{2\sigma^2}\right) dM.$$

We use the substitution $x = M - \mu$ such that $dx = dM$ and the limits are the same since $x(M \to \pm\infty) \to \pm\infty$ for any finite μ. Then we have:

$$\int_{-\infty}^{+\infty} s \, dM = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx.$$

This integral has the form of identity (**??**) with $a = \frac{1}{2\sigma^2}$, hence we see that:

$$\int_{-\infty}^{+\infty} s\, dM = \frac{1}{\sigma\sqrt{2\pi}}\sqrt{2\sigma^2\pi} = 1.$$

For the background term we have:

$$\int_{0}^{+\infty} b\, dM = \int_{0}^{+\infty} \lambda e^{-\lambda M}\, dM = (-e^{-\lambda M})|_{0}^{+\infty} = 1$$

Hence, we see that:

$$\int_{-\infty}^{+\infty} p\, dM = f\int_{-\infty}^{+\infty} s\, dM + (1-f)\int_{0}^{+\infty} b\, dM = f + (1-f) = 1.$$

Thus, the normalisation condition is satisfied for p over $M \in [-\infty, +\infty]$.

Part (b)

To ensure that the signal distribution contributes the correct fraction f to the total probability over $M \in [\alpha, \beta]$, and the background distribution contributes $(1-f)$, we must first normalise each distribution individually over the range $M \in [\alpha, \beta]$, before performing a weighted sum to construct the total PDF. As each distribution is normalised separately, it is guaranteed that any weighted sum of the distributions will also be correctly normalised (provided that the weights sum to unity).

For M restricted to the range $M \in [\alpha, \beta]$, the signal distribution is defined:

$$s(M; \mu, \sigma) = \begin{cases} \frac{A}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right) & \text{for } \alpha < M \leq \beta \\ 0 & \text{otherwise.} \end{cases}$$

Where A is a normalisation factor, μ is the mean of the distribution and σ is the standard deviation. For s to be properly normalised we must have:

$$\int_{\alpha}^{\beta} s(M)\, dM = 1.$$

Hence:

$$1 = \int_{\alpha}^{\beta} s(M)\, dM = \int_{-\infty}^{\beta} s(M)\, dM - \int_{-\infty}^{\alpha} s(M)\, dM$$

$$= A\left(\int_{-\infty}^{\beta} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right) dM - \int_{-\infty}^{\alpha} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right) dM\right)$$

$$= A(F_{norm}(\beta) - F_{norm}(\alpha)),$$

where F_{norm} is the cumulative density function of the normal distribution defined over the range $[-\infty, +\infty]$, which is given by:

$$F_{norm}(X) = \Phi(\frac{X-\mu}{\sigma}).$$

Thus, the normalisation factor is:

$$A = \frac{1}{\Phi(\frac{\beta-\mu}{\sigma}) - \Phi(\frac{\alpha-\mu}{\sigma})}.$$

Assuming that $\alpha, \beta > 0$, the background distribution is defined:

$$b(M;\lambda) = \begin{cases} B\lambda e^{-\lambda M} & \text{for } \alpha < M \leq \beta, \\ 0 & \text{otherwise,} \end{cases}$$

where B is a normalisation factor and λ is the decay constant. For b to be properly normalised, we must have:

$$\begin{aligned} 1 = \int_\alpha^\beta b(M)\,dM &= \int_{-\infty}^\beta b(M)\,dM - \int_{-\infty}^\alpha b(M)\,dM \\ &= B\left(\int_0^\beta \lambda e^{-\lambda M}\,dM - \int_0^\alpha \lambda e^{-\lambda M}\,dM\right) \\ &= B(F_{exp}(\beta) - F_{exp}(\alpha)), \end{aligned}$$

where F_{exp} is the cumulative density function of the exponential decay distribution, which is given by:

$$F_{exp}(X) = \begin{cases} 1 - e^{-\lambda X} & \text{for } X \geq 0, \\ 0 & \text{for } X < 0. \end{cases}$$

Hence we have:

$$B = \frac{1}{F_{exp}(\beta) - F_{exp}(\alpha)} = \frac{1}{e^{-\lambda\alpha} - e^{-\lambda\beta}}.$$

Finally, the total probability density function, is a weighted sum of the individually normalised distributions:

$$p(M) = fs(M;\mu,\sigma) + (1-f)b(M;\lambda).$$

This is guaranteed to be properly normalised since:

$$\int_\alpha^\beta p\,dM = \int_\alpha^\beta fs + (1-f)b\,dM = f\int_\alpha^\beta s\,dM + (1-f)\int_\alpha^\beta b\,dM = f + (1-f) = 1.$$

Assuming $\alpha, \beta > 0$, the full expression is then:

$$p(M) = \frac{f}{\Phi(\frac{\beta-\mu}{\sigma}) - \Phi(\frac{\alpha-\mu}{\sigma})}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right) + (1-f)\frac{\lambda e^{-\lambda M}}{e^{-\lambda\alpha} - e^{-\lambda\beta}}.$$

Part (c)

We coded the normalised signal, background and total distributions using the `scipy.stats` package to implement the expressions found in part (b). Namely:

$$s(M;\mu,\sigma) = \frac{1}{\Phi(\frac{\beta-\mu}{\sigma}) - \Phi(\frac{\alpha-\mu}{\sigma})} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right),$$

$$b(M;\lambda) = \frac{\lambda e^{-\lambda M}}{e^{-\lambda\alpha} - e^{-\lambda\beta}},$$

$$p(M;\boldsymbol{\theta}) = fs(M;\mu,\sigma) + (1-f)b(M;\lambda),$$

where $\boldsymbol{\theta} = (f, \lambda, \mu, \sigma)$ represents the parameters.

Once I had coded the probability density functions, I wrote a script `solve_part_c.py`, which integrates the total probability density over $M \in [5, 5.6]$ using `scipy.integrate.quad` for 1000 random combinations of the parameters $(f, \lambda, \mu, \sigma)$, and checks that each integral comes out to unity. This script was run and all checks passed successfully.

Part (d)

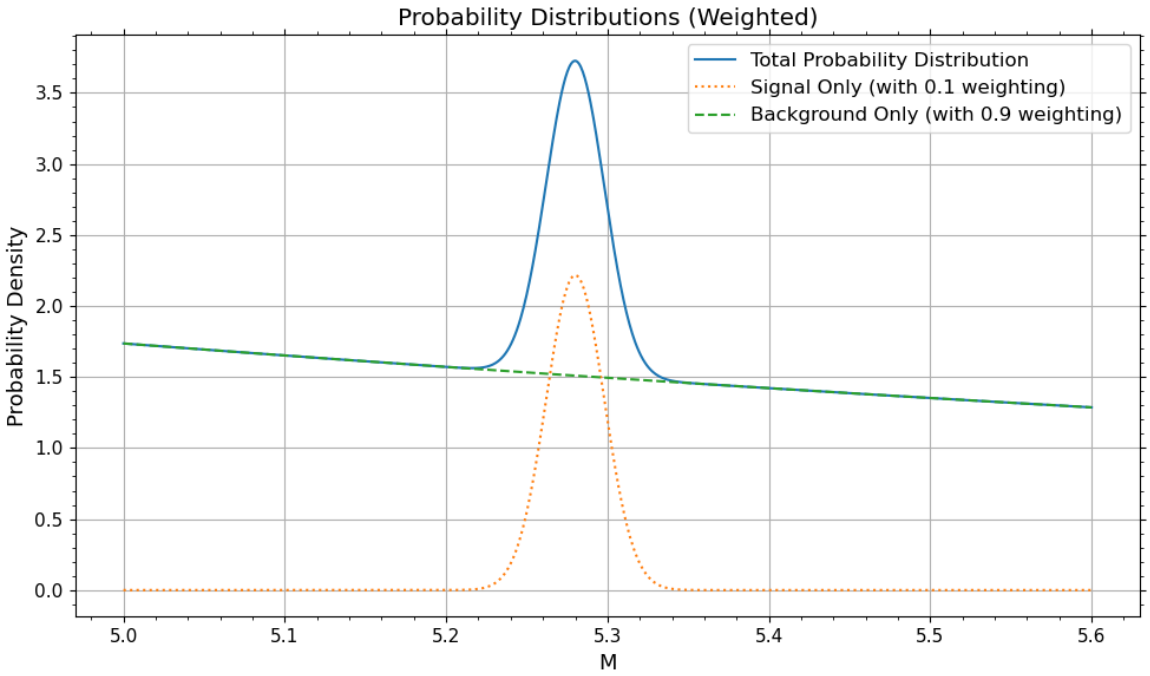

Figure 1: Visualisation of the total, signal and background probability density functions with the weights applied to the signal and background models such that only the total PDF is properly normalised.

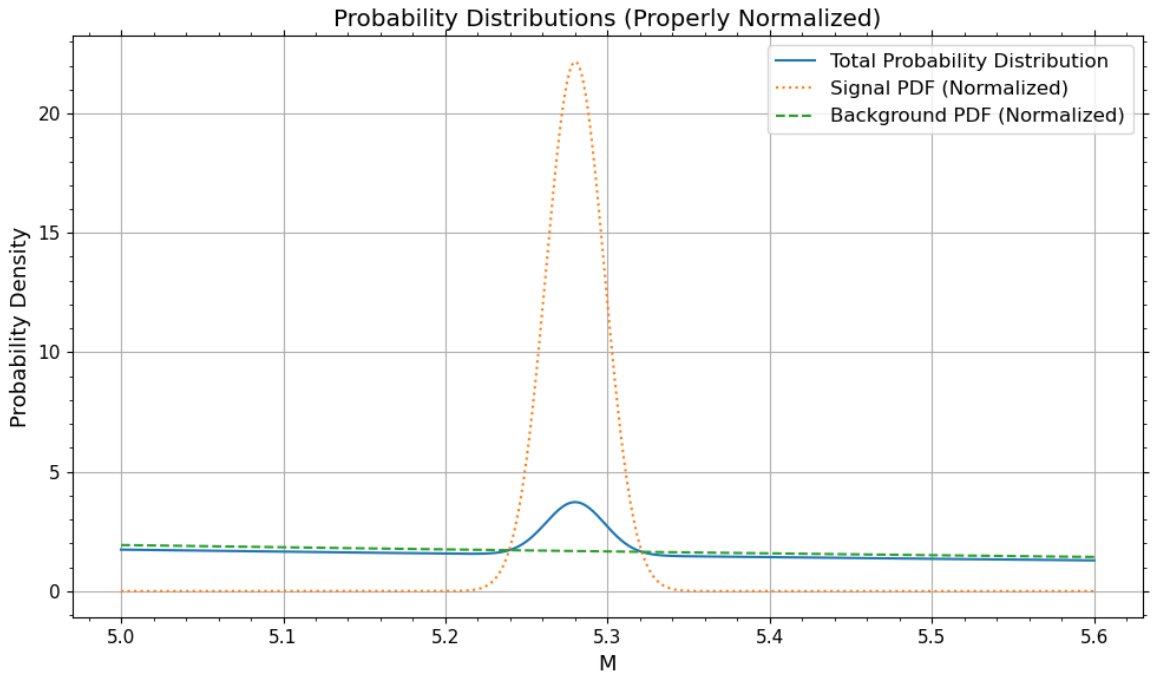

Figure 2: Visualisation of the total, signal and background probability density functions where each distribution is properly normalised.

We created two plots visualising the distributions with the true parameters $f = 0.1, \lambda = 0.5, \mu = 5.28, \sigma = 0.018$. Fig. **??** shows the distributions with the fractions f and $(1 - f)$ applied to the signal and background distributions. Fig. **??** shows the distributions properly normalised.

To re-generate these plots, run `solve_part_d.py` and find the results in `plots/`

Part (e)

To generate samples from the total PDF, we used the inverse CDF method, which works by generating uniform random numbers in [0, 1] and passing them into the inverse CDF (A.K.A percentage point function, PPF) to generate events distributed by the original PDF. Since we did not have access to the PPF of the total distribution, we implemented an algorithm which, for each event, chooses whether to generate from the signal-only or background-only distributions with probability 0.1 and 0.9, then generates the event using the PPF of the chosen model. This method is significantly more efficient than the accept-reject method, since there are no wasted samples. We used the PPFs of the normal and exponential distributions available in `scipy.stats`, which are not automatically normalised over $[\alpha, \beta]$, hence rather than generating uniform random numbers in the interval [0, 1], we generated from the interval $[F(\alpha), F(\beta)]$, where F is the CDF of the distribution we are generating from. This guarantees that data will only be generated in the desired range, and it will be distributed by the correctly normalised PDF (since only the relative probability of two points in the range $[\alpha, \beta]$ matters, which the same for a non-normalised model).

We generated a sample of 100K events, then fitted the total PDF to this data using maximum likelihood estimation of the parameters. That is, we estimated the parameters by minimising the

negative log likelihood:

$$l = -ln\left(\prod_i p(M_i;\boldsymbol{\theta})\right) = -\sum_i ln(p(M_i;\boldsymbol{\theta})) \tag{2}$$

For a given maximum likelihood estimate $\hat{\theta} \in \{\hat{f}, \hat{\lambda}, \hat{\mu}, \hat{\sigma}\}$, the variance of that estimate is given by the minimum variance bound (AKA the Cramér-Rao lower bound) in the asymptotic limit of large sample sizes ($N \to \infty$):

$$V(\hat{\theta}) = \left(NE\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right]\right)^{-1} = \left(NE\left[\frac{\partial^2 l}{\partial \theta^2}\right]\right)^{-1}, \tag{3}$$

where n is the sample size, $E[\cdot]$ is the expectation operator, l is the negative log likelihood. A useful property of maximum likelihood estimation is that the expectation of the double differential of the log likelihood is equal to the evaluation of the double differential at the estimated value. Hence:

$$V(\hat{\theta}) = \left(N\left[\frac{\partial^2 l}{\partial \theta^2}\right]_{\theta=\hat{\theta}}\right)^{-1} \tag{4}$$

Since we are dealing with a sample of 100K events, this asymptotic limit is a reasonable assumption. In this limit, the maximum likelihood estimate produces an estiamte which is normally distributed, unbiased, consistent and maximally efficient (minimum variance as discussed). This estimation was done using `iminuit`, using starting parameters as the true parameters shifted by an appropriate random shift. We obtained the following estimates:

$\hat{f} = 0.0998 \pm 0.0016, \quad \hat{\lambda} = 0.470 \pm 0.019, \quad \hat{\mu} = 5.27980 \pm 0.00033, \quad \hat{\sigma} = 17.99 \times 10^{-3} \pm 0.32 \times 10^{-3}.$

For all estimates, the true values of the parameters lie within the uncertainties of the estimates. We then binned and plotted the sample, and overlaid the fitted PDF, as can be seen in Fig. **??**.

Section B

Introduction

In section B we consider the theory, implementation and results of part (f) and (g). The goal of part (f) was to estimate the size of a dataset sampled from a 'signal plus background' model, which would lead to a 'discovery' of the signal at least 90% of the time when we perform a hypothesis test on it. Similarly, the goal of part (g) was to estimate the size of a dataset sampled from a 'two signals plus background' model, which is needed to discover the two distinct signals at least 90% of the time. For each part, we investigated 50 dataset sizes in an appropriate range found by trial and error. For each dataset size we repeated the following procedure n times: generate a dataset, fit the null (H_0) and alternate (H_1) hypothesis distributions using maximum likelihood estimation, and perform a Neyman-Pearson hypothesis test. The number of discoveries for each dataset size is binomially distributed, which allows us to estimate the probability of discovery and uncertainty on this estimate from our experiment. We then fit a third order polynomial to approximate the relationship between p and N in the range of interest. This allows us to predict the critical dataset sizes, as well as their uncertainties using a Monte Carlo simulation.

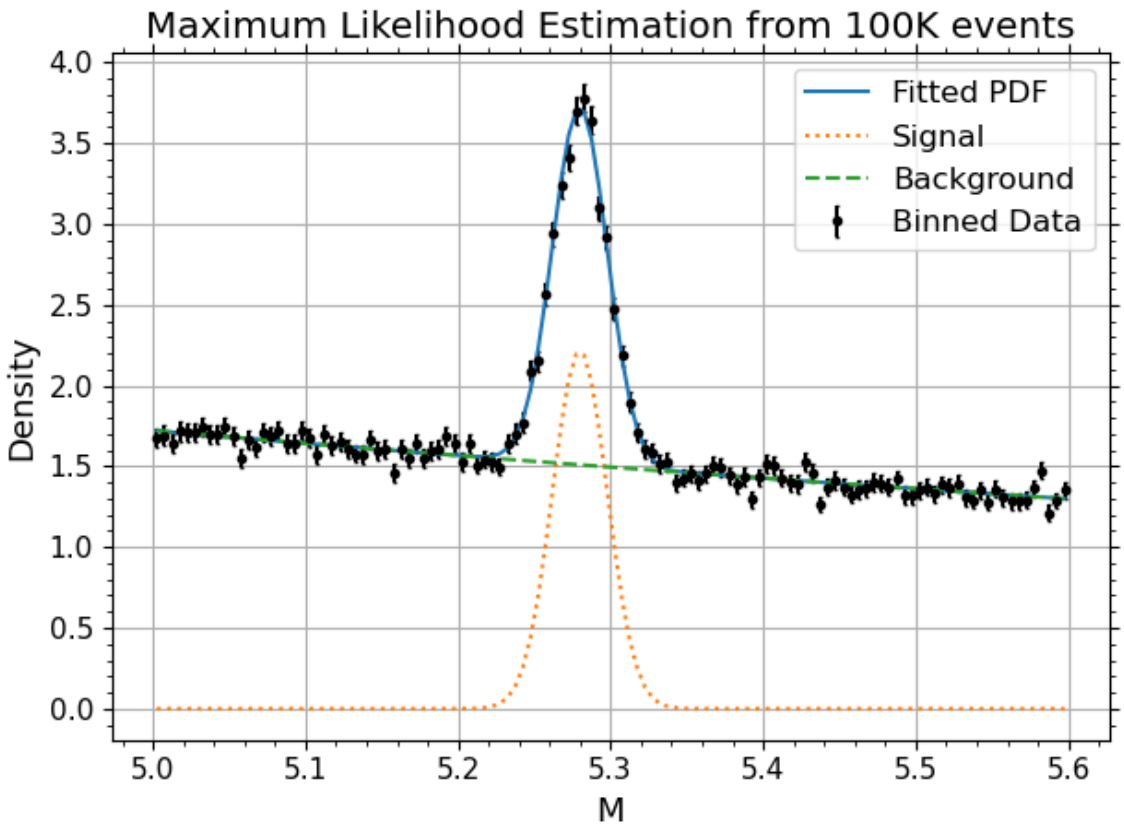

Figure 3: Visualisation of the our 100K sample generated from the true PDF, overlaid by the fitted PDFs of the total, signal and background distributions, whose parameters were estimated via the maximum likelihood method

Methodology

Neyman-Pearson Test

The power, $1-\beta$, of a hypothesis test is the probability of correctly rejecting the null hypothesis when it is false. The Neyman-Pearson Lemma states that the test statistic which maximises the power is given by the ratio of the likelihoods under each hypothesis:

$$T = -2ln\left(\frac{L(\vec{X}|H_0)}{L(\vec{X}|H_1)}\right) \tag{5}$$

According to Wilk's theorem, in the asymptotic limit of large sample sizes $N \to \infty$, this test statistic under the null hypothesis is χ^2 distributed with 1 degree of freedom. We assume that this asymptotic limit is valid, which allows us to find the p value of the hypothesis test from:

$$p = 1 - F_{\chi^2,1}(T),$$

where $F_{\chi^2,1}$ is the cumulative density function of the χ^2 distribution with 1 degree of freedom. Since we cannot have a 'negative' presence of signal (ie. the fraction of signal, $f \geq 0$), this is a one-sided test, and we calculate the significance via the formula:

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

where Φ is the cumulative density function of the normal distribution, and p is the p-value of the test. The threshold for a 'discovery' is $Z \geq 5$

Binomial Distribution

In part (f) and (g), on each dataset size, N_i, we conduct hypothesis tests on n generated datasets and find k_i discoveries. Each hypothesis test has exactly two mutually exclusive outcomes (discovery or no discovery) and the probability of discovery is constant for constant dataset size, hence the experiment can be thought of as a series of Bernoulli trials, and the number of discoveries is binomially distributed. Let p_i be the true probability of discovery for dataset size N_i. The mean of the binomial distribution is $\mu = E[k_i] = p_i n$ and the variance is $V(k_i) = np_i(1-p_i)$.

We can estimate the probability of discovery as the frequency of discovery in the experiment:

$$\hat{p}_i = \frac{k_i}{n} \tag{6}$$

The variance of this estimate is $V(\hat{p}_i) = V(\frac{k_i}{n}) = \frac{V(k_i)}{n^2} = \frac{p_i(1-p_i)}{n}$, hence we can estimate the uncertainty using:

$$\hat{\sigma_i} = \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}} \tag{7}$$

In general, an estimate, $\hat{\theta}$, is consistent if $\hat{\theta} \to \theta$ as $n \to \infty$. The law of large numbers states that the frequency of an outcome $h = \frac{k}{n}$ converges to the probability of that outcome in the limit of infinite experiments: $\frac{k}{n} \to p$ as $n \to \infty$. Hence, our estimate $\hat{p}_i$ is consistent: $\hat{p}_i = \frac{k_i}{n} \to p_i$ as $n \to \infty$

An estimate, $\hat{\theta}$, is unbiased if $E[\hat{\theta}] = \theta$. The expectation of our estimate is $E[\hat{p}_i] = E[\frac{k_i}{n}] = \frac{E[k_i]}{n} = \frac{\mu}{n} = p_i$, hence this estimate is unbiased.

Since $\hat{p}_i \to p_i$ as $n \to \infty$, we see that also $\hat{\sigma_i} = \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}} \to \sqrt{\frac{p_i(1-p_i)}{n}} = \sigma_i$ as $n \to \infty$, hence our uncertainty estimate is also consistent (although it is biased).

According to the central limit theorem, our estimate $\hat{p}_i = \frac{k_i}{n}$ is normally distributed in the limit of large n.

Least Squares Fitting

Once we have collected our data $\{(N_i, \hat{p}_i \pm \hat{\sigma_i})\}$, we need to predict the size of the dataset at which the probability of discovery is 90%, N_{90}. To make this prediction, we can fit some model, $f(N;\boldsymbol{\theta})$, to our data by minimising the 'chi-squared' with respect to the parameters of the model $\boldsymbol{\theta}$, given by:

$$\chi^2 = \sum_i^{50} \frac{(\hat{p}_i - f(N_i|\boldsymbol{\theta}))^2}{\hat{\sigma_i}^2}. \tag{8}$$

If our probability estimates $\hat{p}_i$ are normally distributed with mean $f(N_i|\hat{\boldsymbol{\theta}})$ and standard deviation $\hat{\sigma_i}$, then χ^2 is proportional to the negative log likelihood **??**. Hence, under the assumption that $\hat{p}_i$ is normally distributed (which is valid in the limit of large n), least squares fitting is equivalent to maximum likelihood estimation. As discussed in part (e), if our sample size of $\hat{p}_i$ is large, then the maximum likelihood estimates are maximally efficient, and so we can estimate the errors of the fitted model parameters using the minimum variance bound **??**. However, the asymptotic limit may not be valid as we are dealing with only 50 data points.

Additionally, if this assumption is valid, then the fitted model should make predictions that are within the uncertainties of the measurements 68.3% of the time (since one standard deviation of the normal distribution contains 68.3% of the total probability).

χ^2 test

A useful method to evaluate goodness-of-fit is a χ^2 test. We calculate χ^2 using equation 4. If our probability estimates $\hat{p}_i$ are normally distributed with mean $f(N_i|\hat{\boldsymbol{\theta}})$ and standard deviation $\hat{\sigma}_i$, then 4 is χ^2 distributed with $k = n_{obs} - n_{params}$ degrees of freedom, where n_{obs} is the number of observations and n_{params} is the number of parameters of the fitted model. The expectation value of the χ^2 distribution is k, hence on average, we expect $\chi^2/d.o.f \approx 1$. The p-value for this test is the probability that we obtain a χ^2 as large or larger than the value we measured. This can obtained using the formula:

$$p = 1 - F_{\chi^2,k}(\chi^2), \tag{9}$$

where $F_{\chi^2,k}$ is the cumulative density function of the χ^2 distribution with k degrees of freedom.

The higher the p-value, the better the fitted model agrees with the data. Conventionally, a p-value of less than 5% is considered statistically significant, indicating that the model is not a good fit for the data. A p-value which is too high may indicate that the model has over-fitted to the data.

Monte Carlo Simulation

Once we have a model $f(N|\hat{\boldsymbol{\theta}})$ with parameter estimates $\hat{\boldsymbol{\theta}}$ and uncertainties on those parameter estimates, we can estimate the critical dataset size by finding N_{90} for which $f(N_{90}|\hat{\boldsymbol{\theta}}) = 0.9$.

Since the parameter estimates for our model were obtained by least squares fitting, which is equivalent to maximum likelihood estimation in the limit of large n, the parameter estimates are normally distributed (assuming this limit is valid). Thus, we can estimate the uncertainty of N_{90} via a monte carlo simulation, in which we sample parameters θ_i from a normal distribution with mean $\hat{\theta}_i$ and standard deviation being the error on $\hat{\theta}_i$, then solve for N_{90}. This gives us a sample of measurements of N_{90}, from which we can estimate the error on N_{90} as the sample standard deviation.

Part (f)

We found the range of dataset sizes over which to conduct the experiment via a trial and error approach. We repeatedly generated datasets for a range of sizes, initially on a logarithmic scale, and conducted hypothesis tests on them. This provided loose estimates of the probability of discovery, and allowed us to find the range over which the probability increases from 0% to 100%. After repeating this on different ranges, we determined that the desirable range to perform the experiment is $N \in [500, 900]$. This was chosen as it appeared to contain the true N_{90}, and it is narrow enough to make a precise measurement of N_{90}.

For 50 equally-spaced sizes in this range, we generated $n = 500$ example datasets. For each dataset, we conducted a hypothesis test for the existence of a signal. To construct the null (H_0) and alternate (H_1) hypotheses, we used the procedure detailed in part (e), wherein we fitted a 'background-only model' (for H_0) and a 'signal plus background' model (for H_1) to the dataset using maximum likelihood estimation. We chose to use the Neyman-Pearson test statistic 1 as it is the most powerful test, hence it maximises the probability of discovering the signal. This allowed us to estimate the probability of discovery at each size, leaving us with the dataset $\{(N_i, \hat{p}_i \pm \hat{\sigma}_i) | i = 1, ..., 50\}$.

We fitted a third order polynomial to this dataset, calculated the coverage and performed a χ^2 test. The coverage was calculated to be 70% (proportion of model predictions which lie within uncertainty of true value), the $\chi^2/d.o.f$ was calculated to be , and the p-value of the test was .

The coverage (proportion of model predictions which lie within uncertainty of true value) was calculated to be 70%, which is extremely close to 68.3%, suggesting that our Gaussian assumption for $\hat{p}_i$ is correct, and that our fitted model does capture the true relationship between P and N. The $\chi^2/d.o.f$ was calculated to be , is close to 1, indicating a strong agreement between the model and the data. The p-value of the test was 21.9%, which suggests that the model is a good fit of the data, and approximates the relationship between p and N reasonably well over this range. The p-value is also not too high as to suggest a suspiciously strong agreement between the model and the data.

By performing a Monte Carlo simulation to estimate N_{90} and its uncertainty, we yielded the result:

$$N_{90} = 603 \pm 3$$

However, the error on this result may not be trustworthy, as it was derived by taking the fitted parameter uncertainties to be the minimum variance bound. Since this is only valid in the asymptotic limit, and our sample size was only 50, it is not clear that this assumption is reasonable.

In Fig. 1, we plotted the probability of discovery against the size of the dataset, with the fitted model overlaid. Below this plot, there is a plot of the pulls, which are given by:

$$\text{pull} = \frac{\hat{p}_i - f(N_i)}{\hat{\sigma}_i} \tag{10}$$

where f is the fitted polynomial. If the Gaussian assumption on $\hat{p}_i$ is correct and the fitted model prediction is the mean of its distribution, then the pulls are distributed by the standard normal distribution. On the right hand side of the plot is a density histogram of the pulls, with a standard normal distribution overlaid (the red line). The density histogram appears to agree fairly well with the standard normal distribution, indicating that the fitted model and the Gaussian assumption are both valid.

Part (g)

The methodology for this part was largely the same as that of part (f), except for the hypothesis testing, where the null hypothesis is that there is background and one signal, and the alternate hypothesis is that there are two distinct signals amongst the background. These are constructed by fitting 'signal plus background' and 'two signals plus background' models to the datasets using maximum likelihood estimation.

To find the desired range of dataset sizes to conduct the experiment, we followed the same trial and error procedure as in part (f), and settled on 50 equally-spaced sizes, $\{N_i\}$, in the range $[1500, 3000]$. For each of these sizes, we generated $n = 300$ example datasets. We performed a hypothesis test for the presence of two distinct signals on each dataset using the Neyman-Pearson test statistic (equation 1), and found k_i discoveries. We estimated the probability of discovery with equation 2 and its uncertainty with equation 3. We then fitted a third order polynomial, $f(N)$, to the dataset $\{(N_i, \hat{p}_i \pm \hat{\sigma}_i)|i = 1, ..., 50\}$ by minimising the sum of squared residuals (equation 4) (least squares fitting). We chose a third order polynomial as it is sufficient to approximate the relationship between p and N well within this range. It should be noted that on a wider range of dataset sizes, the relationship is S-shaped (moves from $p = 0$ at low N to $p = 1$ at high N), and certainly cannot be described using a third order polynomial.

If we assume that $\hat{p}_i$ is normally distributed with mean $f(N_i)$ and standard deviation $\hat{\sigma}_i$, then least squares fitting is equivalent to maximum likelihood estimation, and thus we can estimate the uncertainties of the fitted model as the minimum variance bound (equation **??**). Due to the central limit theorem, the assumption that $\hat{p}_i$ is normally distributed is valid in the limit of large n, but since $n = 300$, it is not certain that this assumption holds. The assumption that the mean of $\hat{p}_i$ is $f(N_i)$ holds only if the fitted model actually is the true relationship between N and p (ie. It is a good fit). To assess the validity of these assumptions, we calculate that coverage = 60% and perform a χ^2 test, in which we calculate $\chi^2/d.o.f = 1.197$ and p-value = 16.96%. If these assumptions are correct, then we should get coverage $\approx 68.3\%$, $\chi^2/d.o.f \approx 1$, and p-value $\geq 5\%$. These conditions are all satisfied to a reasonable degree, hence we can make this assumption safely. In Fig. 2 we plot the probability of discovery vs size of dataset with the fitted model overlaid, and the pulls (equation 6) visualised directly below. If the assumptions hold, then the pulls are distributed by the standard normal distribution. We can see that the density histogram of the pulls agrees well with the standard normal distribution, which further reinforces our assumptions.

To estimate the size of dataset needed to discover two distinct signals 90% of the time, N_{90}, we can use the Monte Carlo simulation method described in the methodology section. We obtained the result:

$$N_{90} = 2570 \pm 35$$

As discussed in part (f), the uncertainty on this result may not be trustworthy, as it was calculated assuming that the predictive model's parameter estimates were maximally efficient **??**, but this is only valid with asymptotic sample sizes. Since we had only 50 data-points, this assumption may not be correct.

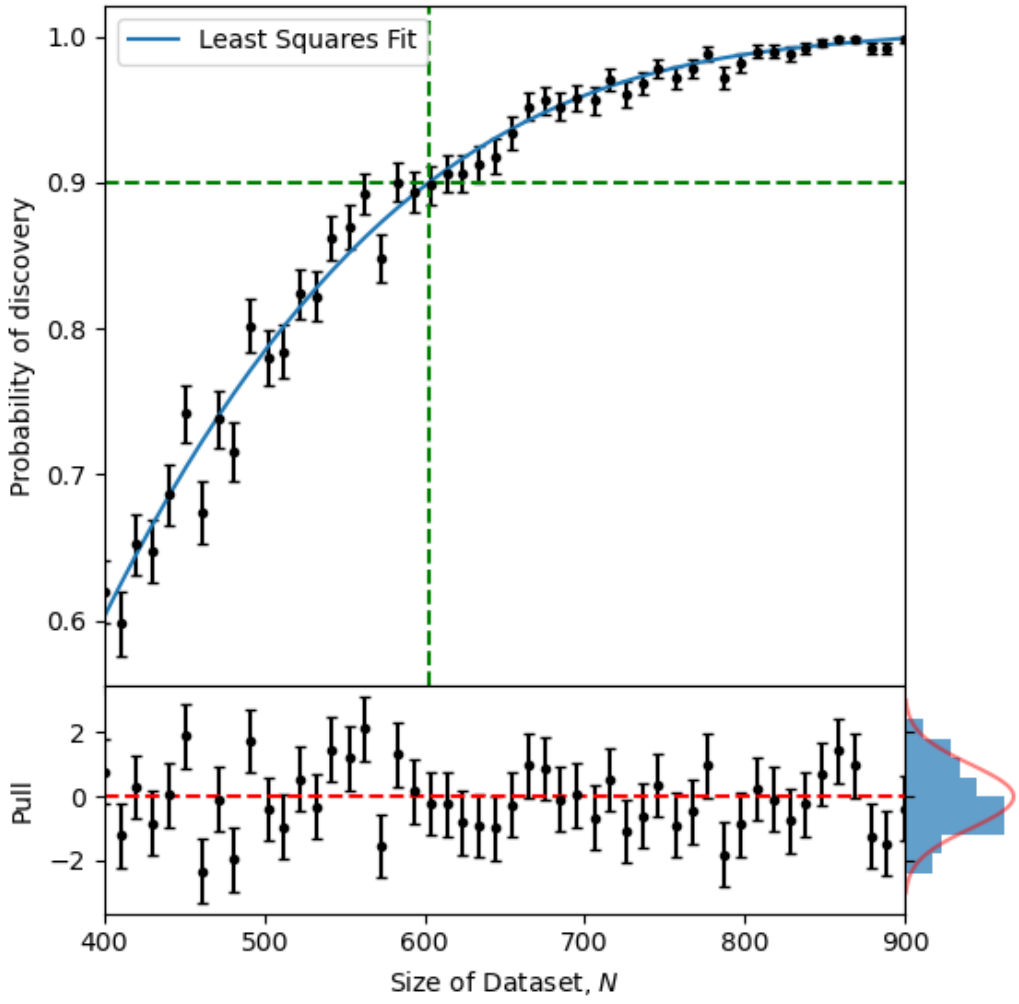

Figure 4: Plot of the estimated probability of discovering a signal vs the size of the dataset. The size of the dataset at which the probability is 0.9 is indicated by the green axis lines. The bottom plot shows the pulls to provide a visualisation of goodness-of-fit. A density histogram is plotted on the right hand edge of the pull plot, with the standard normal distribution overlaid.

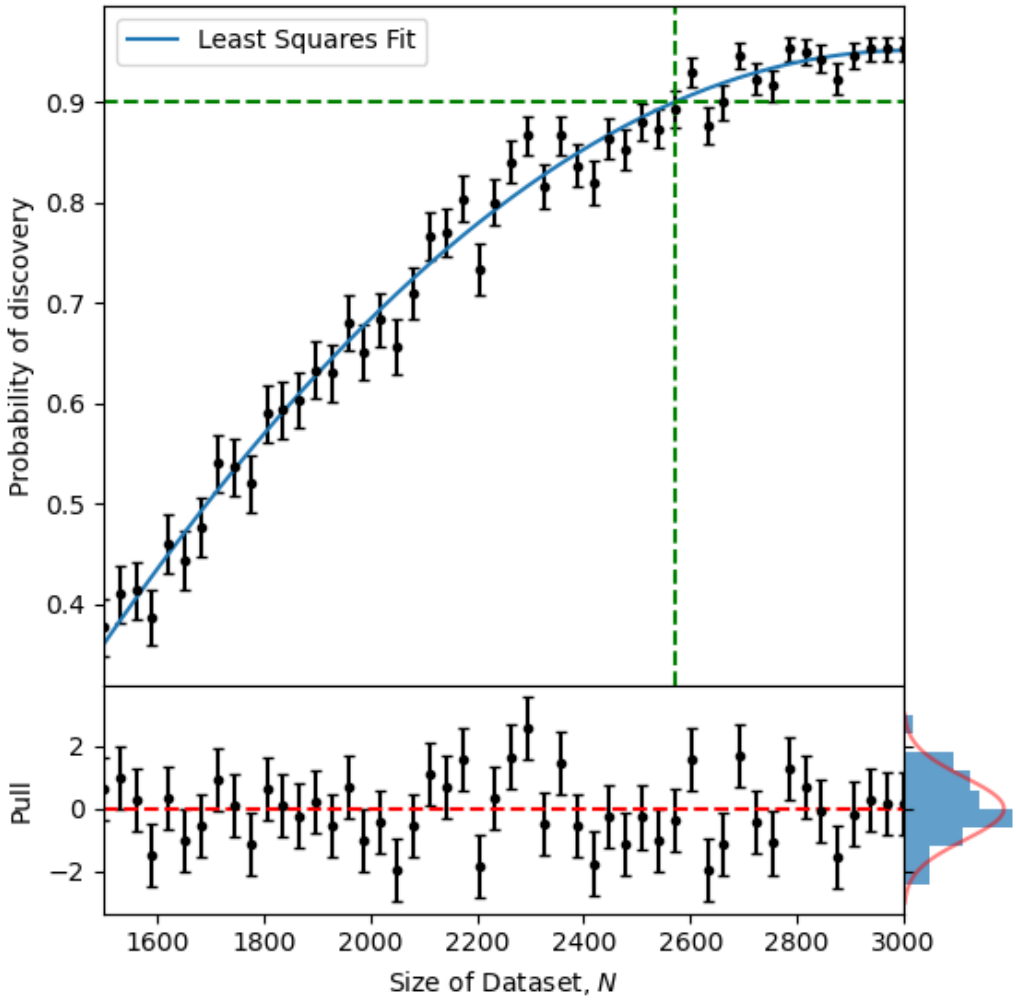

Figure 5: Plot of the estimated probability of discovering two distinct signals vs the size of the dataset. The size of the dataset at which the probability is 0.9 is indicated by the green axis lines. The bottom plot shows the pulls to provide a visualisation of goodness-of-fit. A density histogram is plotted on the right hand edge of the pull plot, with the standard normal distribution overlaid.