PH2221 Data Analysis

Solutions for homework problems

Magnus Lindström
Email: guslinmagc@student.gu.se

April 6, 2018

Contents

1 Chapter 1

1.1 Exercise 1

a) In the first variation of the problem I am shown a picture of one girl. The question is what gender is the other kid.

I will assume that the probability for having a girl is 50%. I have no reason to think otherwise. Since the two events of childbirth are independent (assuming that the children are not Siamese twins, in which case they have the same gender) the probability that the other child is a girl as well is 50%.

b) In the second variation of the problem, as far as I understand it, the woman and I have some sort of agreement that she will pick up two pictures of her two children and that she has to show me one picture of a girl if either of the two children are girls.

A priori, the four different sets of children that the woman can have are {bb}, {bg}, {gb} and {gg}. When the woman later shows me a picture of a girl I know that the scenario is one of the three last: {bg}, {gb} or {gg}. In only one of these three cases will the woman have another picture of a girl. Therefore, the chance of her showing me another girl is 1/3.

1.2 Exercise 2

We could ask many different questions which would produce different probabilities of the observed data. A few examples are

- What is the probability of getting the observed number of heads followed by one/two/three tails? That is, the number of HT/HTT/HTTT observed.
- What is the probability of getting the observed number of HTH combinations?
- What is the probability of the observed even/odd number of heads or tails?
- What is the probability that we the first and the last flips are both heads?

1.3 Exercise 3

Since I do not know what kind of detector is referred to, there is no answer to this question. I would need to know the distribution of the measured energies in order to make an estimate. For example, is it gaussian? When we know the distribution, then we could tell what the probability is for the true value to lie within some interval e.g. $\pm 1\sigma, \pm 2\sigma$. Right now, we don't know the answer.

1.4 Exercise 4

Discussion

In this exercise we need to make use of **Bayes' theorem**:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1}$$

It lends a way of calculating a conditional probability that might otherwise be hard to measure. $P(A|B)$, the **conditional probability** of A given B, is the probability of event A occurring after event B has occurred. The LHS of the equation is called the posterior, $P(B|A)$ is called the likelihood, $P(A)$ is called prior and $P(B)$ is called the evidence. Since $P(B)$ is often constant it is common to make use of $P(A|B) \propto P(B|A)P(A)$ and work with that, since it is often possible to normalise the distribution $P(A|B)$ by other means.

We will also make use of the following property of **independent events** (events that do not affect the probability of the other one happening):

$$P(A_1 \cap A_2) = P(A_1)P(A_2), \quad \text{if } P(A_{1,2}) \neq 0$$

and for two **mutually exclusive events** (two events that can not occur at the same time) A_1, A_2:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

The exercise

The different probabilities of this problem are listed in Table 1. Mongolian swamp fever is abbreviated as MSF. The complement of an event X is written as $\bar{X}$.

Notation	Probability	Description
$P(M)$	0.0001	The prob. that a patient has MSF.
$P(Sp\|M)$	1	The prob. of having spots given that the patient has MSF.
$P(L\|M)$	1	The prob. of having acute lethargy given that the patient has MSE.
$P(T\|M)$	0.6	The prob. of having raging thirst given MSE.
$P(Sn\|M)$	0.2	The prob. of having violent sneezes given MSE.
$P(Sp\|\bar{M})$	0.03	The prob. of a patient having spots given that he/she does not have MSE.
$P(L\|\bar{M})$	0.1	The prob. that a patient complains of lethargy given that he/she does not have MSF.
$P(T\|\bar{M})$	0.02	The prob. that a patient complains of thirst given that he/she does not have MSF.
$P(Sn\|\bar{M})$	0.05	The prob. that a patient complains of sneezing given that he/she does not have MSF.

Table 1: The different events, their respective probabilities of taking place and a description of the events.

To make the notation easier, I'll call the events $Sp, L, T, Sn = A_1, A_2, A_3, A_4$.

a) The first part of the question is: What is the probability that a patient presenting with all of the symptoms for MSF has MSF?

Using Bayes' theorem, we get

$$P(M|A_1,\cap,A_2,\cap,A_3,\cap,A_4) = \frac{P\big(\bigcap_{i=1}^4 A_i|M\big)P(M)}{P(\bigcap_{i=1}^4 A_i)} = \frac{\prod_{i=1}^4 \Big[P(A_i|M)\Big]P(M)}{P\big(\bigcap_{i=1}^4 A_i|M\big)P(M) + P\big(\bigcap_{i=1}^4 A_i|\bar{M}\big)P(\bar{M})}.$$

In the second sign of equality, in the nominator, I made use of the fact that the symptoms given the disease are, from what I understand in the text, independent from one another. Thus the break-up in the nominator is justified. The law of total probability was used for the rewrite of the denominator.

Now, since the probabilities for having the symptoms given MSF were assumed to be independent, and the probabilities for having them given not having MSF are independent according to the exercise, the expression takes the form

$$P(M|\bigcap_{i=1}^4 A_i) = \frac{\prod_{i=1}^4 \Big[P(A_i|M)\Big]P(M)}{\prod_{i=1}^4 \Big[P(A_i|M)\Big]P(M) + \prod_{i=1}^4 \Big[P(A_i|\bar{M})\Big]P(\bar{M})}. \tag{2}$$

Inserting the values from Table 1 gives the answer $\sim .80 = 80\%$

b) The second part of the question is: What is the probability that a patient presenting with three out of 4 symptoms for MSF has MSF? To solve this we use Equation 2 with a substitute $A_j \to \bar{A}_j$. All of the steps taken above are still valid.

The results are: the probabilities of having MSF given all symptoms but spots, lethargy, thirst or sneezing is 0%, 0%, 5.2% and 45.7%, respectively.

2 Chapter 2

2.1 Exercise 8

Discussion

The mean, or **expectation value**, of a discrete stochastic variable X that takes on the values $X = x \in \{x_1, x_2, \ldots, x_n\}$ according to the probability mass function (PMF) $p(x)$ is defined by

$$E[X] = \sum_{i=1}^{n} xp(x).$$

The continuous analog of the discrete expectation value, which is needed for the question at hand, is defined for a continuous stochastic variable X with probability density function (PDF) $f(x)$ as

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx.$$

The **variance** of a stochastic variable X is defined as:

$$Var[X] = \sigma_X^2 = E\big[(X - E[X])^2\big] = E\big[X^2 - 2XE[X] + E[X]^2\big] = \\ = E[X^2] - 2E[X]E[X] + E[X]^2 = E[X^2] - E[X]^2.$$

It describes how much much the PDF/PMF is "spread out" away from the mean. The standard deviation of X, σ_X, is defined as $\sigma_X = \sqrt{Var[X]}$.

To determine the probability for a continuous random variable X to take on different values, one makes use of the following relation:

$$P(a < X < b) = \int_a^b f(x)dx.$$

The probability for X to be a single value is always 0 for continuous variables.

The **mode** x^* of a random variable X is the value for which the PDF/PMF reaches maximum value.

The **median** of a continuous random variable X is the value x_{median} that satisfies:

$$\frac{1}{2} = P(-\infty < X < x_{median}) = \int_{-\infty}^{x_{median}} f(x)dx.$$

We use intervals to describe ranges of X that contain a certain probability, often denoted $1 - \alpha$. An interval containing a certain amount of probability can be chosen in an infinite amount of ways, but there are a couple of standard choices, the following two being among the most common ones.

The **central interval** $\mathcal{O}_{1-\alpha}^c = [x_1, x_2]$ (for a continuous variable X) containing probability $1 - \alpha$ is defined by

$$P(X < x_1) = P(X > x_2) = \alpha/2.$$

That is, the tails of the distribution contain an equal amount of probability.

When dealing with discrete stochastic variables or when, in practice, constructing a central interval from a continuous variable, it is necessary to work with discrete values.

The central interval $\mathcal{O}^c_{1-\alpha} = \{r_{inf}, r_{inf}+1, \ldots, r_{sup}\}$ of a discrete stochastic variable R with the possible values $\{r_1, r_2, ..., r_N\}$ is in this case defined by

$$\begin{aligned}
&\text{If } \; P(r_1) > \alpha/2 : \\
&\quad r_{inf} = r_1 \\
&\text{else:} \\
&\quad r_{inf} = \sup_{r \in r_1,\ldots,r_N} \Big[\sum_{i=0}^{r} P(i) \leq \alpha/2 \Big] + 1
\end{aligned}$$

and

$$\begin{aligned}
&\text{If } \; P(r_N) > \alpha/2 : \\
&\quad r_{sup} = r_N \\
&\text{else:} \\
&\quad r_{sup} = \inf_{r \in r_1,\ldots,r_N} \Big[\sum_{i=r}^{N} P(i) \leq \alpha/2 \Big] - 1.
\end{aligned}$$

In this way, the tails outside the central interval each contain more than $\alpha/2$ probability, but as close to $\alpha/2$ as possible.

The **smallest interval** $[x_1, x_2]$ (for a continuous variable X) containing probability $1-\alpha$ is defined by x_1, x_2 satisfying $f(x_1) = f(x_2)$, $x^* \in [x_1, x_2]$ and $P(x_1 < X < x_2) = 1-\alpha$.

When constructing a smallest interval from a discrete random variable, one uses the following recipe: The smallest interval $\mathcal{O}^s_{1-\alpha} = \{r_{inf}, r_{inf}+1, \ldots, r_{sup}\}$ of a discrete, stochastic variable r with the possible values $\{r_1, r_2, ..., r_N\}$ is defined by the following algorithm:

1. Start with the most probable value of r and set $\mathcal{O}^s_{1-\alpha} = \{r^*\}$. If $P(r^*) \geq 1-\alpha$, we are done. If not, go to point 2.

2. If $P(r \in \mathcal{O}^s_{1-\alpha}) < 1-\alpha$, add the next most probable value of $r = r_{next}$ to the set.

We then keep iterating over point two until $P(r \in \mathcal{O}^s_{1-\alpha}) \geq 1-\alpha$. With this construction algorithm, we ensure that the smallest interval will always contain at least $1-\alpha$ probability. Also, note that the interval does not for discrete stochastic variables have to consist of r-values all lying next to each other. If the PMF has several "peaks" the algorithm above will take pieces of the original set $\{r_1, ..., r_N\}$.

The exercise

a) The mean is given by

$$E[X] = \int_0^\infty x^2 e^{-x} dx = \cancel{\big[-x^2 e^{-x}\big]_0^\infty} + 2\int_0^\infty x e^{-x} dx = \cancel{2\big[-x e^{-x}\big]_0^\infty} + 2\int_0^\infty e^{-x} dx = \tag{3}$$

$$= -2\big[e^{-x}\big]_0^\infty = -2(0-1) = 2. \tag{4}$$

In order to get the standard deviation $\sigma = \sqrt{Var[X]}$, we need to calculate $E[X^2]$ ($E[X]$ is already given):

$$E[X^2] = \int_0^\infty x^3 e^{-x} dx = \cancel{\big[-x^3 e^{-x}\big]_0^\infty} + 3\int_0^\infty x^2 e^{-x} dx =$$
$$\cancel{3\big[-x^2 e^{-x}\big]_0^\infty} + 6\int_0^\infty x e^{-x} dx = \cancel{6\big[-x e^{-x}\big]_0^\infty} + 6\int_0^\infty e^{-x} dx =$$
$$-6\big[e^{-x}\big]_0^\infty = -6(0-1) = 6$$

Now, the standard deviation is

$$\sigma = \sqrt{Var[X]} = \sqrt{E[X^2] - E[X]^2} = \sqrt{6 - 2^2} = \sqrt{2}$$

The probability content within two standard deviations is given by

$$P(2-\sqrt{2} < X < 2+\sqrt{2}) = \int_{2-\sqrt{2}}^{2+\sqrt{2}} x e^{-x} dx = \big[\text{Integration by parts}\big] =$$
$$= [-x e^{-x}]_{2-\sqrt{2}}^{2+\sqrt{2}} - [e^{-x}]_{2-\sqrt{2}}^{2+\sqrt{2}} = \cdots \approx 0.74$$

Answer: $E[X] = 2$, $\sigma = \sqrt{2}$ and the probability contained within two standard deviations is ≈ 0.74.

b) To find the median value, we just have to integrate from 0 to x_{med} over the probability density function and equate that to 0.5:

$$P(X < x_{med}) = \int_0^{x_{med}} x e^{-x} dx = \cdots = -x_{med} e^{-x_{med}} - e^{-x_{med}} + 1 \overset{!}{=} 0.5 \Leftrightarrow$$
$$\Leftrightarrow (x_{med} + 1) e^{-x_{med}} = 0.5 \Rightarrow x_{med} \approx 1.68.$$

Now, in order to find the central interval we basically have to do the same thing again, but with the starting point $P(0 < X < x_1) \overset{!}{=} \frac{\alpha}{2} = 0.16$ and $P(x_2 < X) \overset{!}{=} \frac{\alpha}{2} = 0.16$. Carrying out the calculations in the same manner as with the median, the result is $\mathcal{O}^c_{0.68} = [0.71, 3.29]$.

c) To find the mode, take the derivative of the function and set it to zero:

$$\frac{d}{dx}(x e^{-x}) = e^{-x} - x e^{-x} = (1-x) e^{-x} \overset{!}{=} 0 \Rightarrow x = 1$$

To find the smallest interval, set up the system of equations

$$P(x_1 < X < x_2) = 1 - \alpha$$
$$P(X = x_1) = P(X = x_2).$$

These equations must hold since we want the probability $1 - \alpha$ to be contained in the interval and since $f(x)$ is a unimodal distribution the mode is always contained in the smallest interval. So, we get

$$\int_{x_1}^{x_2} x e^{-x} dx = \Big[\text{integration by parts}\Big] = x_1 e^{-x_1} - x_2 e^{-x_2} + e^{-x_1} - e^{-x_2} \overset{!}{=} 0.68 \Rightarrow$$
$$\Rightarrow x_1 \approx 0.27\,,\; x_2 \approx 2.49$$

2.2 Exercise 10

Discussion

When doing a Bayesian analysis of some data, the starting point is always Bayes' theorem, Eq. 1, with some well motivated priors. If there is no prior belief on the parameters, a constant (flat) prior is often used. That is the case for this assignment. The denominator is also constant so we're left with $P(\boldsymbol{\lambda}|D) \propto P(D|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ stands for the parameters and D stands for the observed data. The mode of the posterior is hence the mode of the likelihood.

It is implied in the problem formulation that the number of successes at each energy follows the Binomial distribution. The Binomial distribution has the PMF

$$p(x) = \text{Bin}(n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{5}$$

where n is the total number of "trials" that are either successes or failures, p is the probability of success and x is the number of successes.

The exercise

This exercise was solved using python (as well as every following exercise where programming is involved). First, for every energy level, a loop over 1000 different p-values spanning $[0, 1]$ was carried out and for every p-value the likelihood for observing the given data was calculated. The posterior for p was then taken to be the normalised likelihood. The smallest intervals for every energy level is given in Tab.2. The posteriors for p for every energy level are shown in Fig. 1 with the 68% credibility intervals shown in between the vertical red lines.

Energy level	r^D/N	Smallest interval for p
0.5	0.000	$[0.000, 0.011]$
1.0	0.040	$[0.024, 0.063]$
1.5	0.200	$[0.163, 0.241]$
2.0	0.580	$[0.531, 0.628]$
2.5	0.920	$[0.891, 0.944]$
3.0	0.987	$[0.983, 0.990]$
3.5	0.995	$[0.993, 0.997]$
4.0	0.998	$[0.997, 0.999]$

Table 2: Smallest intervals on p for every energy value.

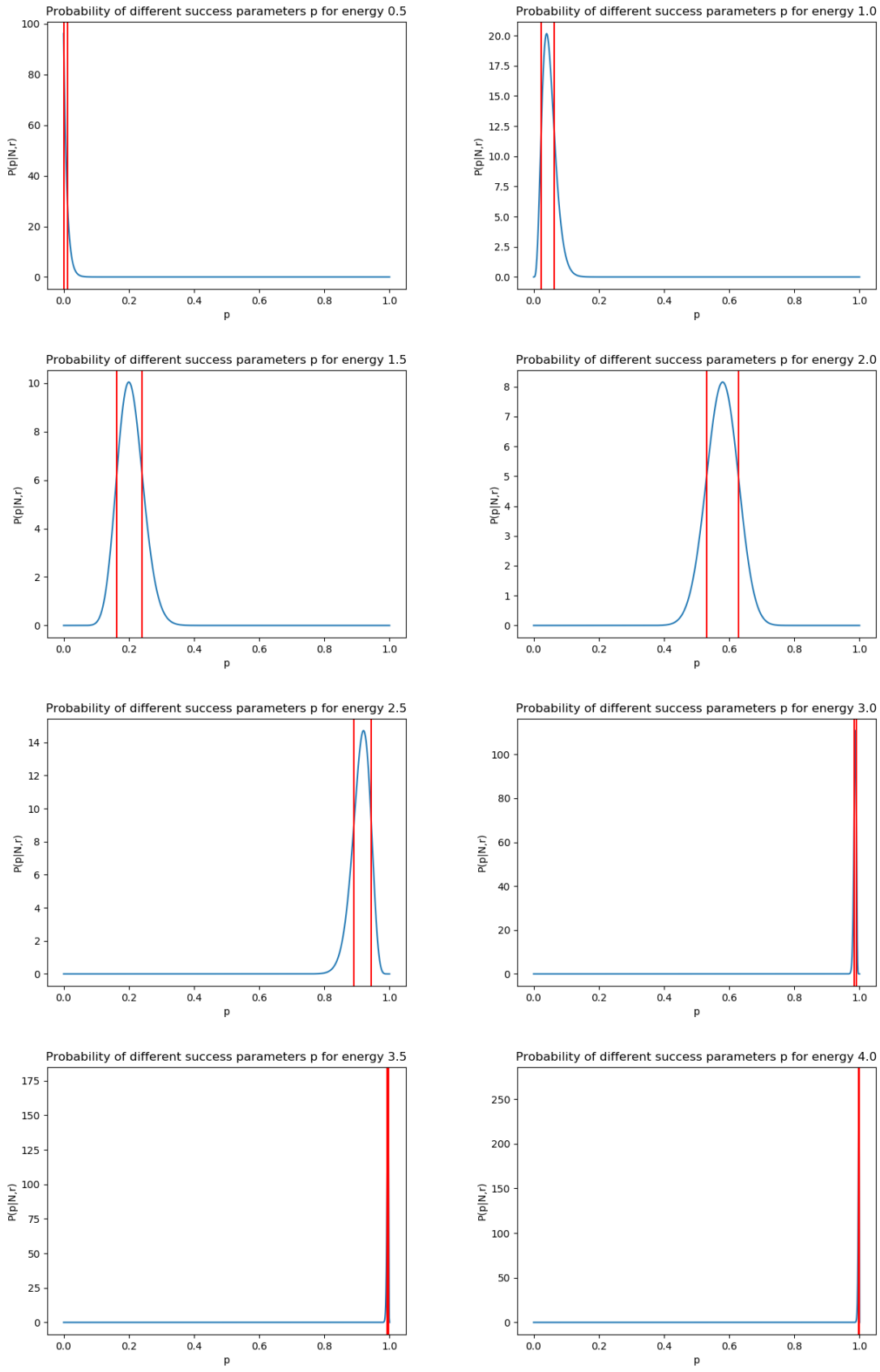

Figure 1: The 68% smallest credibility intervals for the different energy levels. The vertical red lines show the 68% smallest intervals.

2.3 Exercise 11

Discussion

In this exercise we are taking the so called "frequentist's view". We are to analyse the same data as in exercise 10 but in a different way. Instead of constructing credibility intervals we want **confidence level intervals**. In order to make a $1-\alpha$ CL interval, you follow the Neyman construction:

1. Define which interval you want to use. In this case, we're using $\mathcal{O}^c_{0.90}$.
2. Do your experiment and, in this case, count the number of successes r_D in the experiment.
3. For the measured r_D, find out which values of the parameters of interest, here p, that would make $r_D \in \mathcal{O}^c_{0.90}$. One uses the likelihood to determine if the measured number of successes is within the interval of choice.
4. The resulting range of p is our (Neyman) confidence level interval for p.

So what does this CL interval tell us? It tells us that, in this case, we are "90% sure" that the true value of our parameter, p_0, lies within the CL interval that r_D gives rise to. More specifically, if we were to repeat the experiment a large number of times, the true value p_0 would be contained within the CL intervals created by our experiments 90% of the time.

The exercise

For every energy level, $1,000$ values of p between 0 and 1 were tested. For every value of p, the probability to observe $r = 0, 1, \ldots, N$ successes were calculated. For every value of r, a 90% CL interval was created. The results of this are seen in Fig 2. For our measured values $\{r_i^D\}$, the CL intervals are shown in Tab. 3

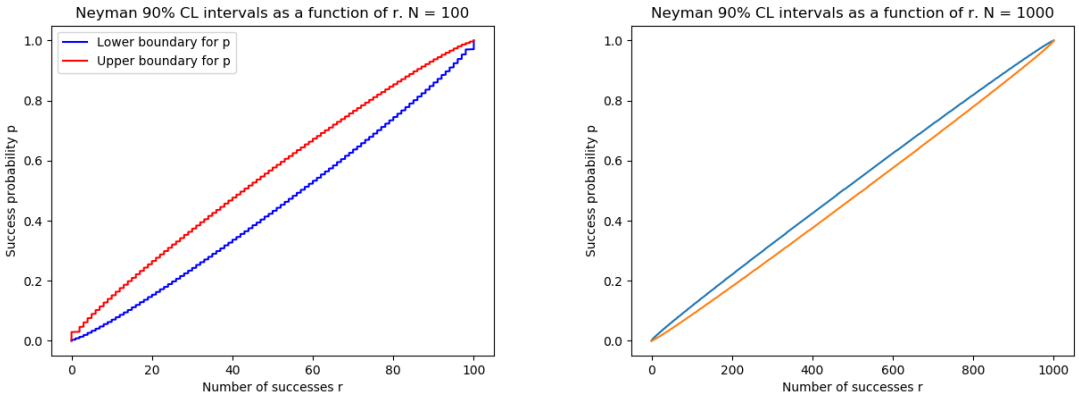

Figure 2: Neyman CL intervals for p as a function of r. Left: $N = 100$. Right: $N = 1000$.

Energy level	r^D/N	$\mathcal{O}^c_{0.90}$ CL interval for p
0.5	0.000	[0.000, 0.029]
1.0	0.040	[0.020, 0.075]
1.5	0.200	[0.146, 0.265]
2.0	0.580	[0.504, 0.654]
2.5	0.920	[0.873, 0.952]
3.0	0.987	[0.981, 0.991]
3.5	0.995	[0.991, 0.997]
4.0	0.998	[0.996, 0.999]

Table 3: The Neyman 90% CL intervals for p given measured data at the different energy levels.

2.4 Exercise 13

Discussion

The exercise is about exploring what happens if we reuse the same data multiple times and take the previous posterior to be our new prior. This is the exact same thing as redoing the same experiment over and over again with the exact same result every time. We would thus expect the posterior to become more and more centered around a specific value and the uncertainty, or variance, to become smaller and smaller since we keep observing the same thing again and again.

The exercise

We perform an experiment where we measure the number of successes from what we understand is a Binomially distributed random variable $\widetilde{R}$. Out of N trials, we measure r successes. Let's now reuse this data n times and see what happens to the posterior $P_n(p|r,N)$.

The first time we create the posterior, we use a flat prior $P_0(p) = 1, p \in [0,1]$:

$$P_1(p|r,N) = \frac{P(r|p,N)\cancel{P_0(p)}}{\int_0^1 P(r|p',N)\cancel{P_0(p')}dp'} = \frac{\cancel{\binom{N}{r}}p^r(1-p)^{N-r}}{\int_0^1 \cancel{\binom{N}{r}}p'^r(1-p')^{N-r}dp'} =$$
$$= \frac{p^r(1-p)^{N-r}}{\beta(r+1,N-r+1)} = \frac{p^r(1-p)^{N-r}}{\frac{r!(N-r)!}{(N+1)!}} = \frac{(N+1)!}{r!(N-r)!}p^r(1-p)^{N-r}$$

Let's now use this distribution over p as our prior for the next experiment:

$$P_2(p|r,N) = \frac{P(r|p,N)P_0(p')}{\int_0^1 P(r|p',N)P_0(p')dp'} = \frac{\cancel{\binom{N}{r}}p^r(1-p)^{N-r}\cancel{\frac{(N+1)!}{r!(N-r)!}}p^r(1-p)^{N-r}}{\int_0^1 \cancel{\binom{N}{r}}p'^r(1-p')^{N-r}\cancel{\frac{(N+1)!}{r!(N-r)!}}p'^r(1-p')^{N-r}dp'} =$$
$$= \frac{p^{2r}(1-p)^{2(N-r)}}{\beta(2r+1,2N-2r+1)} = \frac{(2N+1)!)}{(2r)!(2N-2r)!}p^{2r}(1-p)^{2(N-r)}$$

Repeating this step one more time, doing the same type of calculations, gives us

$$P_3(p|r,N) = ... = \frac{(3N+1)!)}{(3r)!(3N-3r)!}p^{3r}(1-p)^{3(N-r)}.$$

So we see that repeating this n times does indeed give us the expected posterior

$$P_n(p|r,N) = \frac{(nN+1)!)}{(nr)!(nN-nr)!}p^{nr}(1-p)^{n(N-r)}.$$

Now, taking the expectation value of p with respect to $P_n(p|r,N)$ (which is a normalised distribution over $[0,1]$ (it integrates to 1) gives us:

$$E[p] = \int_0^1 pP_n(p|r,N)dp = \frac{(nN+1)!)}{(nr)!(nN-nr)!}\int_0^1 p^{nr+1}(1-p)^{n(N-r)}dp =$$
$$= \frac{(nN+1)!)}{(nr)!(nN-nr)!}\beta(nr+2,nN-nr+1) = \frac{(nN+1)!)}{(nr)!\cancel{(nN-nr)!}}\frac{(nr+1)!\cancel{(nN-nr)!}}{(nN+2)!} =$$
$$= \frac{\cancel{(nN+1)!}}{\cancel{(nr)!}}\frac{(nr+1)\cancel{(nr)!}}{(nN+2)\cancel{(nN+1)!}} = \frac{nr+1}{nN+2} \Rightarrow \frac{r}{N}, \quad \text{as} \quad n \Rightarrow \infty.$$

This is what one intuitively expects. If you do an infinite amount of 10 repetitions of coin flips in which you keep observing 5 heads and 5 tails, you would conclude that the probability p of observing a head should be $5/10 = 0.5$.

Now, for the variance we need $E[p^2]$. Using much the same methods as for the mean, we get

$$E[p^2] = \int_0^1 p^2P_n(p|r,N)dp = \frac{(nN+1)!)}{(nr)!(nN-nr)!}\int_0^1 p^{nr+2}(1-p)^{n(N-r)}dp =$$
$$= \frac{(nN+1)!)}{(nr)!(nN-nr)!}\beta(nr+3,nN-nr+1) = ... = \frac{(nr+2)(nr+1)}{(nN+3)(nN+2)} \Rightarrow \frac{r^2}{N^2}, \quad \text{as} \quad n \Rightarrow \infty.$$

And the variance is $Var[p] = E[p^2] - E[p]^2 = 0$, which is also expected.

3 Chapter 3

3.1 Exercise 4

Discussion

A new concept for this question is the **full width at half maximum** (FWHM) of a function, which is the range of an independent variable between the two points where the function reaches half of its maximum value. So it's the points x_1, x_2 where $f(x_{1,2}) = f(x_{max})/2$, with x_{max} being the value at which f has it's maximum.

The exercise

a) The mean of x under the PDF $f(x) = \frac{1}{2}e^{-|x|}, x \in [-\infty, \infty]$ is 0, since f is a symmetric function around 0 with the mode of x being $x^* = 0$. Therefore $E[x] = 0$.

For the standard deviation, we need to know $E[x^2]$:

$$E[x^2] = \frac{1}{2}\int_{-\infty}^{\infty} x^2 e^{-|x|}dx = \frac{1}{2}\int_{-\infty}^{0} x^2 e^{x} + \frac{1}{2}\int_{0}^{\infty} x^2 e^{-x} = \int_{0}^{\infty} x^2 e^{-x}.$$

But we've seen this integral before in Eq. 3, and it evaluates to 2.

So we get $\sigma = \sqrt{Var[x]} = \sqrt{2 - 0^2} = \sqrt{2}$.

b) We're looking for a x_{HM} that satisfies $f(x_{HM}) = \frac{1}{2}f(0)$. We will just look at positive values of x since the function is symmetric.

$$\frac{1}{2}e^{-x} = \frac{1}{4} \Leftrightarrow e^{-x} = \frac{1}{2} \Leftrightarrow x = ln(2) \Longrightarrow x_{HM} = \pm ln(2)$$

So, the FWHM is 2ln(2). The standard deviation is roughly 1.41 and the FWHM is roughly 1.39, so in this case they are about equal. But that does not have to be the case for every function.

c) The probability contained within $\pm\sigma$ around the peak is

$$\int_{-\sqrt{2}}^{\sqrt{2}} \frac{1}{2}e^{-|x|}dx = \int_{0}^{\sqrt{2}} e^{-x}dx = \big[-e^{-x}\big]_0^{\sqrt{2}} = -e^{-\sqrt{2}} + e^0 \approx 0.76$$

3.2 Exercise 7

Discussion

The **Poisson distribution** is used to model events where there is a large, unknown, number of independent "trials" with a small chance for "success" for every trial. This can for example be the number of atoms in some sample that undergo radioactive decay during a time period, where it can be assumed that the total number of atoms stay the same during the process. A stochastic variable n that counts the number of successes in this scenario will be distributed according to a Poisson distribution with the PDF

$$f(n) = \text{Poi}(n|\nu) = \frac{e^{-\nu}\nu^n}{n!}, \tag{6}$$

where ν is the expected number of successes.

When conducting an experiment, counting successes from something modeled by a Poisson process, and constructing the posterior with the use of a flat prior, it is interesting to note that the posterior takes the same form as the PDF itself. Say we count n_D successes, what is the posterior for ν? Let's examine:

$$P(\nu|n_D) = \frac{P(n_D|\nu)\cancel{P_0(\nu)}}{\int_0^{\nu_{max}} P(n_D|\nu')\cancel{P_0(\nu')}d\nu'} = \frac{\frac{\nu^{n_D}e^{-\nu}}{\cancel{n_D!}}}{\int_0^{\nu_{max}} \frac{\nu'^{n_D}e^{-\nu'}}{\cancel{n_D!}}d\nu'} = \frac{\nu^{n_D}e^{-\nu}}{\int_0^{\nu_{max}} \nu'^{n_D}e^{-\nu'}d\nu'} \tag{7}$$

ν_{max} is an arbitrary large value of ν that is required to make the prior $P_0(\nu) = \frac{1}{\nu_{max}}$ non-zero so that they can cancel each other. When taking $\nu_{max} \Rightarrow \infty$ the numerator of Eq. 7 goes to the familiar value $n_D!$, which gives us the posterior

$$P(\nu|n_D) = \frac{e^{-\nu}\nu^{n_D}}{n_D!}. \tag{8}$$

This is the same form as the original PDF in Eq. 6, but with ν being the variable instead of n.

The exercise

a) To calculate the 95% probability lower limit on ν after an experiment with $n_D = 9$ and a flat prior, we simply have to make use of Eq. 8 and do the usual integral:

$$P(\nu > \nu_{95}|n_D = 9) = \int_0^{\nu_{95}} \frac{e^{-\nu}\nu^9}{9!}d\nu \stackrel{!}{=} 0.05 \Rightarrow \nu_{95} \approx 5.43.$$

So, with 95% certainty we can conclude that ν is larger than 5.43.

b) To find the 68% CL interval using the Neyman construction and the smallest interval definition, we follow the recipe in Sec. 2.3 with 1,000 values for ν in the interval $[0, 50]$. The resulting CL interval is $\mathcal{O}^s_{0.68} = [6.51, 13.26]$.

3.3 Exercise 8

Discussion

One issue, or rather an annoyance, with the smallest and central interval definitions when constructing CL intervals is that the intervals can in some cases be empty, that is $\mathcal{O}^{c/s}_{1-\alpha} = \{\}$. This problem arises when, for example, one wants to model a Poisson process with parameter λ when another noise source following Poi(ν), where ν is known, is also present. The way to go about this is to create a CL interval for $\mu = \lambda + \nu$ with μ being the combined expectation value of the two processes and then simply subtract the value of ν to get the interval for λ. When constructing the smallest/central intervals of different values for n given a ν, it sometimes happens that when subtracting the background, the resulting CL interval for the signal rate can be $\lambda \in [0, 0]$.

There is no problem with this from a mathematical viewpoint, since the interval is designed to contain the true value λ_0 only in $100(1-\alpha)\%$ of the time. But the somewhat poor interpretation of CL intervals being ranges where a variable is likely to be has led

to the appearance of other intervals. The **Feldman-Cousins interval** (FC interval) is designed to take care of this problem.

The FC interval is constructed with a different ranking of data, not strictly on probability of observing the results as was the case in central and smallest intervals. The ranking r of an observed data n for the case of multiple Poisson processes is

$$r = \frac{P(n|\mu = \lambda + \nu)}{P(n|\hat{\mu})},$$

where $\hat{\mu}$ is the value of μ that maximises the probability of observing n subject to any restriction that might apply. For double poisson processes, we want/expect our source rate to be non-negative $\lambda > 0$, so we do not allow the combined rate $\hat{\mu}$ to fall below ν:

$$\hat{\mu} = \begin{cases} n & \text{if } n > \nu \\ \nu & \text{if } n \leq \nu. \end{cases}$$

This ensures that we do not get empty intervals.

räcker den förklaringen?

The exercise

a, b, c) As mentioned above, the way to construct a CL interval for the value of the signal λ is to subtract the known background of $\nu = 3.2$. The resulting 68% CL intervals and credible interval from doing just that are found in Tab. 4.

Type of interval used	Interval
FC interval	[3.156, 9.563]
Smallest interval	[3.307, 10.063]
Credible interval	[3.153, 9.159]

Table 4: The 68% CL intervals and credible intervals for λ given $n_D = 9$.

3.4 Exercise 13

We have for the unbinned likelihood

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda)$$

and for the expectation in bin j

$$\nu_j = \int_{\Delta_j} f(x|\lambda)dx. \tag{9}$$

We are to show that

$$\lim_{K\to\infty} \prod_{j=1}^{K} \frac{e^{-\nu_j}\nu_j^{n_j}}{n_j!} = \frac{1}{e}\prod_{i=1}^{n} f(x_i|\lambda)\Delta, \tag{10}$$

where Δ is the size of the bin interval in x.

Start by looking at the LHS of Eq. 10. The product of all of the exponentials as $K \to \infty$ will lead to a factor

$$\exp\Big(- \sum \nu_j \Big) = \exp\Big(- \int_{-\infty}^{\infty} f(x|\lambda) dx \Big) = \frac{1}{e},$$

since $f(x|\lambda)$ is a proper distribution. Now take the rest of the LHS, $\nu_j^{n_j}/n_j!$. As the number of bins becomes very large, at some point every bin will contain either 1 or 0 events. The bins with 1 event will contribute with a factor $\nu_j^1/1! = \nu_i$ (i because the iterating index now coincides with the registered events that we have) and the ones with 0 events $\nu_j^0/0! = 1$. But, looking at Eq. 9, the value of ν_i approaches

$$\nu_i \to f(x_i|\lambda)\Delta,$$

as $K \to \infty$.

Taken together, this means that the LHS of Eq. 10 becomes

$$\lim_{K\to\infty} \prod_{j=1}^{K} \frac{e^{-\nu_j} \nu_j^{n_j}}{n_j!} = \frac{1}{e} \prod_{i=1}^{n} \nu_i = \frac{1}{e} \prod_{i=1}^{n} f(x_i|\lambda)\Delta,$$

Q.E.D.

3.5 Exercise 16

In this exercise, we're dealing a variable $X \sim \text{Bin}(N, p)$, where $N \sim \text{Poi}(\lambda)$. The goal is to show that

$$P(X) = \frac{e^{\nu p}(\nu p)^X}{X!}.$$

To achieve this, we sum over all possible values of N to get rid of the N dependency:

$$P(X) = \sum_{N=0}^{\infty} P(X|N,p)P(N|\nu) = \big[N < X \text{ is impossible}\big] =$$

$$= \sum_{N=X}^{\infty} P(X|N,p)P(N|\nu) = \sum_{N=X}^{\infty} \binom{N}{X} p^X (1-p)^{N-X} \frac{e^{-\nu}\nu^N}{N!} = \big[\text{set } N = X + l\big] =$$

$$= \sum_{l=0}^{\infty} \binom{X+l}{X} p^X (1-p)^l \frac{e^{-\nu}\nu^{X+l}}{(X+l)!} = \sum_{l=0}^{\infty} \frac{\cancel{(X+l)!}}{X!l!} p^X (1-p)^l \frac{e^{-\nu}\nu^{X+l}}{\cancel{(X+l)!}} =$$

$$= \frac{(\nu p)^X e^{-\nu}}{X!} \underbrace{\sum_{l=0}^{\infty} \frac{(v(1-p))^l}{l!}}_{\text{definition of e}} = \frac{(\nu p)^X \cancel{e^{\nu}}}{X!} \cancel{e^{-\nu p}} = \frac{e^{-\nu p}(\nu p)^X}{X!},$$

which is the sought after result.

4 Chapter 4

4.1 Exercise 8

Discussion

In this exercise we numerically test the **Central Limit Theorem** (CLT) which states that the average of a series of measurements $\bar{x} = \sum_{i=1}^{n} x_i/n$ from any distribution $P(X)$ with mean μ and finite moments will in the limit of $n \to \infty$ approach $N(\bar{x}|\mu, \frac{\sigma_X^2}{n})$. That is, a normal distribution with mean μ and the same variance as the original distribution divided by n. This is a very important result as is tells us that $\bar{x}$ is an **unbiased estimator** of μ_X, which means that the expectation value of $\bar{x}$ is μ_X. On top of that, it follows a well known distribution that we know how to work with, the normal distribution.

There's an important caveat that concerns the applicability of the CLT, which is that the underlying distribution must have finite **moments**. The n-th moment of a random variable is defined as $E[x^n]$.

The Cauchy distribution has the following PDF:

$$f(x) = \frac{1}{\pi\gamma} \frac{\gamma^2}{(x - x_0)^2 + \gamma^2},$$

with x_0 being the mode and median (but not mean) and $\gamma > 0$ being a parameter of the distribution. When trying to calculate any moment, one does not arrive at any defined values. This means that the distribution lacks both a mean and a variance. Let's see what happens to the distribution of $\bar{x}$ under the normal circumstances (finite moments) and under the circumstance of undefined moments.

The exercise

a) The values of n and γ were chosen to be $n = 5, 20, 100$ and $\lambda = 0.5, 5, 50$. For both assignment a and b, $N = 10,000$ means were calculated. The distribution of the means drawn from the exponential distribution can be seen in the top half of Fig. 3. The blue line is the gaussian curve that the CLT tells us that the mean will approach as $n \to \infty$. It is clear that that is indeed the case.

b) The parameters here were $n = 5, 20, 100$, $x_0 = 25$ and $\gamma = 3$. The distribution of $N = 10,000$ means is shown in the lower half of Fig. 3. The range on the x-axis is set to be $\pm 3\sigma$ around the mean of the distribution of $\bar{x}$, the same as for assignment a. It can clearly be seen that the distribution of $\bar{x}$ drawn from the Cauchy distribution does not look the same as when drawn from the exponential distribution. The tails of the Cauchy distribution are large enough to cause the integral over its PDF to diverge. This also causes the variance of the drawn numbers to often lie far away from the mode x_0.

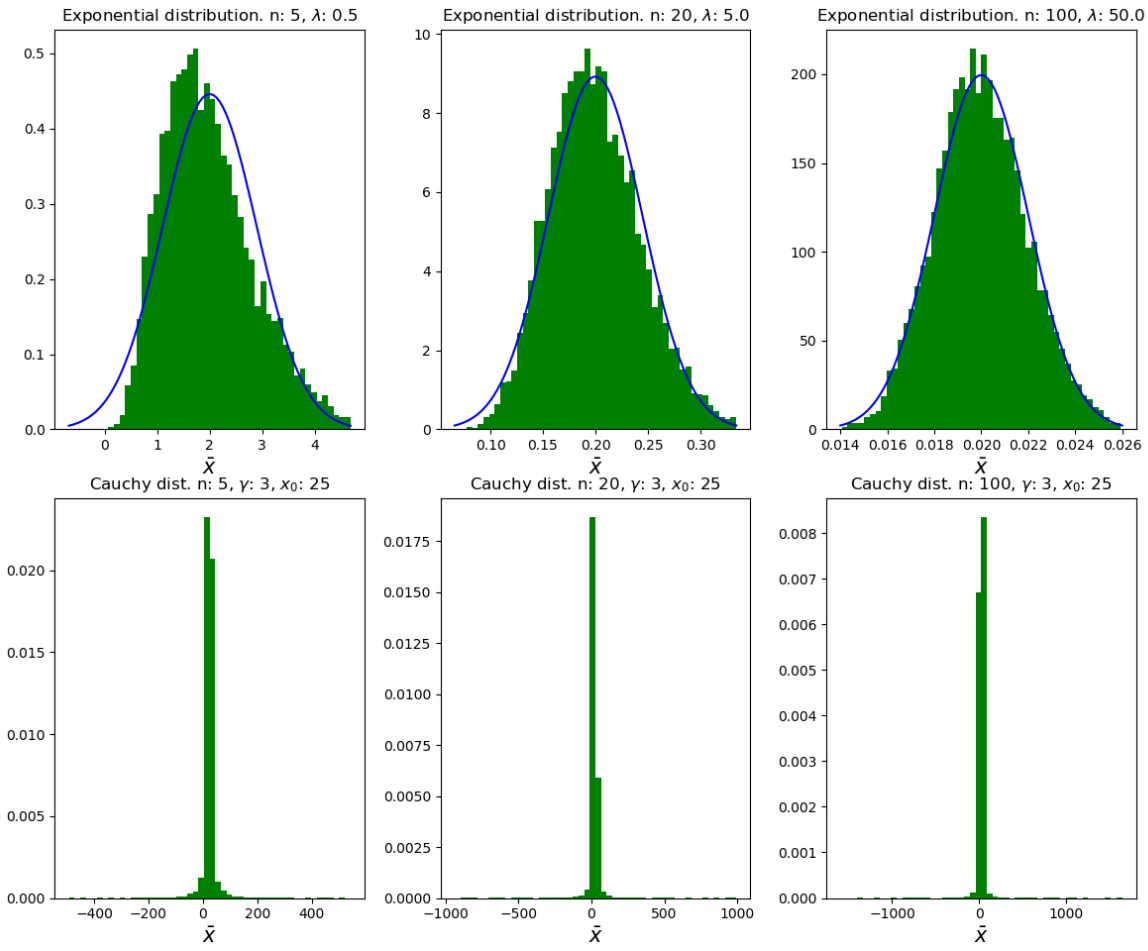

Figure 3: Numerical test of the CLT. $N = 10,000$ means $\bar{x}$ calculated from $n = 5, 20, 100$ points. The green histogram is the distribution of $\bar{x}$ and the blue line is what the CLT predicts when the distribution satisfies its conditions. The range on the x-axis is in all cases chosen to $\pm 3\sigma$ from the mean of the drawn $\{\bar{x}_i\}$. **Top**: Exponential distribution with $\lambda = 0.5, 5, 50$. **Bottom**: Cauchy distribution with $x_0 = 25$ and $\gamma = 3$.

4.2 Exercise 11

Discussion

So far we have only looked at univariate distributions of stochastic variables $f(x)$, with one random variable. But this can be extended to **multivariate distributions** where two or more variables are random: $f(x_1, ..., x_k)$. When having more than one variable, the variance of a single variable no longer explains the shape of the distribution, instead the **covariance matrix $\mathbf{\Sigma}$** is introduced

$$\mathbf{\Sigma} = \begin{pmatrix} E[(x_1 - \mu_1)(x_1 - \mu_1)] & \cdots & E[(x_1 - \mu_1)(x_k - \mu_k)] \\ \vdots & \ddots & \vdots \\ E[(x_k - \mu_k)(x_1 - \mu_1)] & \cdots & E[(x_k - \mu_k)(x_k - \mu_k)] \end{pmatrix},$$

where $\mu_i = E[x_i]_{f(\boldsymbol{x})}, i = 1, ..., k$ are the expectation values of the individual variables taken with respect to the joint PDF $f(\boldsymbol{x})$. An entry $\Sigma_{i,j}$ describes how variables i and j vary together, a large positive value means that if one is big, then the other is as well. A negative value means that one variable takes on low values when the other one takes on high. $\Sigma_{i,j}$ can also be written as $\Sigma_{i,j} = \rho_{i,j}\sigma_x\sigma_y$, where $\rho_{i,j}$ is the **Pearson correlation coefficient** between x_i and x_j. The Pearson correlation coefficient is defined from the same relation. $\rho_{i,j}$ takes on values $\rho_{i,j} \in [-1, 1]$, where 1 indicates perfect linear correlation and -1 negative linear correlation. $\rho_{i,j} = 0$ means no correlation.

Thus, with the Pearson correlation coefficient, the correlation matrix $\mathbf{\Sigma}$ can be written as

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \cdots & \rho_{1,k}\sigma_1\sigma_k \\ \vdots & \ddots & \vdots \\ \rho_{k,1}\sigma_k\sigma_1 & \cdots & \sigma_k^2 \end{pmatrix}, \tag{11}$$

where σ_i^2 being the variance of x_i. In this exercise, there are only two variables so $\rho_{1,2} = \rho_{2,1} = \rho$.

The exercise

The plots are shown in Fig. 4.

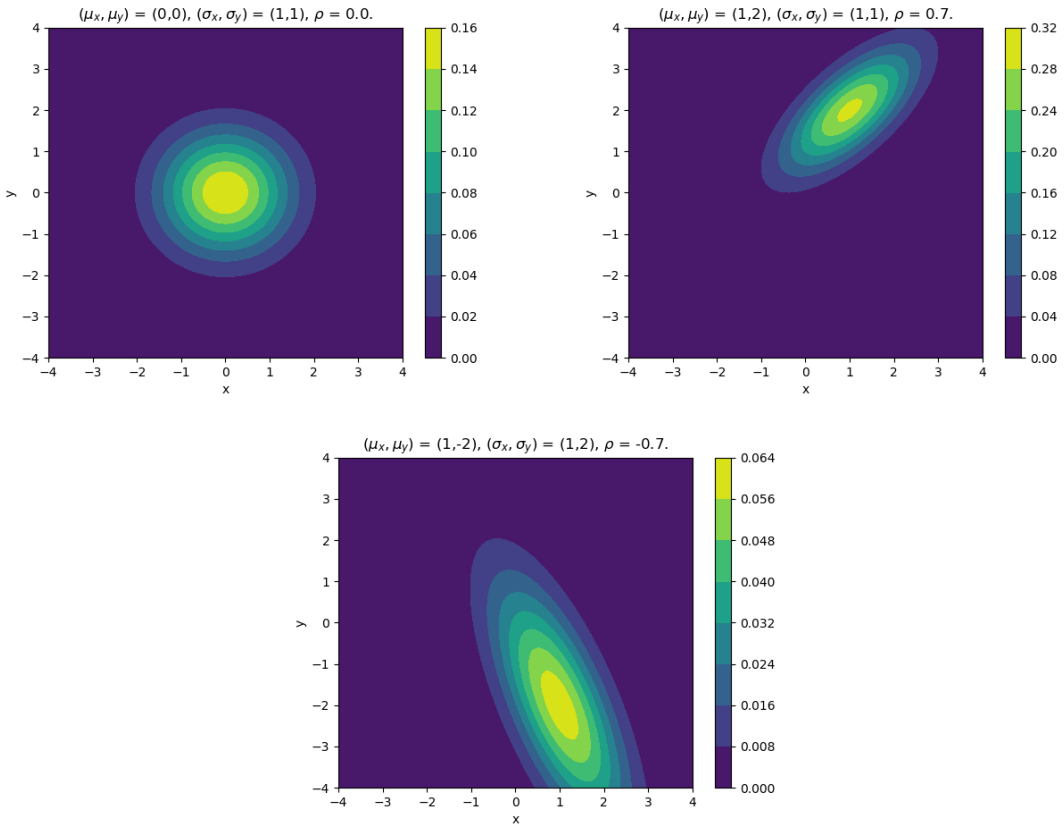

Figure 4: The PDF of a bivariate Gaussian distribution with different sets of parameters.

4.3 Exercise 12

a) We are to show that, for the special case of $k = 2$ dimensions,

$$N\Big(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\Big) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}\exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big) = \\ = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\Big(-\frac{1}{2(1-\rho^2)}\big(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-\frac{2\rho xy}{\sigma_x\sigma_y}\big)\Big), \tag{12}$$

where the first expression is the general multivariate Normal distribution with k dimensions and $|\boldsymbol{\Sigma}|$ being the determinant of the covariance matrix.

Start from the first expression in Eq. 12. $|\boldsymbol{\Sigma}|$ can be expanded to $|\boldsymbol{\Sigma}| = \sigma_x^2\sigma_y^2 - \rho^2\sigma_x^2\sigma_y^2$. With $k = 2$ we have the prefactor

$$\frac{1}{\sqrt{(2\pi)^2|\boldsymbol{\Sigma}|}} = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}. \tag{13}$$

What's left to show is now only the exponent. It contains the inverse of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{|\boldsymbol{\Sigma}|}\begin{pmatrix}\sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2\end{pmatrix}.$$

So, the whole expression inside the exponent evaluates to

$$-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \frac{-1}{2\sigma_x^2\sigma_y^2(1-\rho^2)}\begin{pmatrix}x-\mu_x\\ y-\mu_y\end{pmatrix}^T\begin{pmatrix}\sigma_y^2 & -\rho\sigma_x\sigma_y\\ -\rho\sigma_x\sigma_y & \sigma_x^2\end{pmatrix}\begin{pmatrix}x-\mu_x\\ y-\mu_y\end{pmatrix} =$$
$$= \cdots = -\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right).$$

Setting $\mu_x = \mu_y = 0$ and combining this result with that of Eq. 13, we arrive at the RHS of Eq. 12.

b) We have a bivariate gaussian distribution $P(x,y) = N\big((\mu_x,\mu_y),(\sigma_x^2,\sigma_y^2)\big)$, as described above. We first set $\mu_x = \mu_y = 0$ via a coordinate shift. Now, in order to get a probability distribution for $z = x - y$, we need to make a substitution in $P(x,y)$ and then integrate over the other variable according to

$$y = x - z \quad \text{and} \quad P(z) = \int_{-\infty}^{\infty} P(x, x-z)dx.$$

Let's make $P(z)$ more manageable:

$$P(z) = \int_{-\infty}^{\infty} \underbrace{\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}}_{=A} \exp\left\{ \underbrace{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2}+\frac{(x-z)^2}{\sigma_y^2}-\frac{2\rho x(x-z)}{\sigma_x\sigma_y}\right)}_{=E} \right\} dx.$$

If we collect the terms in E containing different powers of x we get the expression

$$E = -\beta x^2 + \gamma x - \delta,$$

with

$$\beta = \frac{1}{2(1-\rho^2)}\Big(\frac{1}{\sigma_x^2}+\frac{1}{\sigma_y^2}-\frac{2\rho}{\sigma_x\sigma_y}\Big), \quad \gamma = \frac{2z}{2(1-\rho^2)}\Big(\frac{1}{\sigma_y^2}-\frac{\rho}{\sigma_x\sigma_y}\Big)$$
$$\text{and} \quad \delta = \frac{z^2}{2\sigma_y^2(1-\rho^2)}.$$

So, we can write

$$P(z) = A\int_{-\infty}^{\infty}\exp\Big(-\beta x^2+\gamma x-\delta\Big)dx = A\exp\big(-\delta\big)\int_{-\infty}^{\infty}\exp\Big(-\beta x(x-\frac{\gamma}{\beta})\Big)dx =$$
$$= \Big[x\to x+\frac{\gamma}{2\beta}\Big] = A\exp\big(-\delta\big)\int_{-\infty}^{\infty}\exp\Big(-\beta\big(x+\frac{\gamma}{2\beta}\big)\big(x-\frac{\gamma}{2\beta}\big)\Big)dx =$$
$$= A\exp\big(-\delta\big)\int_{-\infty}^{\infty}\exp\Big(-\beta\big(x^2+\frac{\gamma^2}{4\beta^2}\big)\Big)dx = A\exp\big(\frac{\gamma^2}{4\beta^2}-\delta\big)\int_{-\infty}^{\infty}\exp\Big(-\beta x^2\Big)dx =$$
$$= A\exp\big(\frac{\gamma^2}{4\beta^2}-\delta\big)\sqrt{\frac{\pi}{\beta}}.$$

When writing out A and inserting the values of β, γ and δ, we see that $P(z)$ - after some simplification - takes on the expression

$$P(z) = A\exp\big(\frac{\gamma^2}{4\beta^2}-\delta\big)\sqrt{\frac{\pi}{\beta}} = \cdots =$$
$$= \frac{1}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y)}}\exp\Big(-\frac{1}{2}\frac{z^2}{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}\Big) = \frac{1}{\sqrt{2\pi\sigma_z^2}}\exp\Big(-\frac{z^2}{2\sigma_z^2}\Big).$$

We see that in the PDF for z, $\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$ and $\mu_z = 0$. $\mu_z = 0$ comes from the initial coordinate shift $\mu_x = \mu_y = 0$. Due to back transformation to the original coordinates, we get the relation $\mu_z = \mu_x - \mu_y$. These are the desired results of the exercise.

4.4 Exercise 13

Discussion

When one random variable is distributed around another random variable – for example as in this exercise where we have a measurement of a process that follows a gaussian distribution $N(x|x_0, \sigma_x^2)$ and then these measurements go through another gaussian $N(y|0, \sigma_y)$ – the way to get the final distribution is to convolute the two PDFs.

The **convolution** of two functions $f(\tau)$ and $g(\tau)$ is defined as

$$[f * g](t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau.$$

A convolution of two functions can also be performed with a product of the two Fourier transforms of the functions in the frequency domain:

$$[f * g](t) = \mathcal{F}^{-1}\big(\mathcal{F}(f)\mathcal{F}(g)\big), \tag{14}$$

where the Fourier transform $\mathcal{F}\big(f(x)\big)(k) = \hat{f}(k)$ is defined as

$$\hat{f}(k) = \int_{-\infty}^{\infty} f(x)e^{-2\pi ixk}dx. \tag{15}$$

In this exercise, the convolution will be calculated via the use of Fourier transforms.

The exercise

From the exercise, we understand that the "real" data is generated by an underlying normal distribution $f(x) = N(x_0, \sigma_x^2)$ which is then measured up by some device that takes every value and runs it through another normal distribution $g(y) = N(0, \sigma_y^2)$. We would like to obtain $h(z) = [f * g](z)$. To do this, we make use of Eqs. 14 and 15 and start off by taking the fourier transform of $f(x)$:

$$\begin{aligned}\hat{f}(k) = \int_{-\infty}^{\infty} f(x)e^{-2\pi ixk}dx = \frac{1}{\sqrt{2\pi\sigma_x^2}}\int_{-\infty}^{\infty} e^{-\frac{(x-x_0)^2}{2\sigma_x^2}}e^{-2\pi ixk}dx = \big[x' = x - x_0\big] = \\ = \frac{1}{\sqrt{2\pi\sigma_x^2}}\int_{-\infty}^{\infty} e^{-\frac{x'^2}{2\sigma_x^2}}e^{-2\pi i(x'+x_0)k}dx = \frac{e^{-2\pi ix_0k}}{\sqrt{2\pi\sigma_x^2}}\int_{-\infty}^{\infty} e^{-\frac{x'^2}{2\sigma_x^2}}e^{-2\pi ix'k}dx = \\ = \frac{e^{-2\pi ix_0k}}{\sqrt{2\pi\sigma_x^2}}\int_{-\infty}^{\infty} e^{-\frac{x'^2}{2\sigma_x^2}}\big(\cos(2\pi x'k) - i\sin(2\pi x'k)\big)dx.\end{aligned}$$

The last step was taken using Euler's formula. Note that the sine part of the expression constitutes an uneven function integrated over an even range and is equal to zero. We therefore have

$$\hat{f}(k) = \frac{e^{-2\pi ix_0k}}{\sqrt{2\pi\sigma_x^2}}\int_{-\infty}^{\infty} e^{-\frac{x'^2}{2\sigma_x^2}}\cos(2\pi x'k)dx.$$

Now, this is a standard integral

$$\int_{-\infty}^{\infty} e^{-at^2}\cos(2bt)dt = \sqrt{\frac{\pi}{a}}e^{-\frac{b^2}{a}},$$

if we set $a = \frac{1}{2\sigma_x^2}, b = \pi k$. Making use of this, we end up with

$$\hat{f}(k) = \frac{e^{-2\pi i x_0 k}}{\cancel{\sqrt{2\pi\sigma_x^2}}}\cancel{\sqrt{2\pi\sigma_x^2}}\, e^{-2\sigma_x^2\pi^2k^2} = e^{-2\pi i x_0 k}e^{-2\sigma_x^2\pi^2k^2}. \tag{16}$$

This is the shape that all gaussians take in the frequency domain, where the first factor represents the mean of the distribution and the second factor contains the variance. Therefore, the result of doing the Fourier transform of $g(y)$ must be (and is)

$$\hat{g}(k) = e^{-2\sigma_y^2\pi^2k^2}.$$

Multiplying $\hat{f}$ with $\hat{g}$ we get

$$\hat{f}(k)\hat{g}(k) = e^{-2\pi i x_0 k}e^{-2(\sigma_x^2+\sigma_y^2)\pi^2k^2}. \tag{17}$$

Now we compare the shape of $\hat{f}(k)\hat{g}(k)$ in Eq. 17 with that of the generic gaussian in Eq. 16 we see that $\hat{f}(k)\hat{g}(k)$ must be the Fourier transform of a gaussian with $h(z) = N(x_0, \sigma_x^2 + \sigma_y^2)$ as its PDF.

So, the answer to the question "What is the distribution of the observed quantity?" is $h(z) = N(x_0, \sigma_x^2+\sigma_y^2)$, a gaussian with the same mean as the underlying data-generating distribution and variance being the addition of the two variances.

4.5 Exercise 14

Discussion

The exercise

a) The posterior probability density for the parameters $\boldsymbol{\lambda} = \{A, B, C\}$ and the measured cross sections $\{\sigma_i^{(cs)}\}, i = 1, \ldots, 5$ is given by

$$P(\boldsymbol{\lambda}|\{\sigma_i^{(cs)}\}) = \frac{\prod_{i=1}^5 P(\sigma_i^{(cs)}|\boldsymbol{\lambda})\cancel{P_0(\boldsymbol{\lambda})}}{\int \prod_{i=1}^5 P(\sigma_i^{(cs)}|\boldsymbol{\lambda})\cancel{P_0(\boldsymbol{\lambda})}d\boldsymbol{\lambda}} \propto \prod_{i=1}^5 P(\sigma_i^{(cs)}|\boldsymbol{\lambda}) =$$
$$= \prod_{i=1}^5 \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\Big(-\frac{\big(\sigma_i^{(cs)} - \sigma(\theta_i|\boldsymbol{\lambda})\big)^2}{2\sigma_i^2}\Big),$$

where σ_i is the quoted standard deviation for measurement x_i.

b) To get the mode of the posterior, a grid search was made over A, B, C, starting over the interval $[-20, 20]$ for all three parameters and successively narrowing down. The mode was found to be at

$$A = -15.356 \quad B = -1.077 \quad C = -2.567.$$

5 Chapter 5

5.1 Exercise 1

Discussion

The data given in this exercise consists of a measured number of successes, $r_i^{(D)}$, assumed to follow a binomial distribution $\text{Bin}(p)$, given $N_i = N = 100$ number of trials for eight different energy levels of the experiment $E_i, i = 1, \ldots, 8$. We are interested in knowing the success rate p of the binomial distribution as a function of energy level E and we note that the curve when plotted looks like a sigmoid curve, see Fig. 5.

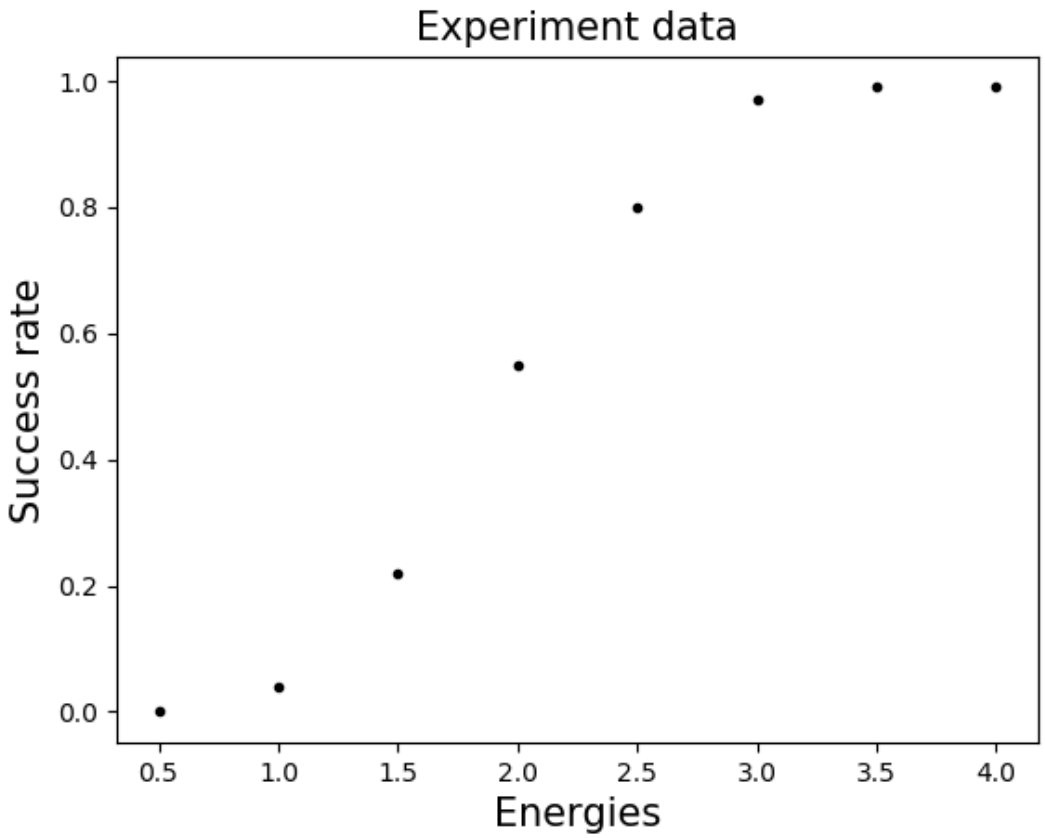

Figure 5: The measured success rates seem to follow a sigmoidal curve.

We therefore decide to model p with a general expression for the sigmoidal curve $p(E|A, E_0) = \sigma(E|A, E_0) = \frac{1}{1+\exp\left(-A(E-E_0)\right)}$ with A to control the steepness of the function and E_0 to control where the function reaches success rate 0.5. When estimating the values of A, E_0 we will take both a Bayesian approach and a frequentist approach.

Bayesian: Assuming that the experimental data follows a binomial distribution with the success rate set by $\sigma(E|A, E_0)$, we can form the likelihood of the data

$$P(\{r_i^{(D)}\}|A, E_0) = \prod_{i=1}^{8} \binom{N}{r_i^{(D)}} \sigma(E_i|A, E_0)^{r_i^{(D)}} \left(1 - \sigma(E_i|A, E_0)\right)^{N - r_i^{(D)}}. \tag{18}$$

To create the posterior we need to assign prior probabilities for A, E_0. To make the data fit as good as possible, which is the real goal here, we simply look at the data and consider which values of the parameters that are most likely and then set priors after that.

Frequentist: We will make use of a so called **test statistic** to determine which values are more or less likely. A test statistic (TS from now on) is a scalar valued parameter that indicates how likely a particular outcome is. With the use of this TS we can determine which regions of (A, E_0) are likely to contain the true values of the parameters.

We choose our TS to be the product of the probabilities of 8 data points, one per energy level (the likelihood):

$$\xi(\{r_i\}; A, E_0) = \prod_{i=1}^{8} \binom{N}{r_i} \sigma(E_i|A, E_0)^{r_i} \big(1 - \sigma(E_i|A, E_0)\big)^{N-r_i}$$

To construct a CL $1 - \alpha$ region for A, E_0, we make a grid over different values for both parameters. At each grid point our $\sigma(E|A, E_0)$ function will determine the success probability for the binomial distribution. For the eight different energies of interest, we randomly generate 8 datapoints r_i, one for each energy level, using the success probability specified by σ. We then determine $\xi(\{r_i\}; A, E_0)$ for these data points and store ξ in a vector. The larger the value of ξ, the more probable the data. This is repeated 100 times and the vector is then sorted.

Now, we construct ξ using our eight real data points and see if $\xi(\{r_i^{(D)}; A, E_0\})$ is in the most $1 - \alpha$ top level of the vector with all ξ values stored. If it is, that means that this combination of A, E_0 is in the $1 - \alpha$ CL region.

The above procedure is repeated for every grid point and the grid is then visualised to show the $1 - \alpha$ CL region.

The exercise

a) Before setting up the posterior, priors are needed. Gaussian priors for both A and E_0 were selected. First we construct the prior for E_0. We note that the success rate is about 0.5 for $E \approx 2$ and so we set $\mu_{E_0} = 2$ and $\sigma_{E_0} = 0.3$. The standard deviation is more or less arbitrarily set.

For A, we notice that the increase in success rate from $2E$ to $2.5E$ is 0.25, which means that the derivative of the function at $E = 2$ should be roughly 0.5. The expression for the derivative is

$$\left.\frac{\partial \sigma(E|A, E_0)}{\partial E}\right|_{E=E_0} = \left.\frac{A\exp(-A(E - E_0))}{\big(1 + \exp(-A(E - E_0))\big)^2}\right|_{E=E_0} = \frac{A}{4} \overset{!}{=} 0.5 \Leftrightarrow \underline{A = 2},$$

and the standard deviation is again arbitrarily set to $\sigma_A = 0.5$.

So we now have

$$P_0(A) = N(A|\mu_A = 2, \sigma_A^2 = 0.5^2)$$
$$P_0(E_0) = N(E_0|\mu_{E_0} = 2, \sigma_{E_0}^2 = 0.3^2)$$

and with the likelihood from Eq. 18 we can form the posterior:

$$P(A, E_0|\{r_i\}) \propto \prod_{i=1}^{8} \binom{N}{r_i} \sigma(E_i|A, E_0)^{r_i} \big(1 - \sigma(E_i|A, E_0)\big)^{N-r_i} N(2, 0.25) N(2, 0.09),$$

where the evidence as usual is taken to be constant. The posterior is then evaluated over an area in parameter space containing all significant values for the posterior and normalised.

The posterior can be seen in the left panel of Fig. 6. The modes of A, E_0 that gives rise to the largest value for the posterior are $(A^*, E_0^*) = (2.626, 1.973)$. The curve $\sigma(E|A^*, E_0^*)$ can be seen in the right panel of Fig. 6 together with the experimental data points.

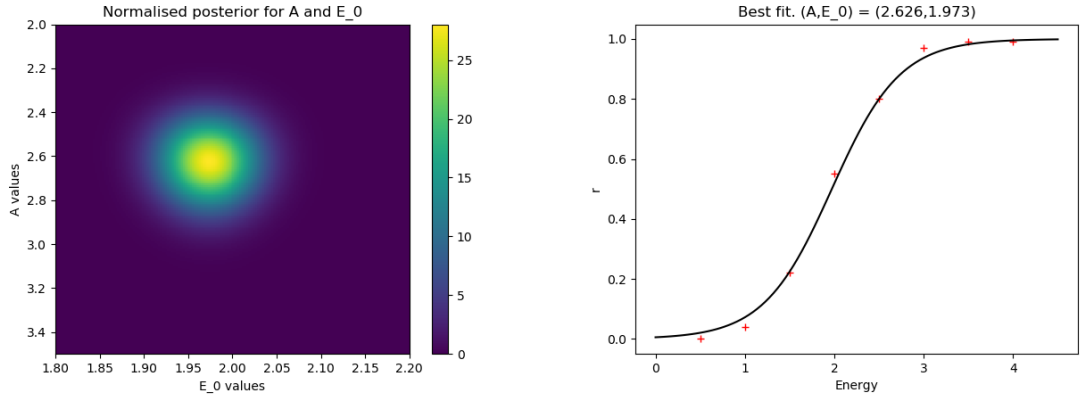

Figure 6: **Left**: The posterior of A and E_0. **Right:** The modes of A, E_0 and their corresponding success rate curve. The red pluses are the observed data.

b) The procedure outlined in the discussion above was performed for $1 - \alpha = 0.68$ and 100 ξ values per grid point. The 68% CL region found through this procedure is shown in Fig. 7

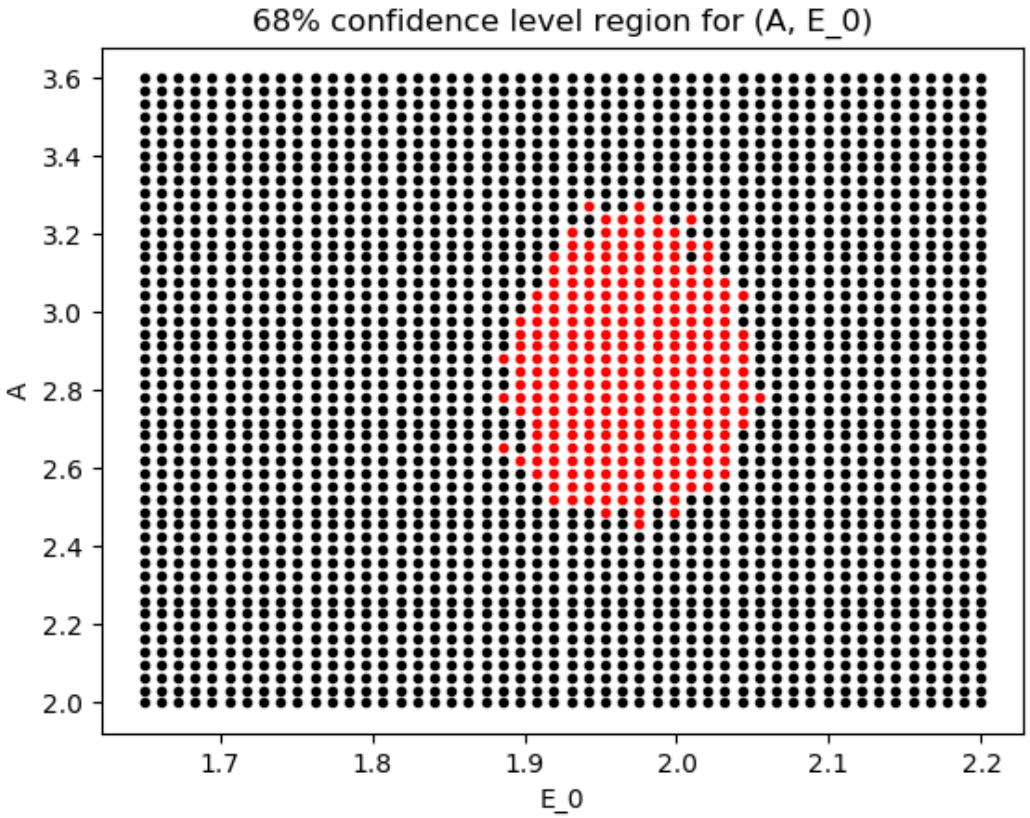

Figure 7: 68% credible region for the parameters A, E_0

5.2 Exercise 2

a) The new modelling function for the success rate p is $p(E|A, E_0) = \sin(A(E - E_0))$. Since my confidence about the values of the parameters were lower in this case, flat priors were used. Flat priors make the posterior be proportional to the likelihood, as we've seen before. And in this case, the posterior is

$$P(A, E_0|N, \{r_i\}) = \frac{P(\{r_i\}|N, A, E_0)P_0(A)P_0(E_0)}{\int P(\{r_i\}|N, A, E_0)P_0(A)P_0(E_0)dAdE_0} \propto P(\{r_i\}|N, A, E_0) = \quad (19)$$

$$= \prod_{i=1}^{8} \binom{N}{r_i} p(E|A, E_0)^r (1 - p(E|A, E_0))^{N-r_i}. \quad (20)$$

Since $p = \sin(A(E - E_0))$ can take on negative values which make no sense from a probabilistic point of view in Eq. 19, in those instances where that happened, p was set to 0 instead.

The posterior can be seen in the left panel of Fig. 8. The modes of A, E_0 are $(A^*, E_0^*) = (2.626, 1.973)$. The curve $\sigma(E|A^*, E_0^*)$ can be seen in the right panel of Fig. 8 together with the experimental data points.

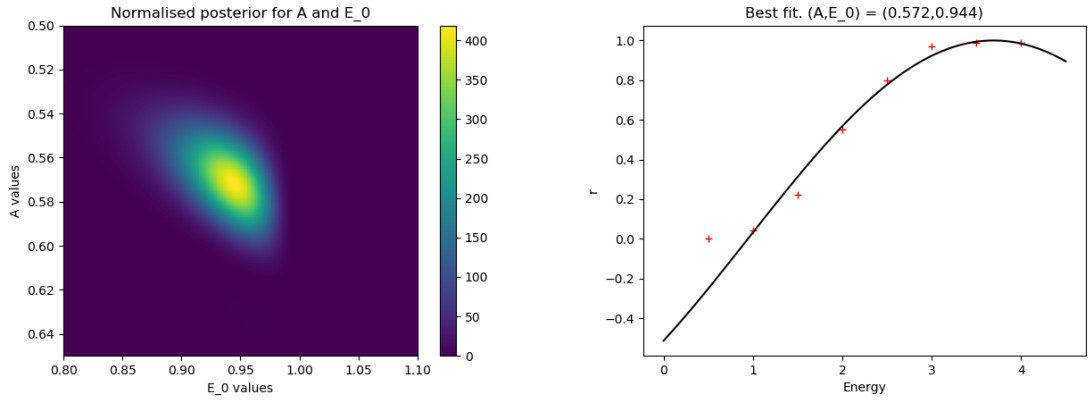

Figure 8: **Left**: The posterior for A and E_0 with $p(E|A, E_0) = \sin(A(E - E_0))$. **Right:** A^*, E_0^* and their corresponding success rate curve. The red pluses is the observed data.

b) Repeating exercise 5.1b with $p(E|A, E_0) = \sin(A(E - E_0))$ produces a plot without any grid points in the 68% CL region, at all. When increasing $1 - \alpha$ to 0.98, a small number of points appear. The two results are shown in Fig. 9.

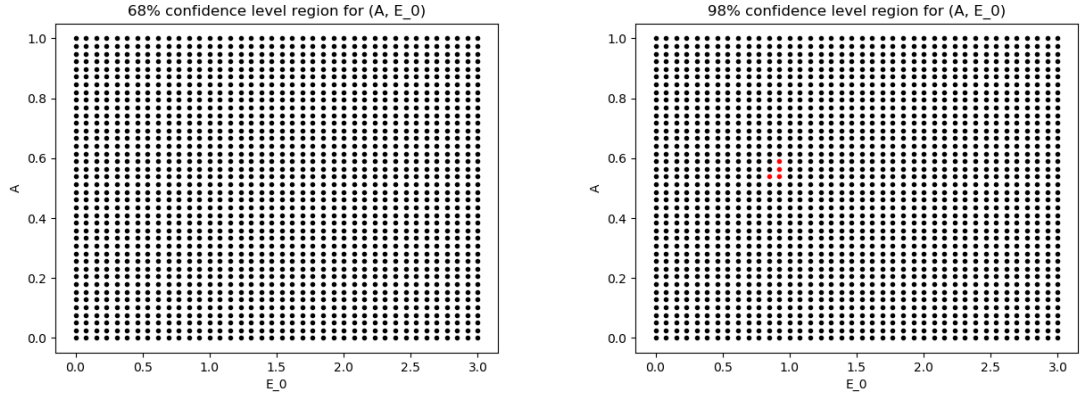

Figure 9: The $1-\alpha$ CL region (shown in red) with $p(E|A, E_0) = \sin(A(E - E_0))$. **Left**: $1 - \alpha = 0.68$. **Right:** $1 - \alpha = 0.98$.

c) It is obvious from the results that the sigmoid function explains the data much better than the sine, both from an intuitive sense since the sine can produce negative probabilities and from the results at hand; the sigmoid produces a much better fit to the data than does the sine. Also, the CL regions speak for themselves with the 68% CL region for the sine being completely empty.

It is interesting to note that the posterior functions in and of themselves to not tell us whether or not a model is right or wrong, they are always normalised and look the same because we make them that way. It is not until we check the fit of the model to the data that we can say anything about the result. The CL regions, in contrast, give us an indication even at the stage of just constructing the regions: If the region is completely empty, the model is probably not the right choice.

5.3 Exercise 3

Discussion

The χ^2 test statistic is in general defined as the weighted sum of the squared distances between the model prediction and the data:

$$\chi^2 = \sum_i \frac{(y_i - f(x_i|\boldsymbol{\lambda}))^2}{\omega_i^2}.$$

f is our model for the data, $\boldsymbol{\lambda}$ our parameters and a common choice for the weights ω_i^2 is the expected variance of the data points.

For data $\{y_i\}$ assumed to follow a gaussian distribution around some function $f(x_i|\boldsymbol{\lambda})$ with variances $\sigma_i(x, |\boldsymbol{\lambda})$, we define the **canonical** χ^2 to be

$$\chi^2 = \sum_i \frac{(y_i - f(x_i|\boldsymbol{\lambda}))^2}{\sigma_i^2}. \tag{21}$$

An interesting thing about this definition is that the PDF of χ^2 defined in this way will not depend on the parameters at all, it's universal for all experiments. Let's show this for one measurement:

$$P(\chi^2)\left|\frac{d\chi^2}{dy}\right| = 2P(y) \quad y \geq f(x|\boldsymbol{\lambda})$$
$$\frac{d\chi^2}{dy} = \frac{2(y - f(x|\boldsymbol{\lambda}))}{\sigma^2}$$
$$\left|\frac{d\chi^2}{dy}\right| = \frac{2\sqrt{\chi^2}}{\sigma}$$
$$P(\chi^2) = 2\frac{\sigma}{2\sqrt{\chi^2}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\chi^2/2} = \frac{1}{\sqrt{2\pi\chi^2}}e^{-\chi^2/2}.$$

The same result can be shown for any number of measurements n (or degrees of freedom if one or more of the parameters have been used to optimize χ^2):

$$P(\chi^2|n) = \frac{(\chi^2)^{n/2-1}}{2^{n/2}\Gamma(n/2)}e^{-\chi^2/2}.$$

This exercise is about deriving the mean variance and mode for the canonical χ^2 distribution for one measurement, let's do that.

The exercise

The probability distribution is

$$P(\chi^2) = \frac{1}{\sqrt{2\pi\chi^2}}e^{-\chi^2/2}.$$

Mean

$$E[\chi^2] = \int_0^\infty \frac{\chi^2}{\sqrt{2\pi\chi^2}}e^{-\chi^2/2} = \frac{1}{\sqrt{2\pi}}\int_0^\infty \sqrt{\chi^2}e^{-\chi^2/2} = \frac{1}{\sqrt{2\pi}}\sqrt{2\pi} = 1$$

Variance We first need $E[\chi^{2^2}]$:

$$E[\chi^{2^2}] = \int_0^\infty \frac{\chi^{2^2}}{\sqrt{2\pi\chi^2}} e^{-\chi^2/2} = \frac{1}{\sqrt{2\pi}} \int_0^\infty \chi^{2^{3/2}} e^{-\chi^2/2} = \frac{1}{\sqrt{2\pi}} 3\sqrt{2\pi} = 3$$

The variance is:

$$\mathrm{Var}[\chi^2] = E[\chi^{2^2}] - E[\chi^2]^2 = 2$$

Mode Take the derivative of the distribution to find the mode:

$$\frac{d}{d\chi^2}(\chi^2) = \frac{d}{d\chi^2}\Big(\frac{1}{\sqrt{2\pi\chi^2}} e^{-\chi^2/2}\Big) = \cdots = -\frac{1}{\sqrt{8\pi\chi^2}} e^{-\chi^2/2}\Big(1 + \frac{1}{\chi^2}\Big) \stackrel{!}{=} 0.$$

We can see that there is no point where the derivative is zero except at $\chi^2 = -1$, but the distribution is only defined for $[0, \infty)$. That, together with the observation that the derivative is negative everywhere tells us that the mode is $\chi^{2^*} = 0$.

5.4 Exercise 8

When comparing different models with one another, there are several ways of going about. One way is to define the **Bayes factor** for two models. In this exercise, we will compare which of two models best explain a dataset. M_1 is the event in which model 1 is correct and M_2 is the event in which model 2 is correct. The prior odds for these two models is defined as the ratio between the priors for both models

$$\mathcal{O}_0 = \frac{P_0(M_1)}{P_0(M_2)}.$$

The magnitude of the prior odds is determined on a case to case basis, given the situation.

The posterior odds are defined as

$$\mathcal{O} = \frac{P(M_1)}{P(M_2)} = \frac{P(D|M_1)P_0(M_1)}{P(D|M_2)P_0(M_2)} = B(M_1, M_2)\mathcal{O}_0,$$

with

$$B(M_1, M_2) = \frac{P(D|M_1)}{P(D|M_2)}$$

being the Bayes factor. The Bayes factor tells us how we should update our beliefs regarding which of the two models is more likely to be correct after the experiment has been carried out. When evaluating B, we make use of the **law of total probability** that states $P(A) = \int P(A|B)P(B)dB$:

utvidga definitionen när du orkar

$$P(D|M_1) = \int P(D|\boldsymbol{\lambda}, M_1)P(\boldsymbol{\lambda}|M_1)d\boldsymbol{\lambda},$$

with $\boldsymbol{\lambda}$ being the parameters of the model. I chose to use the prior distribution for $P(\boldsymbol{\lambda}|M_1)$.

The **p-value** of an observed test statistic ξ^D is defined as the probability of observing the value at hand or smaller/larger:

$$p = F(\xi) = \int_{\xi_{min}}^{\xi^D} P(\xi) d\xi, \quad \text{or} \tag{22}$$

$$p = 1 - F(\xi) = \int_{\xi^D}^{\xi_{max}} P(\xi) d\xi. \tag{23}$$

Which one of the two definitions that should be used depends on the test statistic and the problem at hand. In this exercise it makes sense to use the first of the two since the χ^2 test statistic is used and this has a larger value for more improbable values.

But what does the p-value tell us about how likely a specific model is to be true? To answer this, we need to obtain the distribution of the p-value for the *correct* model:

$$P(p|\lambda)\frac{dp}{d\xi} = P(\xi|\lambda)$$

and

$$\frac{dp}{d\xi} = \frac{d}{d\xi}\int_{\xi_{min}}^{\xi} P(\xi'|\lambda) d\xi' = P(\xi|\lambda),$$

which means that

$$P(p|\lambda) = 1, \quad p \in [0, 1].$$

Hence, the distribution is completely flat. So, any p-value has the same probability density as any other p-value, even very small ones. The thing that people do is make an arbitrary decision and say that "since the probability of observing a very tiny p-value is low and corresponds to a value of the test statistic far out in the tail of its distribution, we will reject models giving very small p-values". It is also possible to show that wrong models generally produce tiny p-values. These two arguments taken together is the reason why people use p-values.

A **marginalised distribution** is the distribution that results from ignoring one or more variables in a multivariate PDF/PMF. For a continuous bivariate distribution $f(x, y)$, one of the two possible marginalised distributions would be

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

a) We have a set of $N = 18$ data points and two different models to explain the data. One that only describes the data as noise

$$f(x|A, B, C) = A + Bx + Cx^2,$$

and one that describes it as noise plus a signal:

$$f(x|A, B, C, D, E, F) = A + Bx + Cx^2 + \frac{D}{(x - E)^2 + F^2}.$$

The event that the "background noise only" model is true is called M_1 and the event that the "background + signal" model is true is called M_2.

χ^2 is for this exercise defined according to Eq. 21. When minimizing the $\chi^{2,D}$, the python package "scipy.optimize" was used, specifically the function "minimize" from that package. The optimal parameters for both M_1 and M_2 are shown in Tab.5.

Model	Parameters	Parameter values
M_1	(A, B, C)	(-7.268, 173.5, -28.88)
M_2	(A, B, C, D, E, F)	(6.706, 56.79, 85.51, 0.1555, 0.4933, 0.06157)

Table 5: The parameter values minimizing $\chi^{2,D}$ for the two models.

Since 3, 6 parameters were used to optimize $\chi^{2,D}$ for M_1, M_2, respectively, the distribution that these values will follow is the one shown in Eq. 21 with n being the degrees of freedom. For M_1: $n = N - 3 = 15$ and for M_2: $n = N - 6 = 12$. The values of $\chi^{2,D}$ for the best fits in both models are shown in Fig. 10 (vertical lines) together with their respective distributions.

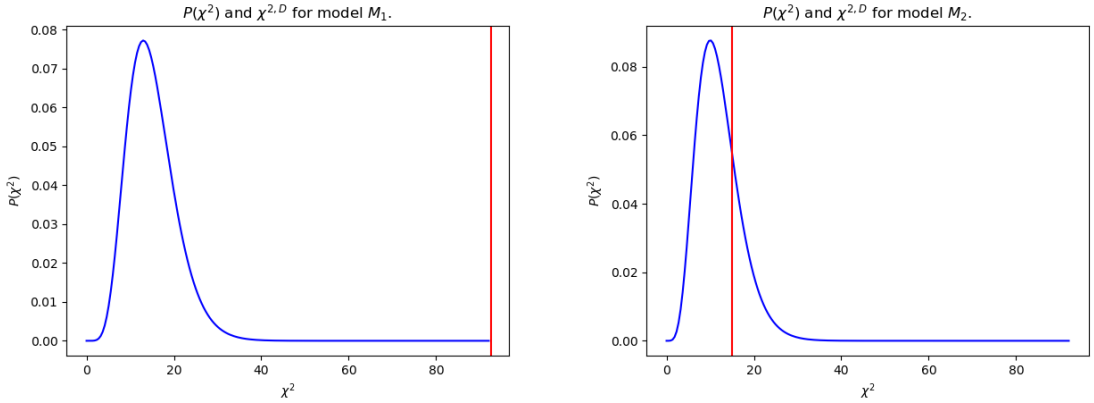

Figure 10: $\chi^{2,D}$ and its distribution for the two models. $\chi^{2,D}$ is shown as a vertical red line. **Left**: M_1. **Right:** M_2.

The two functions produced by our two best fit parameter values are shown in Fig. 11 together with the experimental data.

The values of $\chi^{2,D}$ are 3.70×10^{-13} for M_1 and 0.240 for M_2. This tells us that model M_1 is very unlikely to be the "true model" while M_2 very well might be. This is also reasonable given how much better the right function approximates the data in Fig. 11.

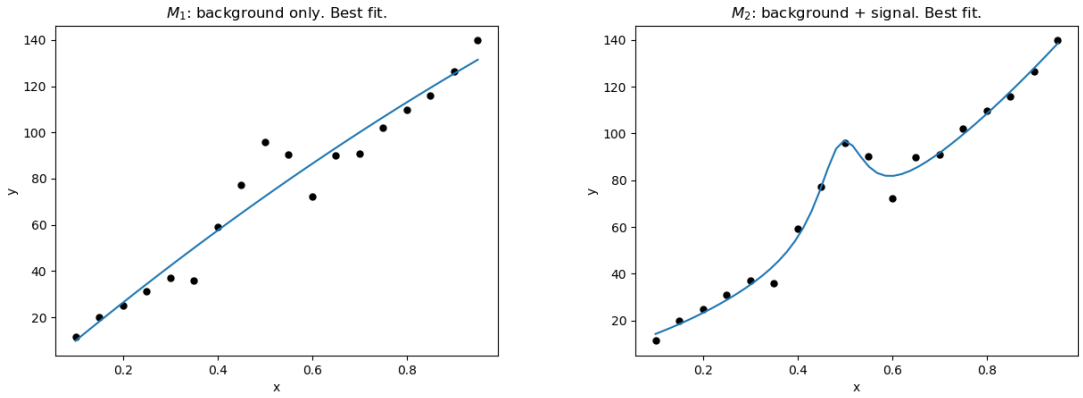

Figure 11: $\chi^{2,D}$ and its distribution for the two models. $\chi^{2,D}$ is shown as a vertical red line. **Left**: M_1. **Right:** M_2.

b) For both M_1 and M_2, the posterior was evaluated over a grid as fine as possible, normalised and then marginalised to get marginalised distributions. The marginalised distributions for the parameters of the noise-only model can be seen in Fig. 12 and the background plus noise parameter distributions in Fig.13. The best fit values for the parameters of both models can be seen in Tab. 6.

The Bayes factor, evaluated as described above, becomes

$$B(M_1, M_2) = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(D|\boldsymbol{\lambda}, M_1)P(\boldsymbol{\lambda}|M_1)d\boldsymbol{\lambda}}{\int P(D|\boldsymbol{\lambda}, M_2)P(\boldsymbol{\lambda}|M_2)d\boldsymbol{\lambda}} = \frac{1.25 \times 10^{-41}}{1.38 \times 10^{-31}} = 9.06 \times 10^{-11},$$

showing that whichever odds we had from the beginning should be updated to be a lot more in favor of M_2.

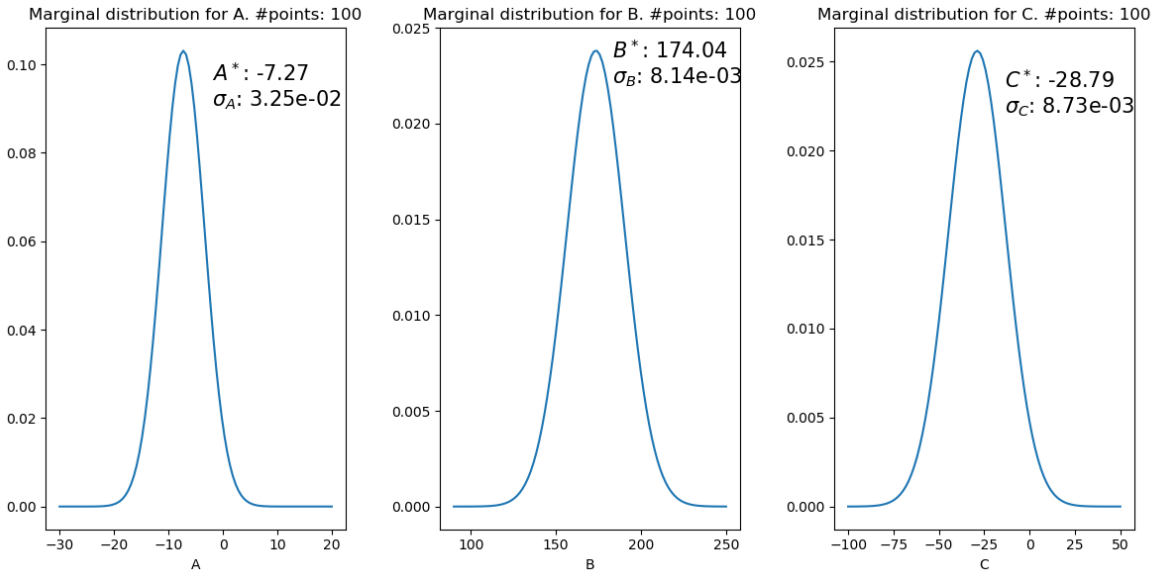

Figure 12: Marginalised distributions for the parameters A, B, C given M_1.

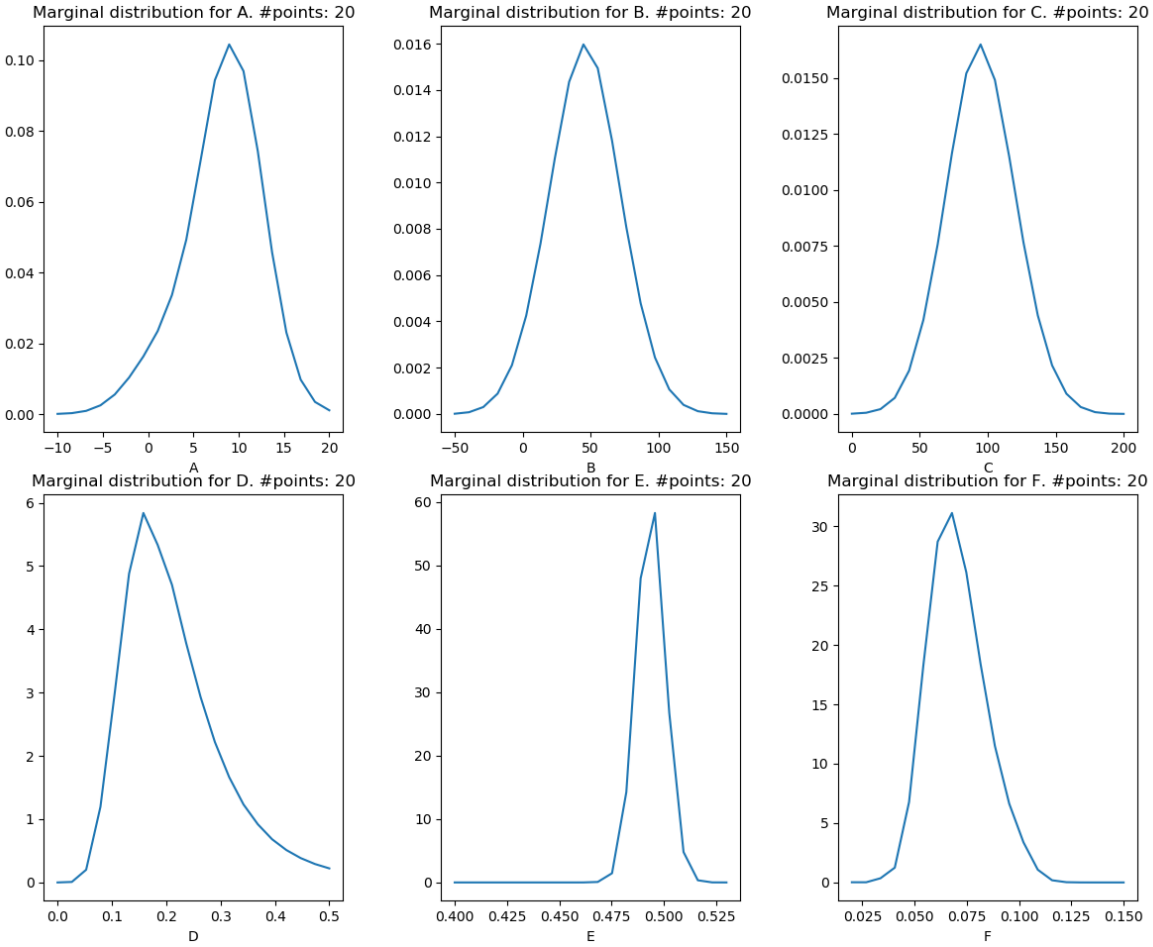

Figure 13: Marginalised distributions for the parameters A, B, C, D, E, F given M_2.

Model	Parameters	Parameter values
M_1	(A, B, C)	(-7.848, 175.1, -29.75)
M_2	(A, B, C, D, E, F)	(7.368, 55.26, 84.21, 0.1580, 0.4963, 0.06152)

Table 6: The parameter values giving the largest posterior for both models.

6 MLE chapter

6.1 Question 1

Discussion

The **Fisher information** tells us how quickly the PDF of a distribution is changing with respect to changes in a dependent variable. For a stochastic variable X with a PDF $P(X|\theta)$ it is defined as

$$I(\theta_0) = -E\Big[\Big(\frac{\partial^2 \ln P(X|\theta)}{\partial\, \theta^2}\Big)\Big]_{\theta_0} \tag{24}$$

or as

$$I(\theta_0) = E\Big[\Big(\frac{\partial \ln P(X|\theta)}{\partial\, \theta}\Big)^2\Big]_{\theta_0}. \tag{25}$$

The expressions are equivalent.

Maximum likelihood estimator

The **maximum likelihood estimation** (MLE) θ_{MLE} of a parameter θ is defined as the mode of the likelihood $\mathcal{L}(\theta)$, that is

$$\theta_{MLE} = \sup_{\theta} \mathcal{L}(\theta).$$

This estimator of θ is a consistent estimator for the true value of the parameter, θ_0. This means that θ_{MLE} converges in probability to θ_0. To show this, we first state the "Law of Large Numbers" (LLN) theorem:

> "For k independent and identically distributed (iid) random variables $X_1, \ldots, X_k$ and $E[X] < \infty$ we have
>
> $$\bar{X_k} = \frac{X_1 + \cdots + X_k}{k} \to E[X]$$
>
> in probability as $k \to \infty$."

We then take n measurements of an arbitrary distribution X with a likelihood dependent on one parameter θ so that $\mathcal{L}(\theta) = P(D|\theta) = \prod_{i=1}^{n} P(X_i|\theta)$ and we define

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln P(X_i|\theta).$$

Now, our MLE of θ is

$$\theta_{MLE} = \theta^* = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} L_n(\theta). \tag{26}$$

We also define, for a continuous distribution $P(X|\theta)$:

$$L(\theta) = E[\ln P(X|\theta)]_{\theta_0} = \int \{\ln P(x|\theta)\} P(x|\theta_0) dx.$$

The LLN now tells us that $L_n(\theta) \to L(\theta)$ as $n \to \infty$.

A summary of our results so far:

1. θ_{MLE} is the mode of $L_n(\theta)$ by definition.
2. θ_0 is the mode of $L(\theta)$.
3. LLN tells us that $L_n(\theta) \to L(\theta)$ as $n \to \infty$,

therefore, we expect that $\theta_0 \to \theta_{MLE}$ as $n \to \infty$. But how about the distribution of $\theta_{MLE} - \theta_0$? We start by making use of the mean value theorem:

$$\frac{f(a) - f(b)}{a - b} = f'(c), \quad c \in [a, b].$$

Let $f(\theta) = \frac{\partial L_n(\theta)}{\partial \theta} = L_n'(\theta)$ and $\theta_{MLE} = a, \theta_0 = b$. Since $f(\theta) = L_n'(\theta_{MLE}) = 0$ we have

$$0 = L_n'(\theta_0) + L_n''(\theta_1)(\theta_{MLE} - \theta_0), \quad \theta_1 \in [\theta_{MLE}, \theta_0]$$

and

$$(\theta_{MLE} - \theta_0) = -\frac{L_n'(\theta_0)}{L_n''(\theta_1)}. \tag{27}$$

We evaluate the numerator and denominator of Eq. 27 separately.

The denominator can, using LLN, be written as

$$L_N''(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 P(X_i|\theta)}{\partial\, \theta^2} \to E\Big[\Big(\frac{\partial^2 \ln P(X|\theta)}{\partial\, \theta^2}\Big)\Big]_{\theta_0}, \quad \text{as} \quad n \to \infty.$$

Since $\theta_1 \in [\theta_{MLE}, \theta_0]$, we also see that $\theta_1 \to \theta_0$ as $n \to \infty$. This means that, recalling the definition of the Fisher information in Eq. 24:

$$L_n''(\theta_1) \to L_n''(\theta_0) \to -I(\theta_0).$$

Now, let's look at the numerator of Eq. 27. Since θ_0 is the mode of $L(\theta)$, it follows that $L'(\theta_0) = 0$. With this in mind we can rewrite the numerator as

$$L_n'(\theta_0) = L_n'(\theta_0) - L'(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \Big(\frac{\partial \ln P(X_i|\theta)}{\partial\theta}\Big|_{\theta_0} - E\Big[\frac{\partial \ln P(X_i|\theta)}{\partial\theta}\Big|_{\theta_0}\Big]\Big).$$

Assuming that the distribution of the random quantity $\frac{\partial \ln P(X_i|\theta)}{\partial\theta}\Big|_{\theta_0}$ is well behaved, we make use of the CLT which gives us a gaussian PDF for $L_n'(\theta_0)$ with mean 0.

The variance of the ratio in Eq. 27 is

$$Var\Big[\frac{L_n'(\theta_0}{L_n''(\theta_0)}\Big] = \frac{Var\big[\frac{\partial P(X_i|\theta)}{\partial\theta}\big|_{\theta_0}\big]}{nI(\theta_0)^2} = \frac{1}{nI(\theta_0)},$$

where Eq. 25 was used in the last step.

Everything taken together, we find that asymptotically and for well behaved PDFs

$$P(\theta_{MLE} - \theta_0) = N(0, \frac{1}{nI(\theta_0)}). \tag{28}$$

The question

a) We have $p(x|p) = p^x(1-p)^{1-x}$. We first need the second derivative of the logarithm of the PDF:

$$\frac{\partial^2 \ln p(x|p)}{\partial p^2} = \frac{\partial}{\partial p}\Big(\frac{\partial}{\partial p}\big(\ln(p^x(1-p)^{1-x})\big)\Big) = \frac{\partial}{\partial p}\Big(\frac{\partial}{\partial p}\big(x\ln p + (1-x)\ln(1-p)\big)\Big) =$$
$$= \frac{\partial}{\partial p}\Big(\frac{x}{p} - \frac{1-x}{1-p}\Big) = -\Big(\frac{1-x}{(1-p)^2} + \frac{x}{p^2}\Big).$$

The Fisher information at the parameter value p_0 is the negative expectation value of the above quantity:

$$I(p) = -E\Big[\Big(\frac{\partial^2 \ln p(x|p)}{\partial^2 p}\Big)\Big]_{p_0} =$$
$$= -\Big(\frac{\partial^2 \ln p(x|p)}{\partial^2 p}\Big|_{x=1} \times p(x=1|p) + \frac{\partial^2 \ln p(x|p)}{\partial^2 p}\Big|_{x=0} \times p(x=0|p)\Big) =$$
$$= \frac{1}{p^2}p + \frac{1}{(1-p)^2}(1-p) = \boxed{\frac{1}{p(1-p)}}.$$

b) Say we have N data points $\{x_1, x_2, \ldots, x_N\}$ with $x_i \in \{0, 1\}$. The likelihood of this data is

$$\mathcal{L}(p) = \prod_{i=1}^{N} p^{x_i}(1-p)^{1-x_i}.$$

Let

$$T = \sum_i I(x=1) \quad \text{and} \quad F = \sum_i I(x=0) = N - T,$$

the likelihood then takes on the expression

$$\mathcal{L}(p) = p^T(1-p)^F = p^T(1-p)^{N-T}.$$

To find the maximum, differentiate with respect to p and set to zero. The p that satisfies this is p_{MLE}:

$$\frac{d\,\mathcal{L}(p)}{d\,p} = Tp^{T-1}(1-p)^{N-T} - p^t(N-T)^{N-T-1} \stackrel{!}{=} 0 \iff$$
$$\iff T(1-p) = p(N-T) \iff \boxed{p_{MLE} = \frac{T}{N}}.$$

c) As was shown in the discussion above, the distribution is given by Eq. 28 and

$$P(p_{MLE} - p_0) = N(0, \frac{1}{nI(p_0)}) = N(0, \frac{p_0(1-p_0)}{n}).$$

6.2 Exercise 2

a) The three plots can be seen in Fig. 14.

b) The expected distribution from the law of large numbers and the CLT is $P(\hat{\lambda}) = N(\lambda_0, \frac{1}{nI(p_0)})$. Judging from Fig. 14, we see that this seems to hold well in this case.

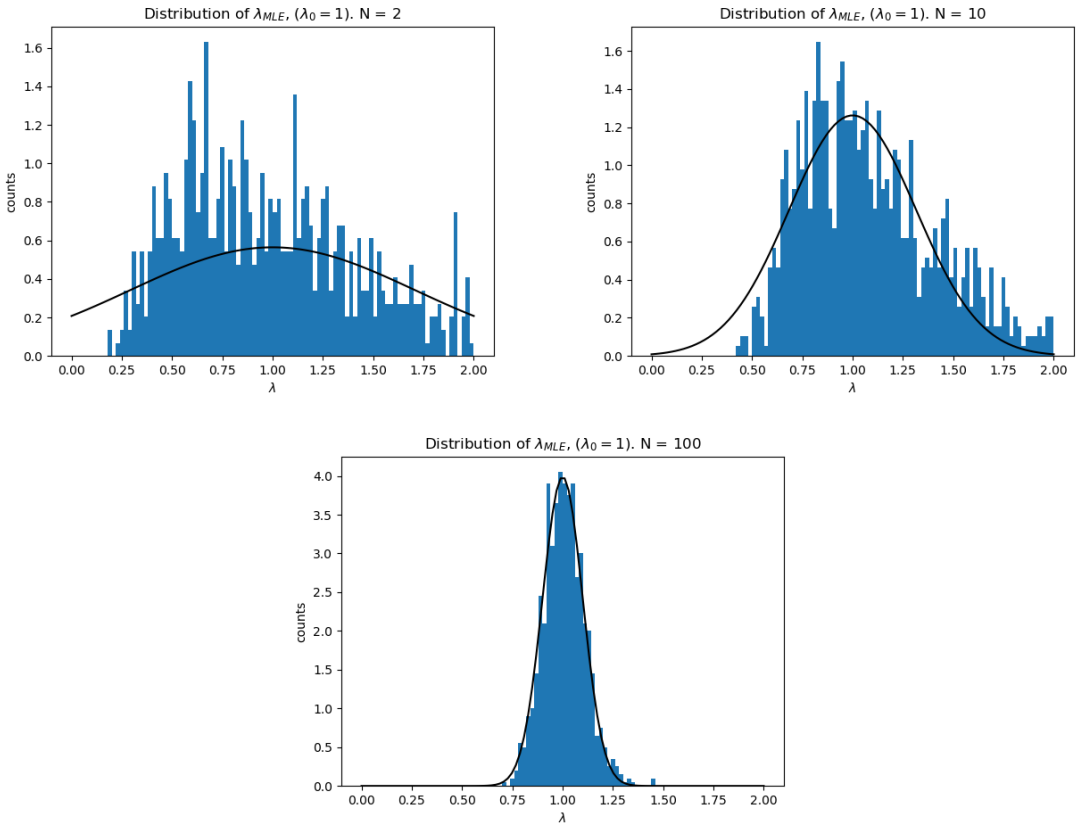

Figure 14: The distribution of the maximum likelihood estimator $\hat{\lambda}$. $\lambda_0 = 1$.