Technical university of Munich
Department of Physics

PH2221 Data Analysis

Final Coursework Report

Radu Bors

Matriculation Number: 03673933

April 7, 2018

Contents

1 Chapter 2 Exercises

1.1 Exercise 2.8

For the following function:

$$P(x) = xe^x \qquad\qquad 0 \leq x < \infty$$

(a) Find the mean and standard deviation. What is the probability content in the interval (mean - standard deviation, mean + standard deviation)?

(b) Find the median and 68% central interval.

(c) Find the median and 68% smallest interval.

a) The mean μ, also known as the expectation value $E[x]$ is given by:

$$E[x] = \int_0^\infty xP(x)dx$$

In our case, an analytical solution is possible:

$$\begin{aligned} E[x] &= \int_0^\infty xP(x)dx = \int_0^\infty x^2 e^{-x} dx \\ &= -e^{-x}(x^2 + 2x + 2)\Big|_0^\infty \\ &= 2 \end{aligned} \tag{1}$$

Due to the nature of the exponential of decreasing faster than the second degree polynomial as $x \to \infty$, the evaluation of the expression at ∞ is 0, and therefore the final result is 2. The standard deviation σ is defined as:

$$\begin{aligned} \sigma &= \sqrt{Var[x]} = \sqrt{E[x^2] - (E[x])^2} \\ &= \sqrt{6-4} \\ &= \sqrt{2} \end{aligned} \tag{2}$$

where $Var[x]$ is the variance. The probability content in the interval $(\mu - \sigma, \mu + \sigma)$ can be evaluated by integrating the function within the limits of the interval. Therefore, the probability content p is:

$$p = \int_{2-\sqrt{2}}^{2+\sqrt{2}} P(x)dx \approx 0.737 \tag{3}$$

where the integral has been computed using the `scipy.integrate.quad` function. The interval has been plotted using the `matplotlib` package, and is presented below in fig. 1.

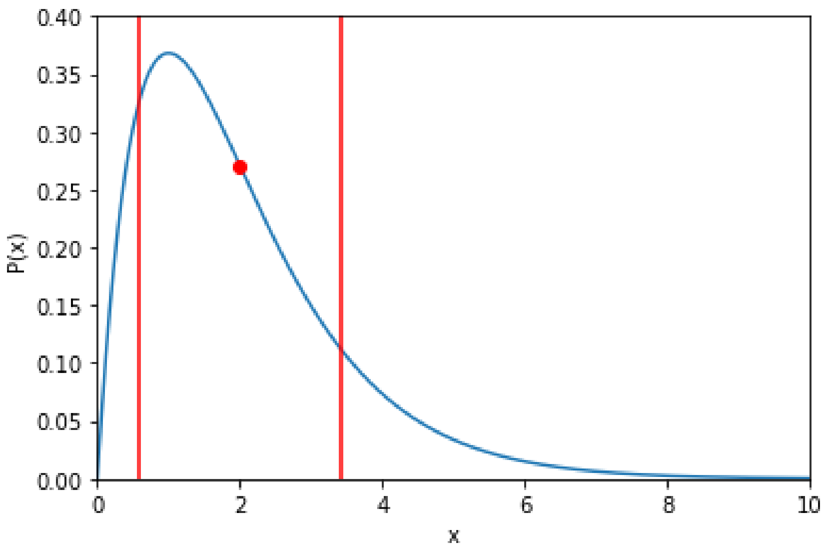


Figure 1: Plot of the given function, together with a red marker indicating the mean $\mu = 2$, and vertical bars to indicate the $(2+\sqrt{2}, 2-\sqrt{2})$ interval, with ≈ 0.737 probability content.

Computing the integral $\int_0^{10} P(x)dx$ numerically, the observed value is > 0.9995. Therefore, adding more values for $x > 10$ makes little sense, and the plot is considered representative for the $x \in [0, 10]$ interval.

b) the median m is defined as the value in the $[0, \infty)$ domain of our function, for which the equality holds:

$$\int_0^m P(x)dx = \int_m^\infty P(x)dx = 0.5 \tag{4}$$

In this case, to find m, a vector of 100001 equally spaced values in the $[0, 10]$ interval was generated using `numpy.linspace`. For each element e in the vector, the integral of $P(x)$ from 0 to e was computed, and substracted from 0.5. The smallest absolute value of this difference was observed for $m \approx 1.678$. The difference is:

$$0.5 - \int_0^m P(x)dx \approx 4.606 \times 10^{-5} \tag{5}$$

and therefore $m = 1.678$ was considered a good enough value for the median.

For a continuous function, such as the one in our case, the central $1-\alpha$ interval $[r_1, r_2]$ is defined for the values r_1 and r_2 which satisfy the set of conditions:

$$\begin{cases} \int_0^{r_1} P(x)dx = \frac{\alpha}{2} \\ \int_{r_2}^{\infty} P(x)dx = \frac{\alpha}{2} \end{cases} \tag{6}$$

In our case, setting $\alpha = 0.32$ (to get the 68% interval), and using the same technique of computing the integral at each point in our previously mentioned vector, yields the interval $[0.712, 3.289]$. Computing the difference:

$$0.68 - \int_{0.712}^{3.289} P(x)dx \approx -2.433 \times 10^{-5} \quad (7)$$

that our interval is acceptable. The plot in fig. 2 presents the median, and the central 68% interval.

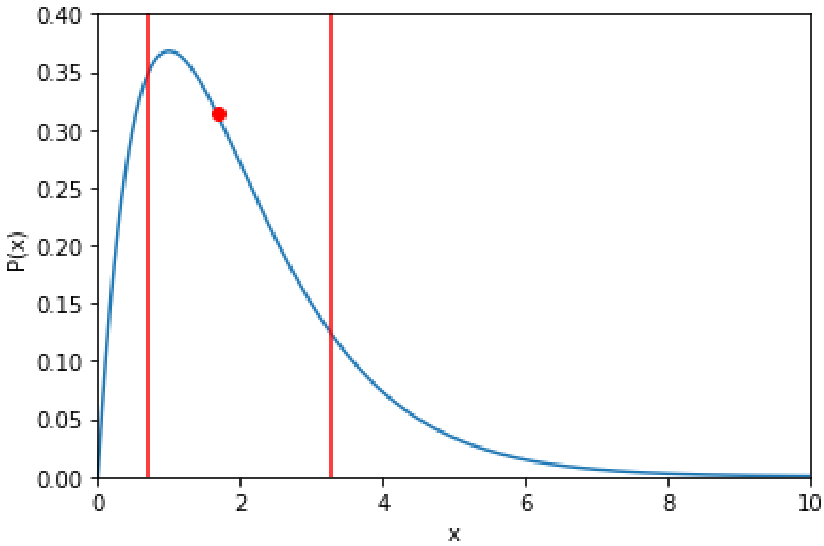


Figure 2: Plot of the given function, together with a red marker indicating the median $m = 1.678$, and vertical bars to indicate the central 68% interval.

c) The mode x^* is defined as the value of x for which our function $P(x^*)$ takes the maximum value. In our case, the mode is at $x^* = 1$, with $P(x^*) = 1$. To compute the smallest 68% interval, again a vector $\{x_i\}$ of 100001 equally spaced values in the $[0, 10]$ interval was generated, and $P(x_i)$ was computed at each of these values, resulting in the $\{P_i\}$ vector. Starting with the top value ($P_i = P(x^*)$), values were added together in descending order of P_i, until the sum was ≥ 0.68. The smallest interval is defined by the minimum and the maximum values of the corresponding $\{x_i\}$ values for which their $P(x_i)$ was added to the sum. In our case, the smallest 68% interval is $[0.271, 2.490]$. Computing the difference:

$$0.68 - \int_{0.271}^{2.490} P(x)dx \approx -1.241 \times 10^{-5} \quad (8)$$

which shows that the found interval is acceptable. The plot in fig. 3 shows the function, together with the mode $x^* = 1$ and vertical bars for the limits $[0.271, 2.490]$ of the smallest 68% interval.

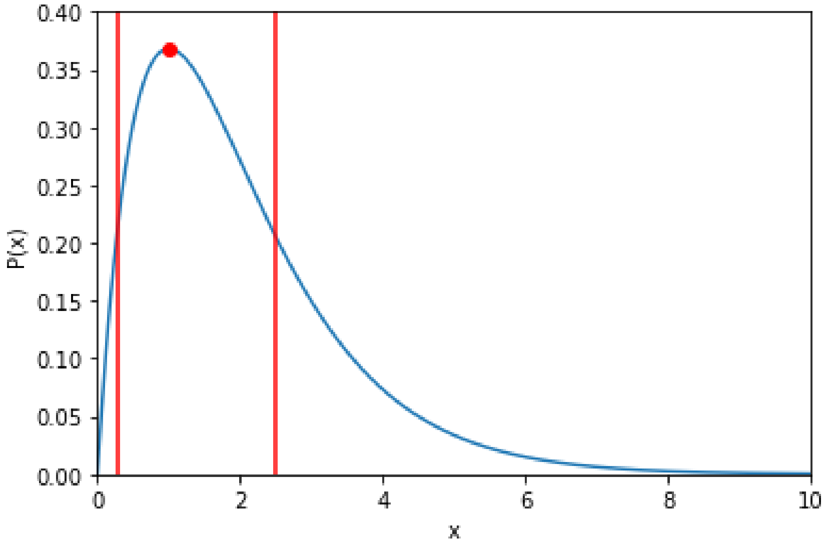


Figure 3: Plot of the given function, together with a red marker indicating the mode $x^* = 1$*, and vertical bars to indicate the smallest* 68% *interval.*

1.2 Exercise 2.10

Consider the data in the table: Starting with a flat prior for each energy, find an estimate for the efficiency (success parameter p) as well as an uncertainty. For the estimate of the parameter, take the mode of the posterior probability for p and use the smallest interval to find the 68% probability range. Make a plot of the result.

Table 1: Data for exercise 2.10

Energy	*Trials*	*Successes*
0.5	100	0
1.0	100	4
1.5	100	20
2.0	100	58
2.5	100	92
3.0	1000	987
3.5	1000	995
4.0	1000	998

The binomial distribution for r (number of successful trials), given N (total trials) and p, probability of success for an individual trial is given by:

$$P(r|N,p) = \frac{N!}{r!(N-r)!}p^r(1-p)^{(N-r)} \tag{9}$$

When applying Bayes Theorem to get the probability distribution of parameter p, given some data $D = r, N$, the equation is:

$$P(p|N,r) = \frac{P(r|N,p)P_0(p)}{P(r|N)} \tag{10}$$

where $P_0(p)$ is a prior distribution for parameter p. By law of total probability:

$$P(r|N) = \int_0^1 P(r|N)P(p)dp \tag{11}$$

which allows us to rewrite eq. 10 as:

$$P(p|N,r) = \frac{P(r|N,p)P_0(p)}{\displaystyle\int_0^1 P(r|N)P(p)dp} \tag{12}$$

For a flat prior $P_0(p) = 1$ (which is already normalized), and observing that $\displaystyle\int_0^1 p^r(1-p)^{N-r}dp$

is the β-function $\beta(r+1, N-r+1) = \dfrac{r!(N-r)!}{(N+1)!}$ the equation becomes:

$$P(p|N,r) = \frac{(N+1)!}{r!(N-r)!}\, p^r(1-p)^{N-r} \tag{13}$$

A vector $\{p_i\}$ of 100001 equally spaced values in the $[0,1]$ interval was generated, and $P(p_i|N_{energy}, r_{energy})$ was plotted for all sets of values of $\{N_{energy}, r_{energy}\}$ in table 1. The plots are presented in fig. 4.

To compute the smallest interval for a discrete variable, such as p, the same procedure as in *exercise 2.8* was applied. $P(p_i)$ was computed at each of the p_i values, resulting in the $\{P_i\}$ vector. Starting with the top value at mode p^* ($P_i = P(p^*)$), values were added together in descending order of P_i, until the sum was ≥ 0.68. The p_i values for which $P(p_i)$ was added to the sum were recorded. The smallest interval is defined by the minimum and the maximum values of these values. The smallest 68% intervals ($O^S_{0.68}$) and the modes p^* are presented in table 2 for each energy.

Table 2: Mode of posterior and smallest 68% *interval for exercise 2.10*

$Energy$	$Trials$	$Successes$	p^*	$O^S_{0.68}$
0.5	100	0	0.000	[0.000, 0.011]
1.0	100	4	0.040	[0.023, 0.063]
1.5	100	20	0.200	[0.162, 0.241]
2.0	100	58	0.580	[0.531, 0.628]
2.5	100	92	0.920	[0.890, 0.944]
3.0	1000	987	0.987	[0.983, 0.990]
3.5	1000	995	0.995	[0.992, 0.997]
4.0	1000	998	0.998	[0.996, 0.999]

As expected, the mode p^* for each energy is at $p^* = \dfrac{r_{data}}{N_{data}}$. By no coincidence, this is also the mode of the likelihood, since there is a direct connection between performing Bayesian Analysis with a flat prior, and a likelyhood analysis.

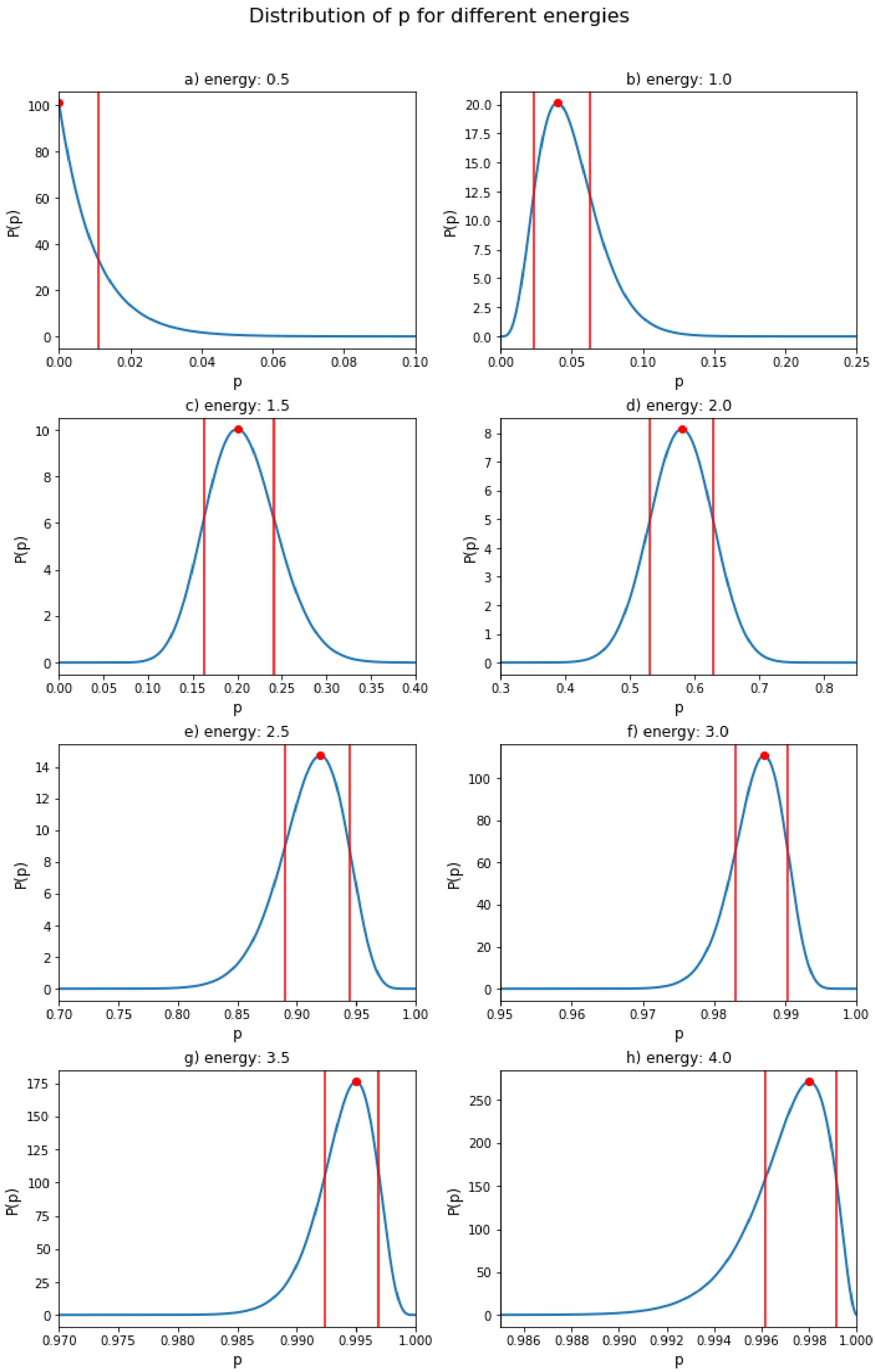


Figure 4: Plot of the posterior $P(p|N, r)$, for different energies, together with a red marker indicating the mode p^*, and vertical bars to indicate the smallest 68% interval.

1.3 Exercise 2.11

Analyze the data in the table from a frequentist perspective by finding the 90% confidence level interval for p as a function of energy. Use the Central Interval to find the 90% CL interval for p.

Table 3: Data for exercise 2.11

Energy	*Trials*	*Successes*
0.5	100	0
1.0	100	4
1.5	100	20
2.0	100	58
2.5	100	92
3.0	1000	987
3.5	1000	995
4.0	1000	998

The probability distribution of getting r successes out of N trials, given an individual trial success probability p is expressed by the binomial distribution in the equation:

$$P(r|N,p) = \frac{N!}{r!(N-r)!}p^r(1-p)^{(N-r)} \tag{14}$$

For a frequentist analysis in this case, one starts by constructing the Neyman Confidence Level, using the central 90% interval (O_{90}^C) for the construction. The central $1-\alpha$ interval ($O_{1-\alpha}^C$) is given by the smallest r_1 and r_2 values possible which satisfy the conditions:

$$\begin{cases} P(r < r_1|N,p) \leq \dfrac{\alpha}{2} \\ P(r < r_2|N,p) \leq \dfrac{\alpha}{2} \end{cases} \tag{15}$$

where r_1 and r_2 are defined by:

$$\begin{cases} r_1 = 0 \; if \; P(r=0|N,p) > \dfrac{\alpha}{2} \\ \\ r_1 = \sup\limits_{r\in\{0,1..N\}} \left[\displaystyle\sum_{i=0}^{r} P(i|N,p) \leq \dfrac{\alpha}{2}\right] + 1 \; else \end{cases} \tag{16}$$

$$\begin{cases} r_2 = N \; if \; P(r=N|N,p) > \dfrac{\alpha}{2} \\ \\ r_2 = \inf\limits_{r\in\{0,1..N\}} \left[\displaystyle\sum_{i=r}^{N} P(i|N,p) \leq \dfrac{\alpha}{2}\right] - 1 \; else \end{cases} \tag{17}$$

For the 90% central interval, one takes $\alpha = 0.05$ and treats the problem in 2 cases: for 100 trials and for 1000 trials. Several values have been considered for p, between 0 and 1 (both 0 and 1 included), with a step of 0.001. For each p in this range, r_1 and r_2 as defined above were computed and plotted in fig. 5.

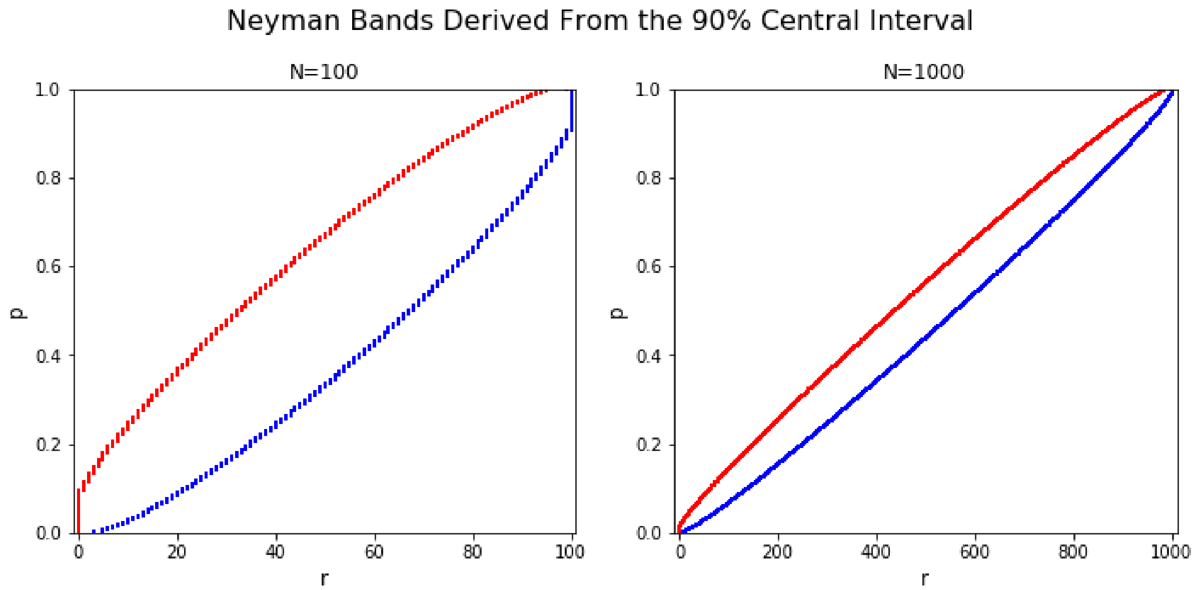


Figure 5: Neyman Construction, using the 90% central interval. The blue band indicates r_1 values, and the red band r_2 values for each p.

The confidence level at each r_{data} from table 3 is given by the minimum and maximum values of p for which $r_{data} \in [r_1, r_2]$ corresponding to that p. The results are presented in table 4.

Table 4: Confidence intervals of p for each r_{data}, for exercise 2.11

Energy	*Trials*	*Successes*	90% *CL*
0.5	100	0	[0, 0.095]
1.0	100	4	[0.002, 0.163]
1.5	100	20	[0.083, 0.368]
2.0	100	58	[0.402, 0.745]
2.5	100	92	[0.780, 0.986]
3.0	1000	987	[0.966, 0.997]
3.5	1000	995	[0.978, 1]
4.0	1000	998	[0.987, 1]

A comparison between the confidence levels for p and the central intervals extracted using Bayesian analysis yields nothing significant since the conditions for the probability contained in the intervals were different: 90% in this case and 68% in the case of the Bayesian Analysis.

1.4 Exercise 2.13

Let us see what happens if we reuse the same data multiple times. We have N trials and measure r successes. Show that if you reuse the data n times, starting at first with a flat prior and then using the posterior from one use of the data as the prior for the next use, you get:

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

What are the expectation value and variance for p in the limit $n \to \infty$?

The binomial distribution is described by:

$$P(r|N,p) = \frac{N!}{r!(N-r)!}p^r(1-p)^{(N-r)} \tag{18}$$

We will make use of mathematical induction to prove the equation for n uses of the same data. Starting with the base case $n=1$ and using a flat prior $P_0(p)=1$ for the Bayesian analysis:

$$\begin{aligned}
P_1(p|N,r) &= \frac{P(r|N,p)P_0(p)}{P(r|N)} \\
&= \frac{P(r|N,p)P_0(p)}{\displaystyle\int_0^1 P(r|N)P_0(p)dp} \\
&= \frac{\dfrac{N!}{r!(N-r)!}\,p^r(1-p)^{N-r}}{\dfrac{N!}{r!(N-r)!}\displaystyle\int_0^1 p^r(1-p)^{N-r}dp} \\
&= \frac{p^r(1-p)^{N-r}}{\displaystyle\int_0^1 p^r(1-p)^{N-r}dp}
\end{aligned} \tag{19}$$

where we note that the integral in the denominator of the final form is the β function $\beta(x,y) = \displaystyle\int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{(x-1)!(y-1)!}{(x+y-1)!}$. By changing notation $x \to r+1$, $y \to N-r+1$ and $t \to p$, the denominator can written as $\dfrac{r!(N-r)!}{(N+1)!}$.

Therefore:

$$P_1(p|N,r) = \frac{(N+1)!}{r!(N-r)!}\, p^r(1-p)^{N-r} \tag{20}$$

This is the already familiar case of Bayesian Analysis with a flat prior. Next, we move to the general case of the induction proof, where $P_{n-1}(p|N,r)$ is used as a prior to get the posterior $P_n(p|N,r)$. Our prior is described by the equation:

$$P_{n-1}(p) = \frac{[(n-1)N+1]!}{[(n-1)r]![(n-1)N-(n-1)r]!}\, p^{(n-1)r}(1-p)^{(n-1)(N-r)} \tag{21}$$

We apply Bayes theorem to get $P_n(p|N,r)$:

$$\begin{aligned} P_n(p|N,r) &= \frac{P(r|N,p)P_{n-1}(p)}{\int_0^1 P(r|N)P_{n-1}(p)dp} \\ &= \frac{p^r(1-p)^{N-r}\ C_{n-1}\ p^{(n-1)r}(1-p)^{(n-1)(N-r)}}{\int_0^1 p^r(1-p)^{N-r}\ C_{n-1}\ p^{(n-1)r}(1-p)^{(n-1)(N-r)}dp} \end{aligned} \tag{22}$$

where $C_{n-1} = \frac{[(n-1)N+1]!}{[(n-1)r]![(n-1)N-(n-1)r]!}$ was used for convenience of notation. Simplfication of C_{n-1} and regrouping the terms yields:

$$P_n(p|N,r) = \frac{p^{nr}(1-p)^{n(N-r)}}{\int_0^1 p^{nr}(1-p)^{n(N-r)}dp} \tag{23}$$

Again, noticing the β function in the denominator, and using notation $x \to nr+1$, $y \to n(N-r)+1$ and $t \to p$ one reaches the result:

$$P_n(p|N,r) = \frac{nN+1!}{(nr)!(nN-nr)!}\, p^{nr}(1-p)^{n(N-r)} \tag{24}$$

Upon a closer inspection of the result, one notices that this is the equivalent of taking a data set of nN trials and nr successes, and analyzing it with a flat prior. We have not gained any new knowledge because we have not added any new data. As we have seen, Bayesian analysis yields the same results in cases with the same knowledge. The expectation value $E_n[p]$ for such a posterior is:

$$
\begin{aligned}
E_n[p] &= \int_0^1 pP_n(p)dp \\
&= \frac{nN+1!}{(nr)!(nN-nr)!}\int_0^1 p^{nr+1}(1-p)^{n(N-r)}dp
\end{aligned}
\tag{25}
$$

where we again observe the β function. Using a property of this function $\beta(x+1,y) = \beta(x,y)\frac{x}{x+y}$, the expectation value becomes:

$$
E_n[p] = \frac{nr+1}{nN+2} \tag{26}
$$

Analogously, to find the expectation value of p^2, one uses:

$$
\beta(x+1,y) = \beta(x,y)\frac{(x+1)}{(x+1)+y}\frac{x}{x+y} \tag{27}
$$

and the same notation change as before: $x \to nr+1$, $y \to n(N-r)+1$ and $t \to p$. The result is:

$$
E_n[p^2] = \frac{nr+1}{nN+2}\frac{nr+2}{nN+3} \tag{28}
$$

and the found variance $Var_n[p] = E_n[p^2] - (E_n[p])^2$ is:

$$
\begin{aligned}
Var_n[p] &= \frac{nr+1}{nN+2}\frac{nr+2}{nN+3} - \frac{(nr+1)^2}{(nN+2)^2} \\
&= \frac{(nr+1)[(nr+2)(nN+2)-(nr+1)(nN+3)]}{(nN+3)(nN+2)^2} \\
&= \frac{(nr+1)(nN-nr+1)}{(nN+3)(nN+2)^2} \\
&= \frac{E_n[p](1-E_n[p])}{nN+3}
\end{aligned}
\tag{29}
$$

As $n \to \infty$, we consider the limits:

$$\lim_{n\to\infty} \frac{nr+1}{nN+2} = \lim_{n\to\infty} \frac{r+1/n}{N+2/n}$$
$$\lim_{n\to\infty} \frac{(nr+1)(nN-nr+1)}{(nN+3)(nN+2)^2} \tag{30}$$

In the first case, the expectation value $E[p]$ approaches the mode $p^* = r/N$. This is consistent with our prior knowledge that for large number of trials and successes, the Gaussian distribution is a good approximation for the binomial distribution. Since the expectation value for a Gauss distributed variable is also its mode, and the posterior when using a flat prior is still a Gaussian, this result is expected. For the second limit, we notice that n appears to a higher power in the denominator, and therefore the variance $Var[p]$ approaches 0.

2 Chapter 3 Exercises

2.1 Exercise 3.4

Consider the function $f(x) = \frac{1}{2}e^{|x|}$ for $-\infty < x < \infty$.

(a) Find the mean and standard deviation of x.

(b) Compare the standard deviation with the FWHM (Full Width at Half Maximum).

(c) What probability is contained in the ± 1 standard deviation interval around the peak ?

a) The mean, or expectation value $E[x]$, is defined as:

$$\begin{aligned} E_n[x] &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_{-\infty}^{\infty} x \, \frac{1}{2} \, e^{|x|} dx \\ &= \int_{-\infty}^{0} x \, \frac{1}{2} \, e^{x} dx + \int_{0}^{\infty} x \, \frac{1}{2} \, e^{-x} dx \\ &= 0 \end{aligned} \tag{31}$$

which is 0 due to the odd function $g(x) = xf(x)$, meaning that $g(-x) = -g(x)$. By comparison $f(x)$ is even (symmetric around the $x = 0$ line) as it can be seen in fig. 6. The standard deviation σ is defined as $\sqrt{Var[x]}$, where $Var[x]$ is the variance:

$$\begin{aligned} Var_n[p] &= E[x^2] - (E[x])^2 \\ &= 2 - 0 \\ &= 2 \end{aligned} \tag{32}$$

and therefore the standard deviation $\sigma = \sqrt{2}$.

b) The full width at half maximum (FWHM) is defined by the interval $x_2 - x_1$ for which $f(x_1) = 1/2 f(x^*)$ and $f(x_2) = 1/2 f(x^*)$, where x^* is the mode and $x_1 < x^* < x_2$. In our case, the mode is at $x^* = 0$, with value $f(x^*) = 0.5$. By solving for x in $f(x) = 1/4$, one gets $x_1 = -\ln(2)$ and $x_2 = \ln(2)$. Hence, the FWHM is $2\ln 2$. The $[x_1, x_2]$ interval has been indicated with vertical green bars in fig. 6. To find the probability contained in the $[x_1, x_2]$ interval, one needs to compute the integral

$$
\begin{aligned}
p_{FWHM} &= \int_{-\ln(2)}^{-\ln(2)} f(x)dx \\
&= \int_{-\ln(2)}^{-\ln(2)} \frac{1}{2}e^{-|x|}dx = 0.5
\end{aligned} \tag{33}
$$

c) The $\pm\sigma$ around the mode $x^* = 0$ has been indicated with red vertical bars in fig. 6. To find the probability contained in this interval, one needs to compute the integral:

$$
\begin{aligned}
p_{\pm\sigma} &= \int_{-\sqrt{2}}^{-\sqrt{2}} f(x)dx \\
&= \int_{-\sqrt{2}}^{0} \frac{1}{2}e^{x}dx + \int_{0}^{\sqrt{2}} \frac{1}{2}e^{-x}dx = \\
&= \left.\frac{e^{x}}{2}\right|_{-\sqrt{2}}^{0} + \left.\frac{e^{-x}}{2}\right|_{0}^{\sqrt{2}} \approx 0.757
\end{aligned} \tag{34}
$$

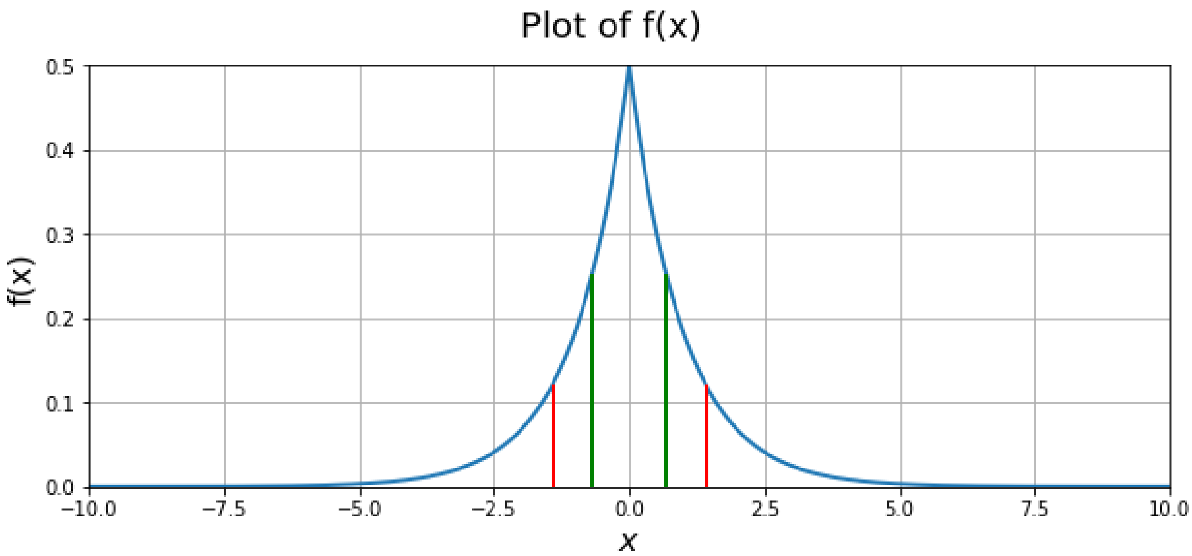


Figure 6: Plot of $f(x)$, with red bars to indicate the $\pm\sigma$ interval around the peak, and blue bars for the interval of the FWHM

2.2 Exercise 3.7

9 events are observed in an experiment modeled with a Poisson probability distribution.

(a) What is the 95% probability lower limit on the Poisson expectation value ? Take a flat prior for your calculations.

(b) What is the 68% confidence level interval for ν using the Neyman construction and the smallest interval definition?

A Poisson distribution is used to model events for which we do not have any information on the number of trials N, except that it is a very large number, and where the probability p of each trial yielding an event is fixed and small enough such that $\nu = Np$ is finite. It is described by the equation:

$$P(n|\nu) = \frac{e^{-\nu}\nu^n}{n!} \tag{35}$$

where n is the number of successful events out of the N trials, and $\nu = Np$ is the expectation value of n. Perform a Bayesian analysis on such a distribution results in the posterior $P(\nu|n)$. For the case of a flat prior $P_0(\nu) = C$, one needs to ensure that it is normalized. Assuming that ν can go up to a maximum value of ν_{max} and setting the condition:

$$\int_0^{\nu_{max}} P_0(\nu)d\nu = 1 \tag{36}$$

shows that $C = \dfrac{1}{\nu_{max}}$. Bayes theorem states that:

$$\begin{aligned} P(\nu|n) &= \frac{P(n|\nu)P_0(\nu)}{P(n)} \\ &= \frac{P(n|\nu)P_0(\nu)}{\displaystyle\int_0^{\nu_{max}} P(n|\nu)P_0(\nu)d\nu} \\ &= \frac{e^{-\nu}\nu^n}{\displaystyle\int_0^{\nu_{max}} e^{-\nu}\nu^n d\nu} \end{aligned} \tag{37}$$

For large ν_{max} such that $\nu_{max} >> n$, the integral in the denominator can be approximated to $n!$, yielding in the posterior result:

$$P(\nu|n) = \frac{e^{-\nu}\nu^n}{n!} \tag{38}$$

To find the 95% probability lower limit ν_{min}, one needs to select the integrating interval of the cumulative distributed of the posterior $P(\nu|n)$ and set the condition:

$$\int_0^{\nu_{min}} \frac{e^{-\nu}\nu^9}{9!} d\nu = 1 - 0.95 \tag{39}$$

which results in a 10^{th} degree polynomial, and therefore a numerical solution is preferred. The lower limit was found to be $\nu_{min} \approx 5.425$. It has been indicated with red bars in fig. 7.

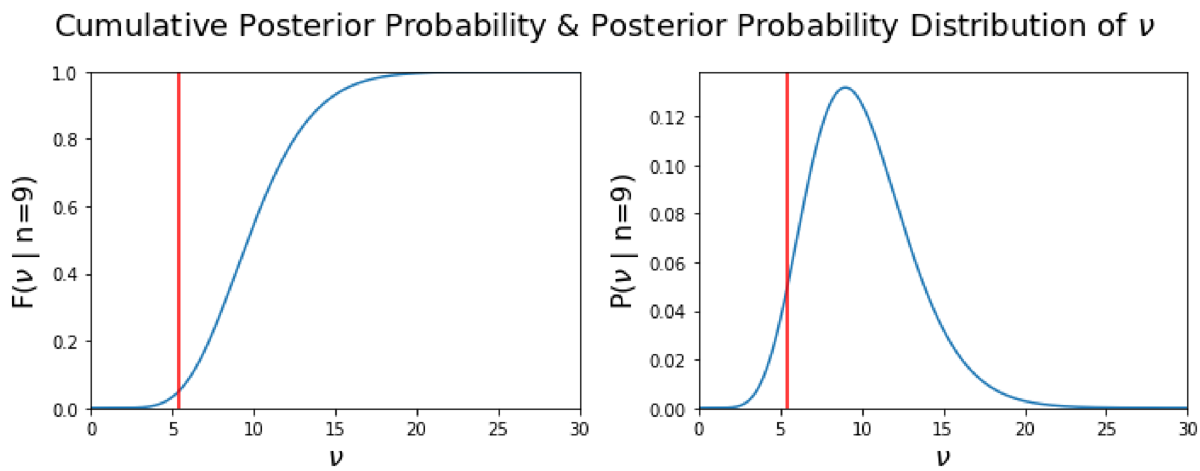


Figure 7: Plot of $P(\nu|n)$ (left) and its cumulative function (right). A red bar was used to indicate the 95% probability lower limit $\nu_{min} = 5.425$

b) To perform a Frequentist analysis, all values for ν were considered from 0 to 50 in steps of 0.005. For each ν, $P(n|\nu)$ was computed for all values of n from 0 to 100, and O^s_{68} the smallest 68% interval $[n_1, n_2]$ was determined as outlined in the procedure below:

1. Starting with $O^s_{68} = n^*$, if $P(n = n^*|\nu)$ at the mode n^* is larger than 0.68, then $O^s_{68} = n^*$ has a single value.

2. Else, add the next largest $P(n_i|\nu)$, and update O^s_{68} by including n_i in the set. Repeat this step until the sum of all probabilities for all the n values in our O^s_{68} is ≥ 0.68. The smallest interval is defined by $[n_1, n_2]$, where n_1 and n_2 are the smallest and the largest, respectively, members of our set of n_i.

The values of n_1 and n_2 are plotted for each considered ν in fig. 8.

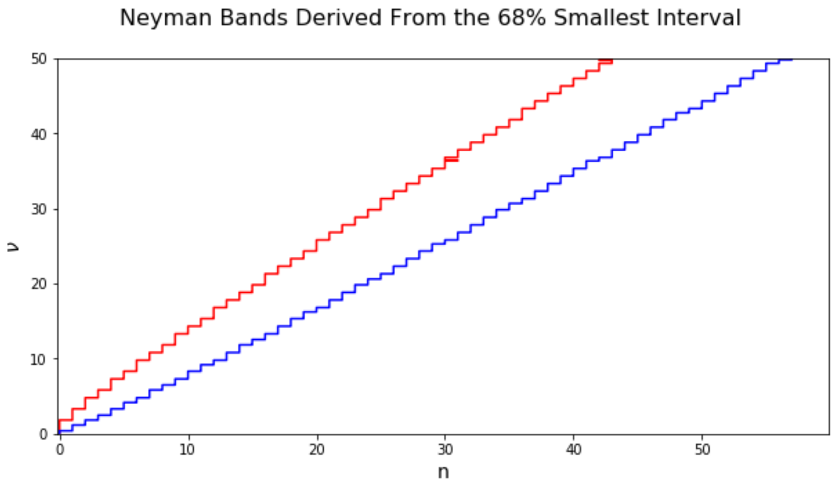


Figure 8: Neyman Bands plot, constructed using the smallest 68% *interval* $[n_1, n_2]$. *Red line indicates* n_2 *values and blue line* n_1 *values*

The 68% confidence level for $n = 9$ is given by the interval $[\nu_1, \nu_2]$, where ν_1, ν_2 are the smallest, respectively largest values for ν, for which $n = 9 \ \in O^s_{68}$. This interval is $[6.495, 13.300]$.

2.3 Exercise 3.8

Repeat the previous exercise, assuming you had a known background of $\lambda = 3.2$ events.

(a) Find the Feldman-Cousins 68% Confidence Level interval.

(b) Find the Neyman 68% Confidence Level interval.

(c) Find the 68% Credible interval for ν.

A Poisson process with a background yields another Poisson process, with expectation value $\mu = \nu + \lambda$, where ν is the expectation value of the signal and λ, the expectation value of the background. Therefore, frequentist analysis for the case of a Poisson distribution with a known background is treated the same as in the previous exercise, with the exception that the variable over which one iterates is $\mu = \nu + \lambda$. Information on ν is then extracted by $\nu = \mu - \lambda$. In our case, we have 9 events recorded, and $\lambda = 3.2$.

a) The Feldman-Cousins interval is constructed in a similar manner to the smallest interval, for which the procedure was outlined above. The main difference is that the probabilities $P(n_i|\mu = \nu + \lambda)$ are not added in descending order of magnitude, but in descending order of a ranking based on r:

$$r_i = \frac{P(n_i|\mu = \nu + \lambda)}{P(n_i|\hat{\mu})} \tag{40}$$

where $\hat{\mu}$ is the mode of the likelihood. Again, all values for μ were considered from 0 to 50 in steps of 0.005. For each μ, $P(n|\mu)$ was computed for all values of n from 0 to 100, and the 68% interval $[n_1, n_2]_{F-C}$ was determined. The Feldman-Cousins band plots are presented in fig. 9.

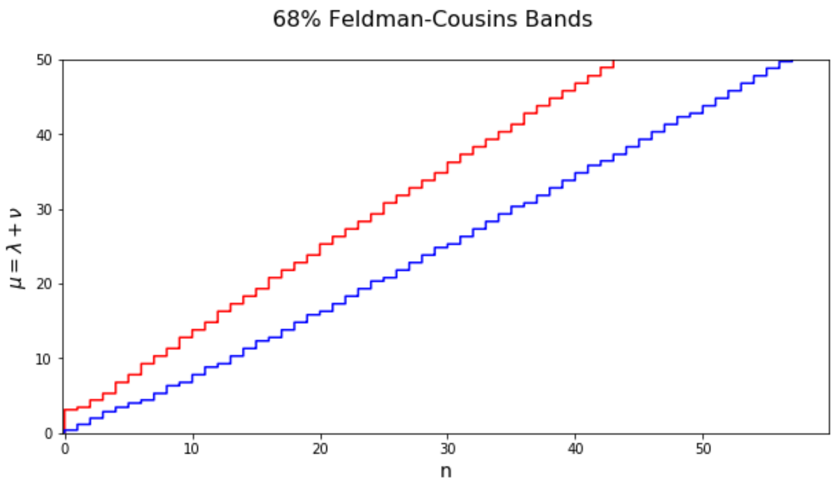


Figure 9: Feldman-Cousins Bands plot constructed using the 68% interval $[n_1, n_2]$. Red line indicates n_2 values and blue line n_1 values.

Computing the μ confidence level for $n_{data} = 9$ results in $[6.335, 12.790]$. The confidence level for ν is determined by subtracting $\lambda = 3.2$: $[3.135, 9.190]$.

b) For the Neyman 68% Construction, both the smallest interval and the central interval were computed, for comparison purpose. The 68% smallest interval Neyman bands are presented in fig. 10, showing a confidence interval for μ of $[6.495, 13.300]$ and for ν of $[3.295, 10.100]$.

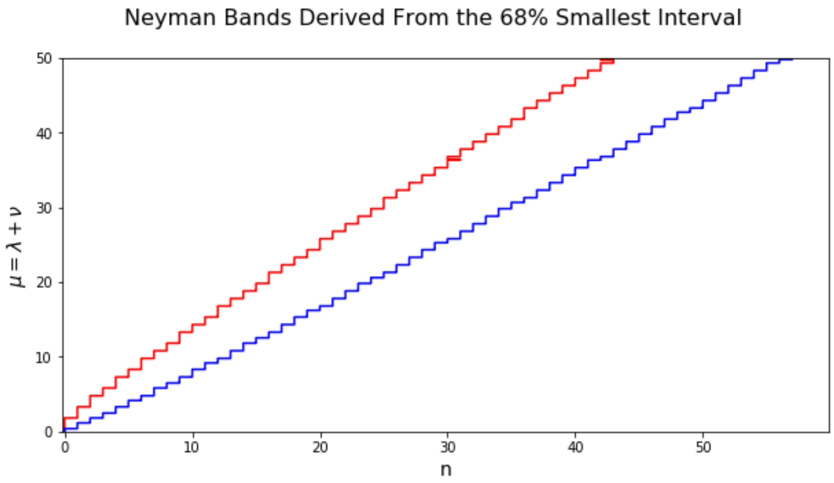


Figure 10: Neyman Bands plot constructed using the 68% smallest interval $[n_1, n_2]$. Red line indicates n_2 values and blue line n_1 values.

The 68% central interval Neyman bands are presented in fig. 11, showing a confidence interval for μ of $[6.070, 14.245]$ and for ν of $[2.870, 11.045]$.

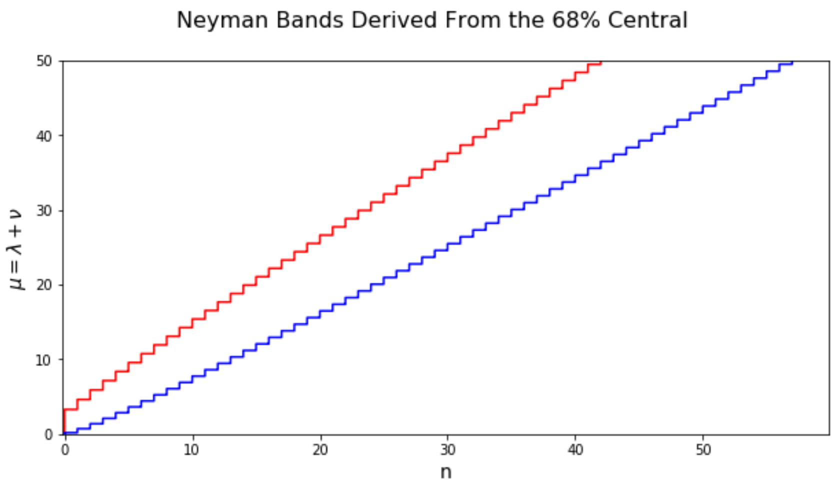


Figure 11: Neyman Bands plot constructed using the 68% central interval $[n_1, n_2]$. Red line indicates n_2 values and blue line n_1 values.

The Neyman bands and the Feldman-Cousins bands have been plotted together in fig. 12 for comparison. It clearly shows, as expected, that the Neyman central interval bands are the broadest of all three. The confidence levels corresponding to each frequentist construction are presented in a plot of the likelihood in fig. 13.

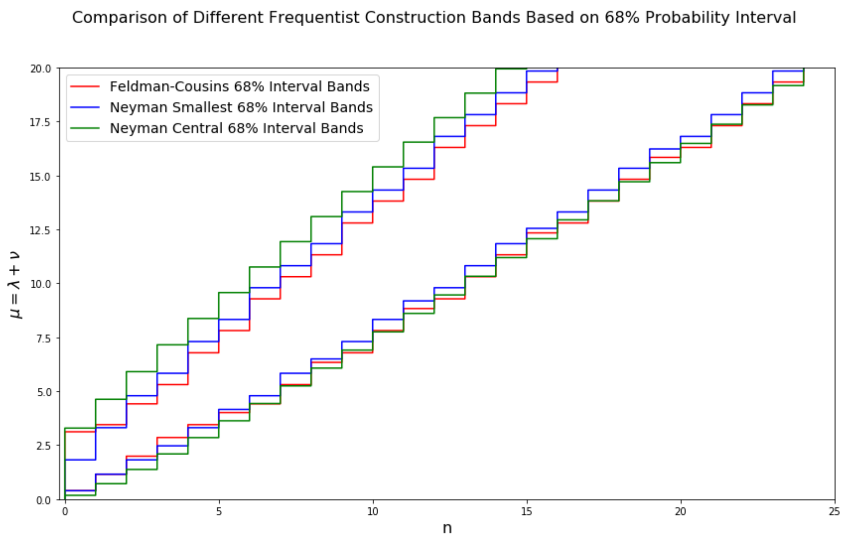


Figure 12: 68% Feldman-Cousins bands (red), 68% smallest interval Neyman bands (blue) and 68% central interval Neyman bands (green).

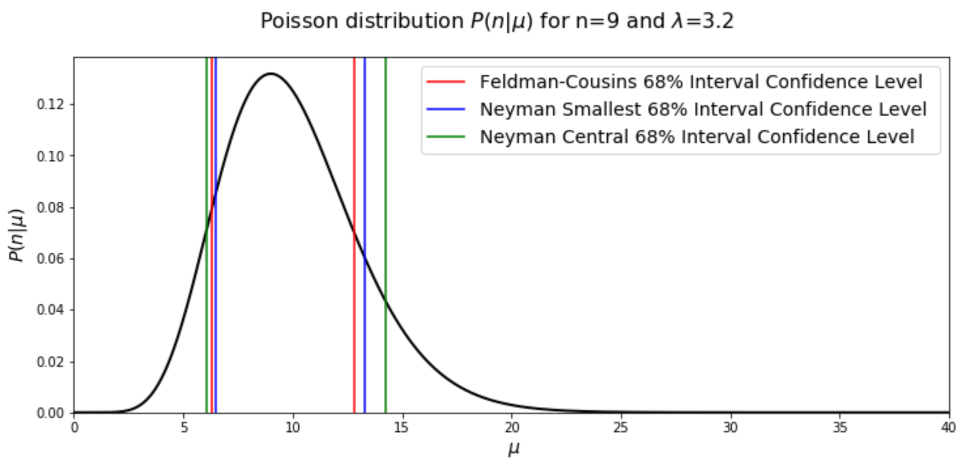


Figure 13: 68% Feldman-Cousins confidence level (red), 68% smallest interval Neyman confidence level (blue) and 68% central interval Neyman confidence level (green).

c) The credible intervals are based on the posterior Poisson distribution. For the posterior Poisson distribution with known background, there are 2 parameters to consider: n and λ. In this case, the Bayes theorem with a flat prior $P_0(\nu) = 1/\nu_{max}$ takes the form:

$$\begin{aligned} P(\nu|n,\lambda) &= \frac{P(\nu|n,\lambda)P_0(\nu)}{\displaystyle\int_0^{\nu_{max}} P(\nu|n,\lambda)P_0(\nu)d\nu} \\ &= \frac{e^{-(\nu+\lambda)}(\nu+\lambda)^n}{\displaystyle\int_0^{\nu_{max}} e^{-(\nu+\lambda)}(\nu+\lambda)^n d\nu} \end{aligned} \tag{41}$$

which can be solved by integration by parts, and approximating for $\nu \to \infty$, to reach the form:

$$P(\nu|n,\lambda) = \frac{e^{-\nu}(\nu+\lambda)^n}{n! \sum_{i=0}^{n} \frac{\lambda_i}{i!}} \tag{42}$$

Two credible intervals were calculated, based on the smallest 68% interval and on the central 68% interval. To this end, 500001 equally spaced apart values for $\nu \in [0, 50]$ were generated. The smallest 68% credible interval was found to be $[3.120, 9.147]$ and the central 68% credible interval $[3.717, 9.890]$. The difference of the sizes of these intervals is -0.146, showing that indeed the smallest interval is narrower than the central, as expected. The credible intervals are shown in fig. 14. All the results from a), b) and c) are summarized below, in table 5.

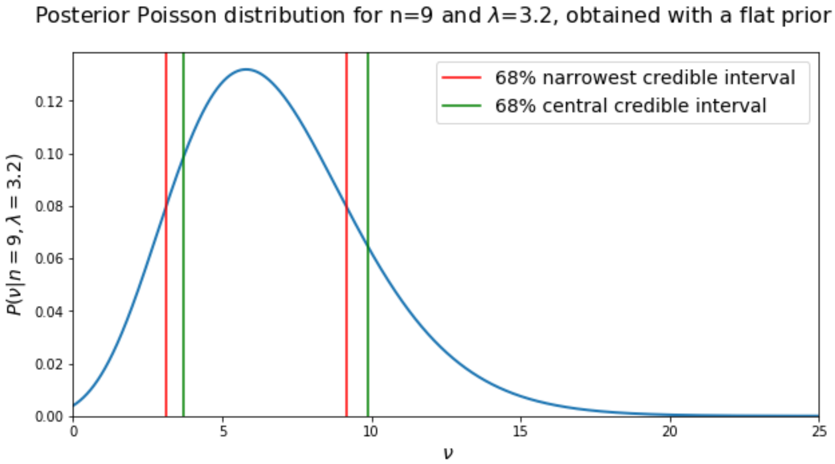


Figure 14: Posterior $P(\nu|n,\lambda)$, for $n = 9$ and $\lambda = 3.2$. Red bars indicate the 68% narrowest credible interval, and green bars the 68% central credible

Table 5: Confidence intervals and credible intervals for ν, for exercise 3.8

Type	*Interval*
68% Feld-Cousins Interval	[3.125, 9.590]
68% Neyman Smallest Interval	[3.295., 10.100]
68% Neyman Central Interval	[2.870, 11.045]
68% Smallest Credible Interval	[3.120, 9.147]
68% Central Credible Interval	[3.717, 9.890]

2.4 Exercise 3.13

In this problem, we look at the relationship between an unbinned likelihood and a binned Poisson probability. We start with a one dimensional density $f(x|\lambda)$ depending on a parameter λ and defined and normalized in a range $[a, b]$. n events are measured with x values x_i $i = 1, , n$. The unbinned likelihood is defined as the product of the densities

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

Now we consider that the interval $[a, b]$ is divided into K subintervals (bins). Take for the expectation in bin j

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

where the integral is over the x range in interval j, which is denoted as Δ_j. Define the probability of the data as the product of the Poisson probabilities in each bin.

We consider the limit $K \to \infty$ and, if no two measurements have exactly the same value of x, then each bin will have either $n_j = 0$ or $n_j = 1$ event. Show that this leads to

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

where Δ is the size of the interval in x assumed fixed for all j. I.e., the unbinned likelihood is proportional to the limit of the product of Poisson probabilities for an infinitely fine binning.

Starting from the LHS term, one can write:

$$\begin{aligned}
\prod_{j=1}^{K} \frac{e^{-\nu_j}\nu_j^{n_j}}{n_j!} &= \prod_{j=1}^{K} (e^{-\nu_j}\nu_j^{n_j}) \prod_{j=1}^{K} \frac{1}{n_j!} \\
&= \exp\left(-\sum_{j=1}^{K} \nu_j\right) \prod_{j=1}^{K} \nu_j^{n_j} \prod_{j=1}^{K} \frac{1}{n_j!} = \\
&= e^{-1} \prod_{j=1}^{K} \nu_j^{n_j} \prod_{j=1}^{K} \frac{1}{n_j!}
\end{aligned} \tag{43}$$

where the sum in the exponential was evaluated to be 1 due to the normalization condition on f: $\displaystyle\int_a^b f(x|\lambda)dx = 1$. One can rewrite this integral as a sum over all ν_j, leading to the result $\sum_{j=1}^{K} \nu_j = 1$.

Replacing $\nu_j = \int_{\Delta_j} f(x|\lambda)dx$ in eq. 43, yields

$$\prod_{j=1}^{K} \frac{e^{-\nu_j}\nu_j^{n_j}}{n_j!} = \frac{1}{e}\prod_{j=1}^{K}\left(\int_{\Delta_j} f(x|\lambda)dx\right)^{n_j}\prod_{j=1}^{K}\frac{1}{n_j!} \tag{44}$$

Taking $K \to \infty$, the last factor in the above expression becomes $\prod_{j=1}^{K}\frac{1}{n_j!} \to 1$

$$\begin{aligned}\lim_{K\to\infty}\prod_{j=1}^{K} \frac{e^{-\nu_j}\nu_j^{n_j}}{n_j!} &= \lim_{K\to\infty}\frac{1}{e}\prod_{j=1}^{K}\left(\int_{\Delta_j} f(x|\lambda)dx\right)^{n_j} \\ &= \lim_{K\to\infty}\frac{1}{e}\int_{\Delta_1}\int_{\Delta_2}..\int_{\Delta_K} g(x_1, x_2...x_{n_j}|\ \lambda)\ dx_{n_j}..dx_1\end{aligned} \tag{45}$$

where the function $g(g(x_1, x_2...x_{n_j}) = \prod_{i=1}^{n} f(x_i|\lambda) = \mathscr{L}(\lambda)$. For $K \to \infty$, the Δ_j become infinitely small, and the multiple integral can be written as a product

$$\begin{aligned}\lim_{K\to\infty}\prod_{j=1}^{K} \frac{e^{-\nu_j}\nu_j^{n_j}}{n_j!} &= \lim_{K\to\infty}\frac{1}{e}\int_{\Delta_1}\int_{\Delta_2}..\int_{\Delta_K} \mathscr{L}(\lambda)\ dx_{n_j}..dx_1 = \\ &= \frac{1}{e}\mathscr{L}(\lambda)\Delta \propto \mathscr{L}(\lambda)\end{aligned} \tag{46}$$

which is the proportionality we were looking to prove.

2.5 Exercise 3.16

We consider a thinned Poisson process. Here we have a random number of occurrences, N , distributed according to a Poisson distribution with mean ν. Each of the N occurrences, X_n , can take on values of 1, with probability p, or 0, with probability $(1-p)$. We want to derive the probability distribution for

$$X = \sum_{n=1}^{N} X_n$$

Show that the probability distribution is given by:

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

We will prove the distribution by finding the generating function of X. For the discrete random variable X:

$$X = \sum_{n=1}^{N} X_n \tag{47}$$

the expectation value $E[X]$ can be determined using Wald's identity which states that $E[X_1 + X_2 + .. + X_N] = E[N]E[X_1]$. Therefore

$$E[X] = E[X_i]E[N] = p\nu \tag{48}$$

where p is the probability that an event yields 1, and ν is the expectation value of the Poisson process generating the events. The generating function for X as a discrete random variable is:

$$\begin{aligned} G_X(z) = E[z^X] = \sum_X = 0^\infty P(X) z^X \\ = \sum_{n=0}^{\infty} E[z^X|N=n]\ \frac{\nu^n}{n!} e^{-\nu} \end{aligned} \tag{49}$$

where $E[z^X|N=n] = E[z^{X_1}]^n = (1+p(z-1))^n$. Therefore the generating function $G_X(z)$ is:

$$\begin{aligned} G_X(z) = (1+p(z-1))^n \frac{\nu^n}{n!} e^{-\nu} = \\ = \frac{[\nu(1+p(z-1))]^n}{n!} e^{-\nu} \end{aligned} \tag{50}$$

Using the approximation $\dfrac{[\nu(1+p(z-1))]^n}{n!} \to e^{\nu(1+p(z-1))}$, the generating function takes the form:

$$G_X(z) \approx e^{p\nu(s-1)} \tag{51}$$

which is the generating function of a Poisson distribution with mean $p\nu$. Hence:

$$P(X) = \frac{e^{-\nu p}(\nu p)^X}{X!} \tag{52}$$

3 Chapter 4 Exercises

3.1 Exercise 4.8

In this problem, you try out the Central Limit Theorem for a case where the conditions under which it was derived apply, and a case under which the conditions do not apply.

(a) In this exercise, try out the CLT on the exponential distribution. First, derive what parameters of a Gauss distribution you would expect from the mean of n samples taken from the exponential distribution with

$$p(x) = \lambda e^{-\lambda x}$$

Then, try out the CLT for at least 3 different choices of n and λ and discuss the results. To generate random numbers according to the exponential distribution, you can use

$$x = -\frac{\ln U}{\lambda}$$

where U is a uniformly distributed random number between $[0, 1)$.

(b) Now try out the CLT for the Cauchy distribution:

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

Argue why the CLT is not expected to hold for the Cauchy distribution. You can generate random numbers from the Cauchy distribution by setting

$$x = \gamma \tan(\pi U - \pi/2) + x_0$$

Try $x_0 = 25$ and $\gamma = 3$ and plot the distribution for x. Now take $n = 100$ samples and plot the distribution of the mean. Discuss the results.

a) The exponential distribution $p(x) = \lambda e^{-\lambda x}$ is defined only for $x \geq 0$ and has mean $1/\lambda$ and variance $\sigma^2 = 1/\lambda^2$. To derive the parameters of the Gauss distribution we expect from the mean of n samples taken from the exponential distribution, we start with the characteristic function of the distribution.

$$
\begin{aligned}
\phi_x(k) &= \int_0^\infty e^{ikx}\ \lambda e^{-\lambda x} dx \\
&= \lambda \int_0^\infty e^{(ik-\lambda)x} dx \\
&= \frac{\lambda}{ik-\lambda} e^{(ik-\lambda)x}\Big|_0^\infty \\
&= \frac{\lambda}{\lambda - ik}
\end{aligned}
\tag{53}
$$

Consider n samples X_j from the exponential distribution, with X their sum

$$
X = \sum_{j=1}^{n} X_j \tag{54}
$$

Each X_j is distributed exponentially with mean $1/\lambda$ and variance $1/\lambda^2$, therefore, by Wald's identity, the mean of X, $E[X] = \dfrac{n}{\lambda}$ and $Var[X] = \dfrac{n}{\lambda^2}$. Consider the random variable, with mean 0 and unit standard deviation (and therefore unit variance):

$$
\begin{aligned}
Z_n &= \frac{X - \frac{n}{\lambda}}{\frac{\sqrt{n}}{\lambda}} \\
&= \sum_{i=1}^{n} \frac{X_i - \frac{n}{\lambda}}{\frac{\sqrt{n}}{\lambda}} \\
&= \sum_{i=1}^{n} \left(\frac{1}{\sqrt{n}} \frac{X_i - \frac{1}{\lambda}}{\frac{\sqrt{1}}{\lambda}} \right)
\end{aligned}
\tag{55}
$$

Using the random variable $Y_i = \dfrac{X_i - \frac{1}{\lambda}}{\frac{\sqrt{1}}{\lambda}}$, the characteristic function of Z_n is:

$$
\begin{aligned}
\phi_{Z_n}(k) &= \phi_{Y_1}\left(\frac{k}{\sqrt{n}}\right)\ \phi_{Y_2}\left(\frac{k}{\sqrt{n}}\right)\ ...\ \phi_{Y_n}\left(\frac{k}{\sqrt{n}}\right) \\
&= \left[\phi_{Y_1}\left(\frac{k}{\sqrt{n}}\right)\right]^n
\end{aligned}
\tag{56}
$$

The characteristic function $\phi_{Y_1}(k)$ is complex valued, and therefore, one can expand its expression using Taylor's theorem for complex functions:

$$\phi_{Y_1}(k) = 1 - \frac{k^2}{2n} + \mathscr{O}\left(\frac{k^3}{n^{3/2}}\right) \quad , \qquad \frac{k}{\sqrt{n}} \to 0 \tag{57}$$

By discarding higher order terms, which vanish as $n \to \infty$, and using the approximation $\left(1 + \frac{x}{n}\right)^n \approx e^x$, one finds the characteristic function for Z_n:

$$\phi_{Z_n}(k) \approx e^{-\frac{1}{2}k^2} \tag{58}$$

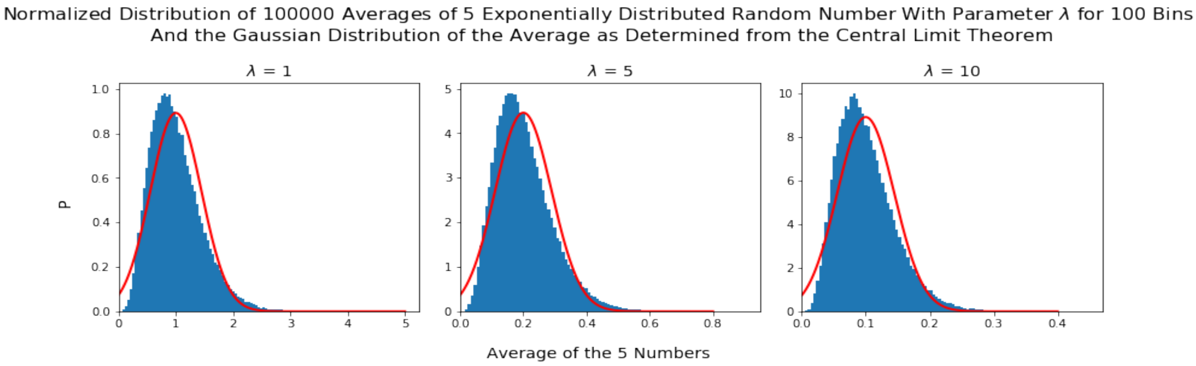
We recognize the characteristic function of a normal distribution with mean $\mu = 0$ and $\sigma^2 = 1$. Therefore, for very large n ($n \to \infty$), $P(Z_n) \to \mathscr{G}(\mu = 0, \sigma = 1)$.

Since Z_n depends on X, in such a case the distribution of X, $P(X) \to \mathscr{G}(\mu = \frac{n}{\lambda}, \sigma = \frac{\sqrt{n}}{\lambda})$. The average of the sum in X as well approaches a Gaussian $\frac{X}{n} \to \mathscr{G}(\mu = \frac{1}{\lambda}, \sigma = \frac{1}{\lambda\sqrt{n}})$. For a random variable $A_n = X/n$, its distribution can be approximated to

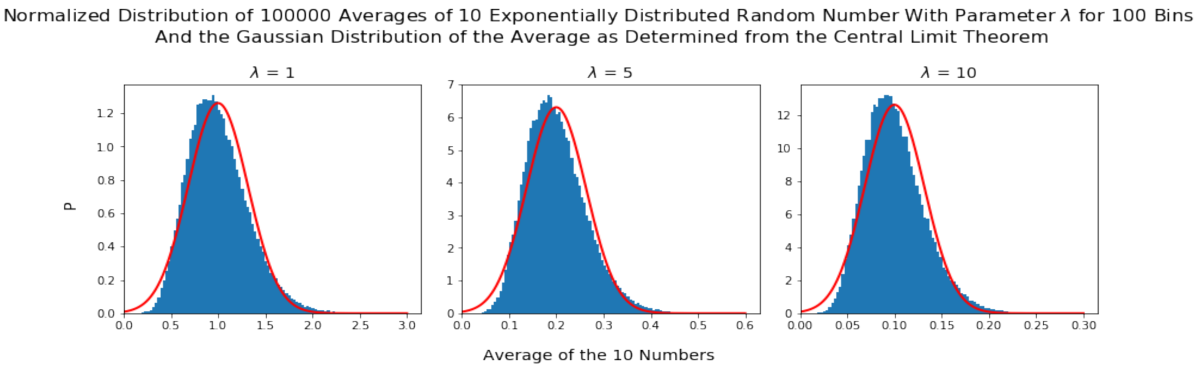
$$\begin{aligned} P(A_n) &= \mathscr{G}(A_n|\mu = \frac{1}{\lambda}, \sigma = \frac{1}{\lambda\sqrt{n}}) \\ &= \frac{\lambda\sqrt{n}}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{A_n - \frac{1}{\lambda}}{\frac{1}{\lambda\sqrt{n}}}\right)^2\right] \end{aligned} \tag{59}$$

To ilustrate the degree to which this Gaussian successfully approximates the distribution of the average, several values of n have been chosen: $n \in 5, 10, 50, 100$. For each value of n, 10^6 averages of numbers generated from the exponential distribution were simulated, and their histograms plotted together with the Gaussian meant to approximate the distribution. To show that the approximation depends solely on n and not on λ, 3 values of $\lambda \in 1, 5, 10$ were taken into account for each value of n. The plots are presented below in fig. 15e. One can observe that the best fit occurs for the highest value $n = 100$, while varying λ changes the width (through σ) and the mode (through μ), but has no impact on the fit.

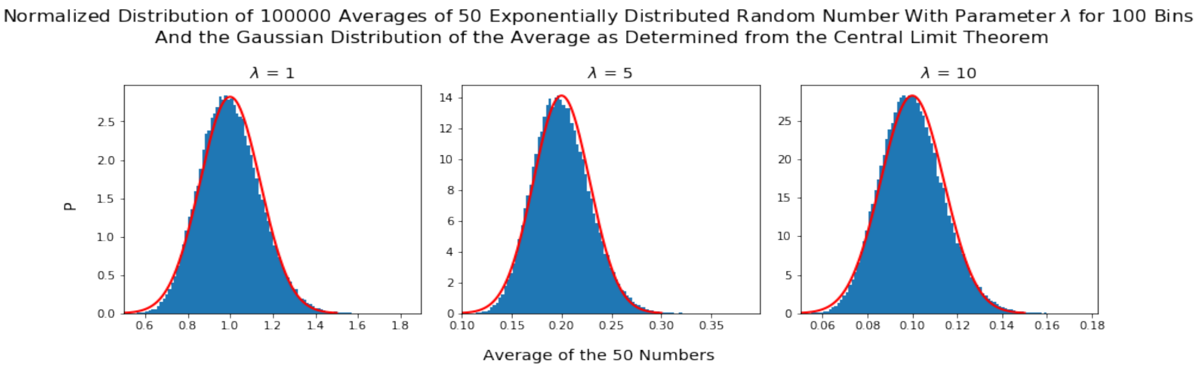
(a) $n = 5$



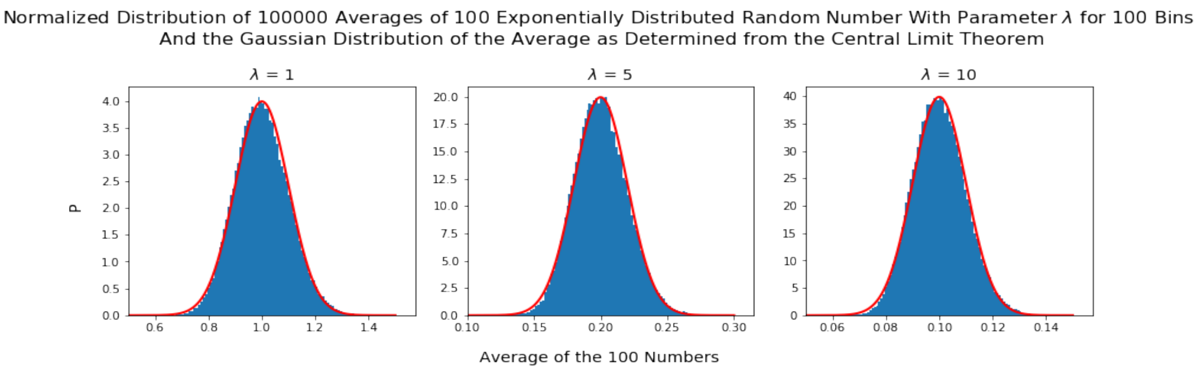
(b) $n = 10$



(c) $n = 50$



(d) $n = 100$



(e) Plots of the normalized histograms of 10^6 averages (blue) for n numbers from the exponential distribution, together with the Gaussians used to approximate them (red).

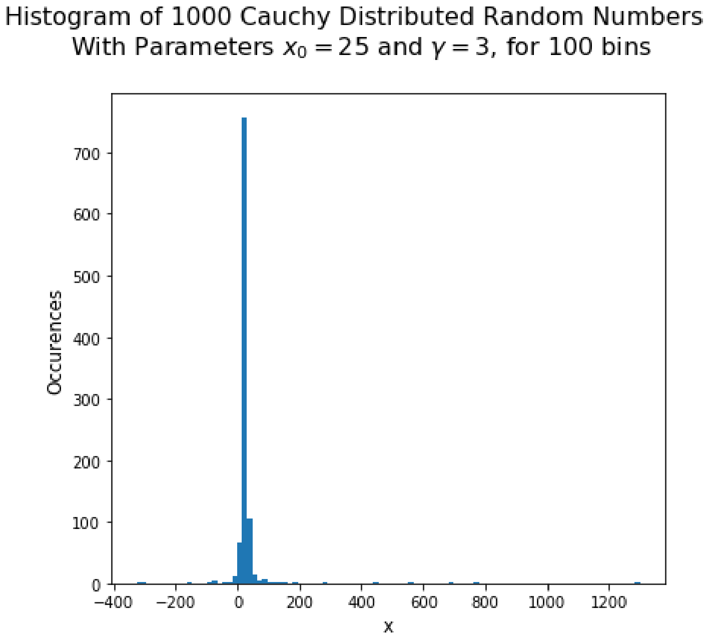


Figure 16: Histogram of 1000 numbers generated from the Cauchy distribution, given the parameters $x_0 = 25$ and $\gamma = 3$. The mean is $\mu = 28.917$ and the standard deviation is $\sigma = 61.22$.

b) Let X_j be a Cauchy distributed random variable. The proof of CLT starts from the premise that the expectation value is defined. However, the Cauchy distribution has no expectation value due to the fact that the integral below does not converge:

$$E[X_j] = \int_{-\infty}^{\infty} x \, \frac{1}{\pi\gamma} \, \frac{\gamma^2}{(x - x_0)^2 + \gamma^2} \, dx \tag{60}$$

In fact, none of the higher order moments of the Cauchy distribution are defined, therefore the distribution has no analytical expression for the variance either. To ilustrate the distribution, 1000 Cauchy distributed variables have been generated, given the parameters $x_0 = 25$ and $\gamma = 3$, and plotted in a histogram in fig. 16. Without a mean and a variance, the CLT cannot be applied in the case of a Cauchy distribution.

A number of 10^5 averages of $n = 100$ (fig. 17) and $n = 1000$ (fig. 18) Cauchy distributed variables have been generated, and plotted in a histogram, as scatter plots and together with a Gaussian with the same mean and variance as the 10000 averages. The plots presented below show that even at large $n = 1000$ (fig. 18), the Gaussian fails to approximate the Cauchy distribution properly. One notices that in both cases, the large standard deviation of the Cauchy averages spreads the Gaussian too much to be a good approximation. The scatter plots give some idea of how spreaded are the averages.

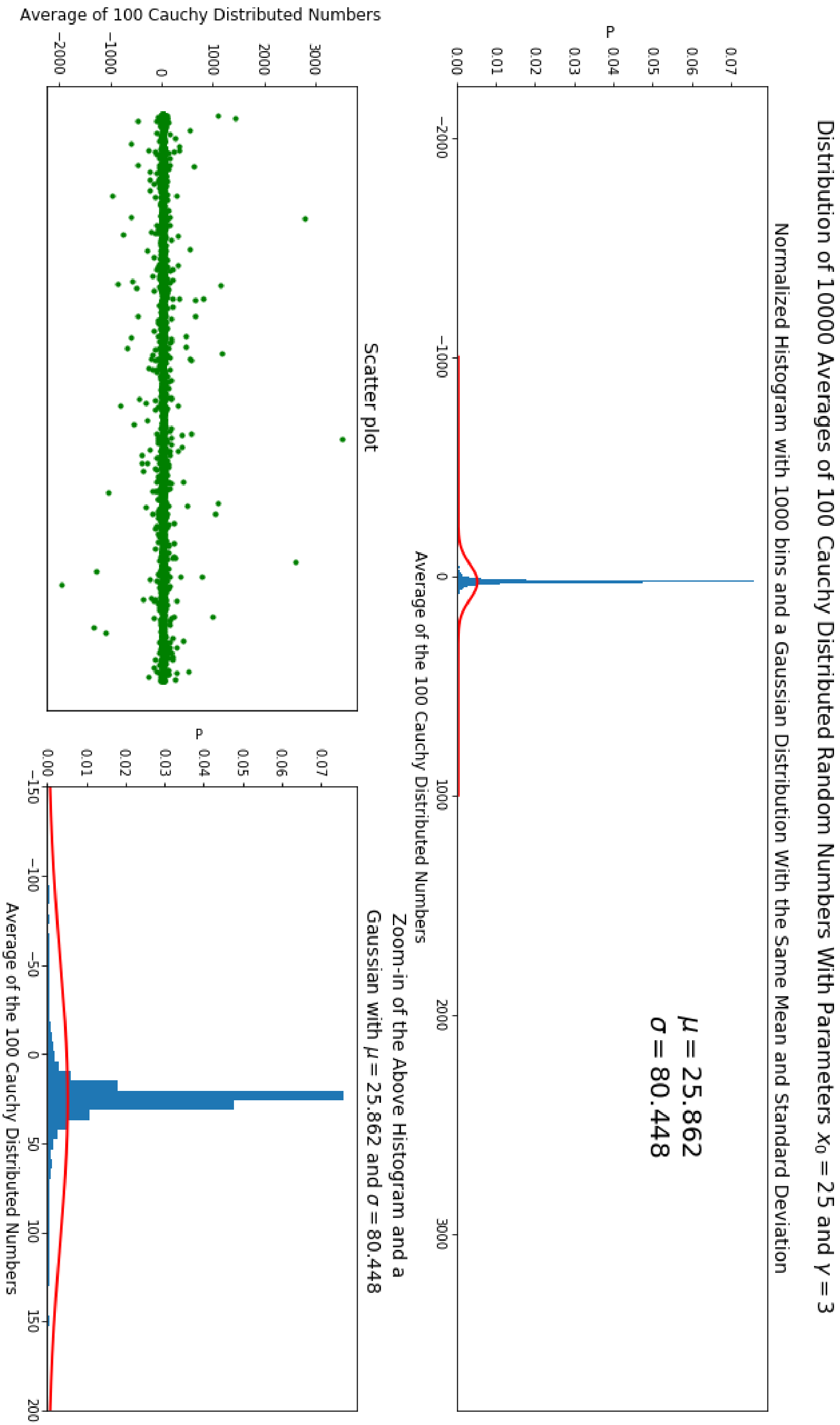


Figure 17: Histogram of 10000 averages of $n = 100$ *numbers generated from the Cauchy distribution, given the parameters* $x_0 = 25$ *and* $\gamma = 3$*. The mean is* $\mu = 25.862$ *and the standard deviation is* $\sigma = 80.448$*. A Gaussian distribution with the same mean and variance was plotted in red to ilustrate the poorness of the approximation*

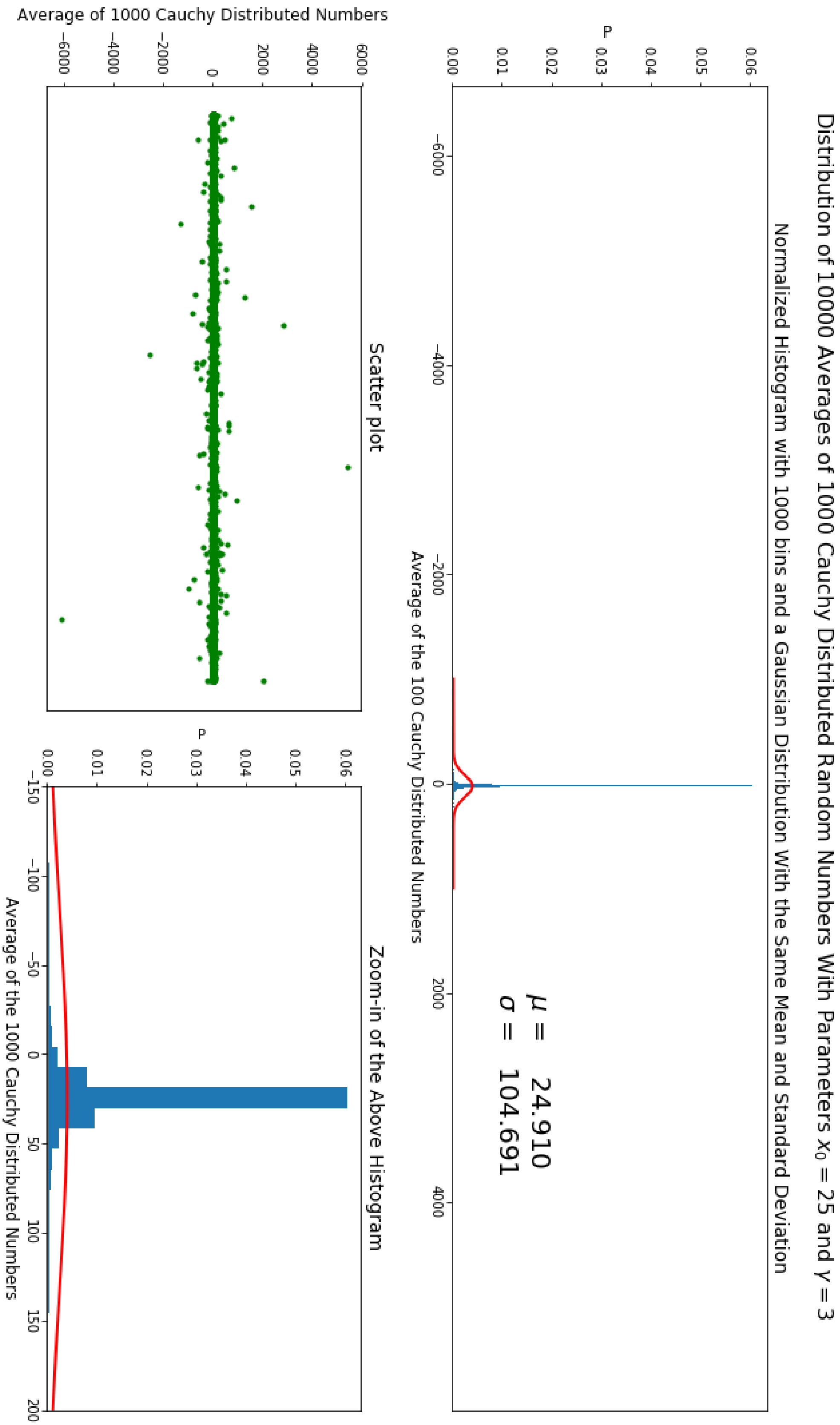


Figure 18: Histogram of 10000 averages of $n = 1000$ numbers generated from the Cauchy distribution, given the parameters $x_0 = 25$ and $\gamma = 3$. The mean is $\mu = 24.910$ and the standard deviation is $\sigma = 104.691$. A Gaussian distribution with the same mean and variance was plotted in red to ilustrate the poorness of the approximation

3.2 Exercise 4.11

With a plotting program, draw contours of the bivariate Gauss function (see next exercise for the definition of the function) for the following parameters:

(a) $\mu_x = 0$, $\mu_y = 0$, $\sigma_x = 1$, $\sigma_y = 1$, $\rho_{xy} = 0$

(b) $\mu_x = 1$, $\mu_y = 2$, $\sigma_x = 1$, $\sigma_y = 1$, $\rho_{xy} = 0.7$

(c) $\mu_x = 1$, $\mu_y = -2$, $\sigma_x = 1$, $\sigma_y = 2$, $\rho_{xy} = -0.7$

The correct bivariate Gaussian distribution is given by:

$$P(x,y) = \frac{\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right]}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \tag{61}$$

where ρ is the covariance coefficient of x and y.

a) For the evaluation of the bivariate Gaussian, 100 equally spaced values were generated for x and y in the interval $[-3, 3]$. The `matplotlib` package was used to plot the surface determined by the distribution, and its contour plots, using eq. 61 and parameters $\mu_x = 0$, $\mu_y = 0$, $\sigma_x = 1$, $\sigma_y = 1$, $\rho = 0$. The plot is presented in fig. 19. One may observe that for the projections in the $P(x,y) - x$ and $P(x,y) - y$ plane, the lines for negative values of y and x respectively overlap with the positive ones. This is to be expected, since $\rho = 0$ and x and y are independent. The projection in the $x - y$ plane is therefore fully symmetric.

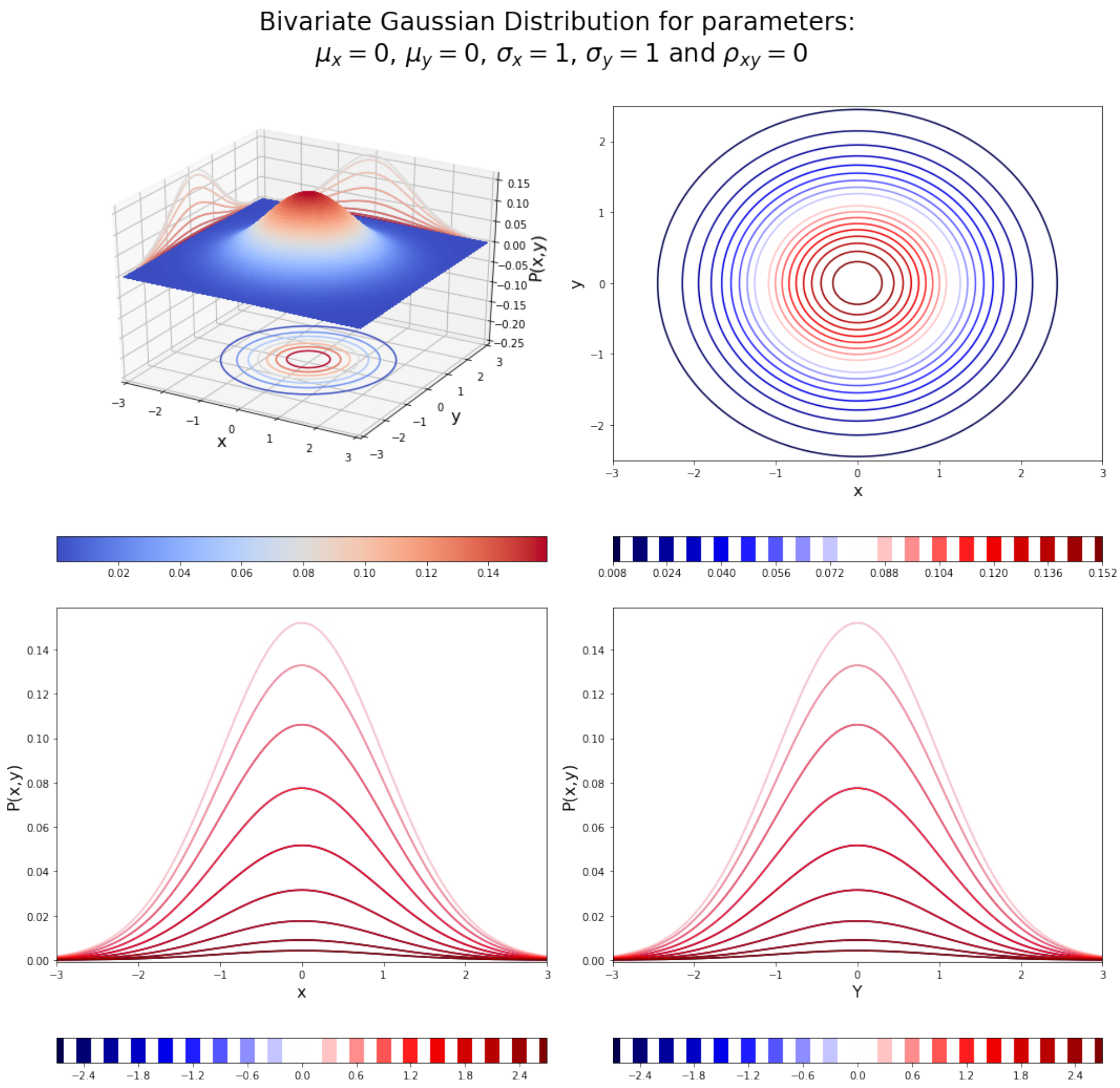


Figure 19: Bivariate Gaussian distribution and its contour plots for $\mu_x = 0$, $\mu_y = 0$, $\sigma_x = 1$, $\sigma_y = 1$, $\rho_{xy} = 0$.

b) For the evaluation of the bivariate Gaussian, 100 equally spaced values were generated for x and y in the interval $[-2, 4]$ and $[-1, 5]$, respectively. For parameters $\mu_x = 1$, $\mu_y = 2$, $\sigma_x = 1$, $\sigma_y = 1$, $\rho_{xy} = 0.7$, the plot is presented in fig. 20.

In this case, x and y are not independent anymore and for the projections in the $P(x, y) - x$ and $P(x, y) - y$ plane, the lines for negative values of y and x respectively do not overlap with the positive ones. The projection in the $x - y$ plane indicates a positive non-zero correlation between x and y.

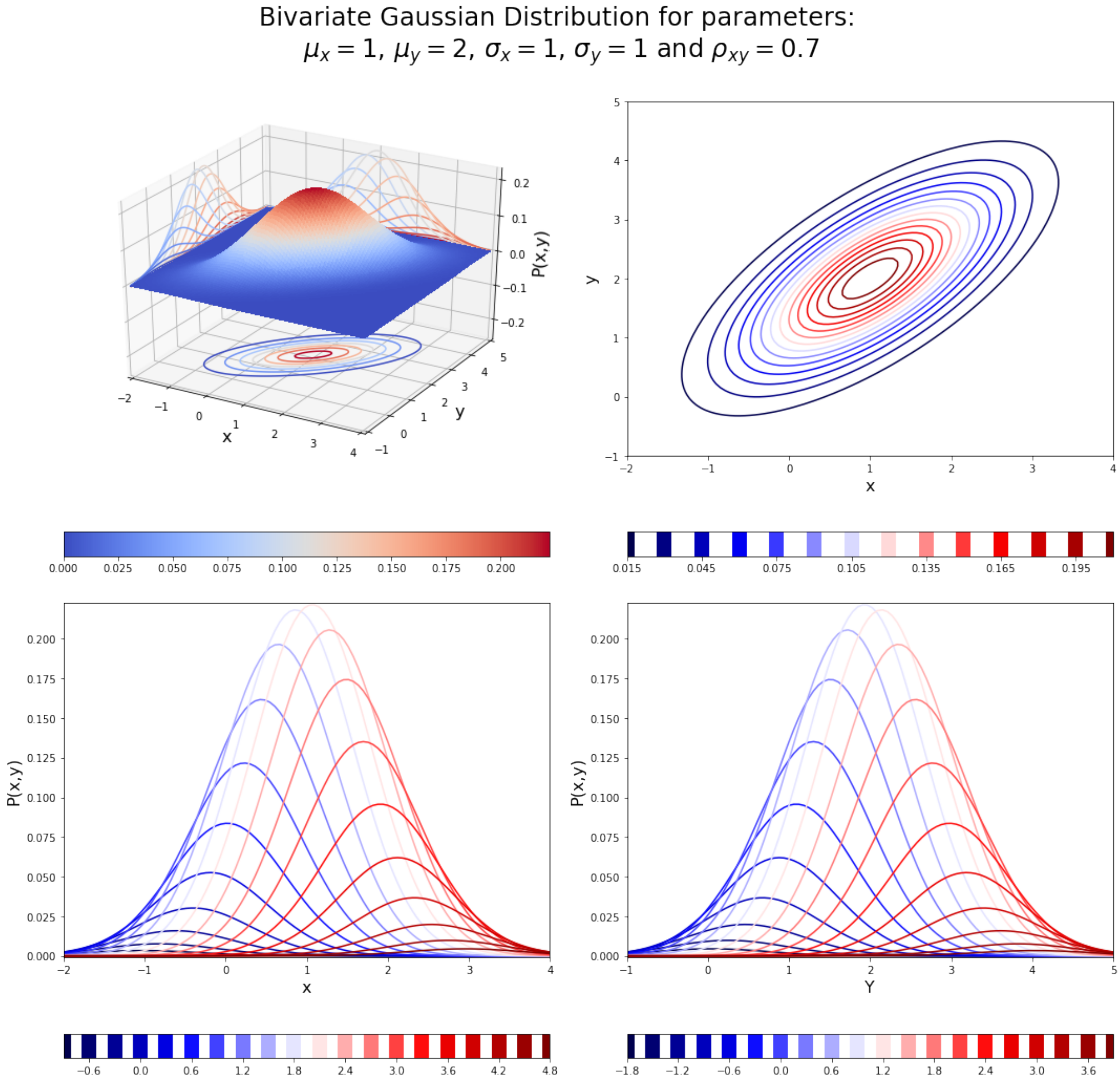


Figure 20: Bivariate Gaussian distribution and its contour plots for $\mu_x = 1$, $\mu_y = 2$, $\sigma_x = 1$, $\sigma_y = 1$, $\rho_{xy} = 0.7$.

c) For the evaluation of the bivariate Gaussian, 100 equally spaced values were generated for x and y in the interval $[-2, 4]$ and $[-7, 3]$, respectively. For parameters $\mu_x = 1$, $\mu_y = -2$, $\sigma_x = 1$, $\sigma_y = 2$, $\rho_{xy} = -0.7$, the plot is presented in fig. 21.

Again, x and y are not independent anymore, as indicated by $\rho = -0.7$. This is to be expected, since $\rho = 0$ and x and y are independent. The projection in the $x - y$ plane indicates a non-zero correlation between x and y. The higher standard deviation for y than in previous cases resulted in a broader distribution along this axis. The contour plot of the projection in the $x - y$ shows a negative non-zero correlation coefficient.

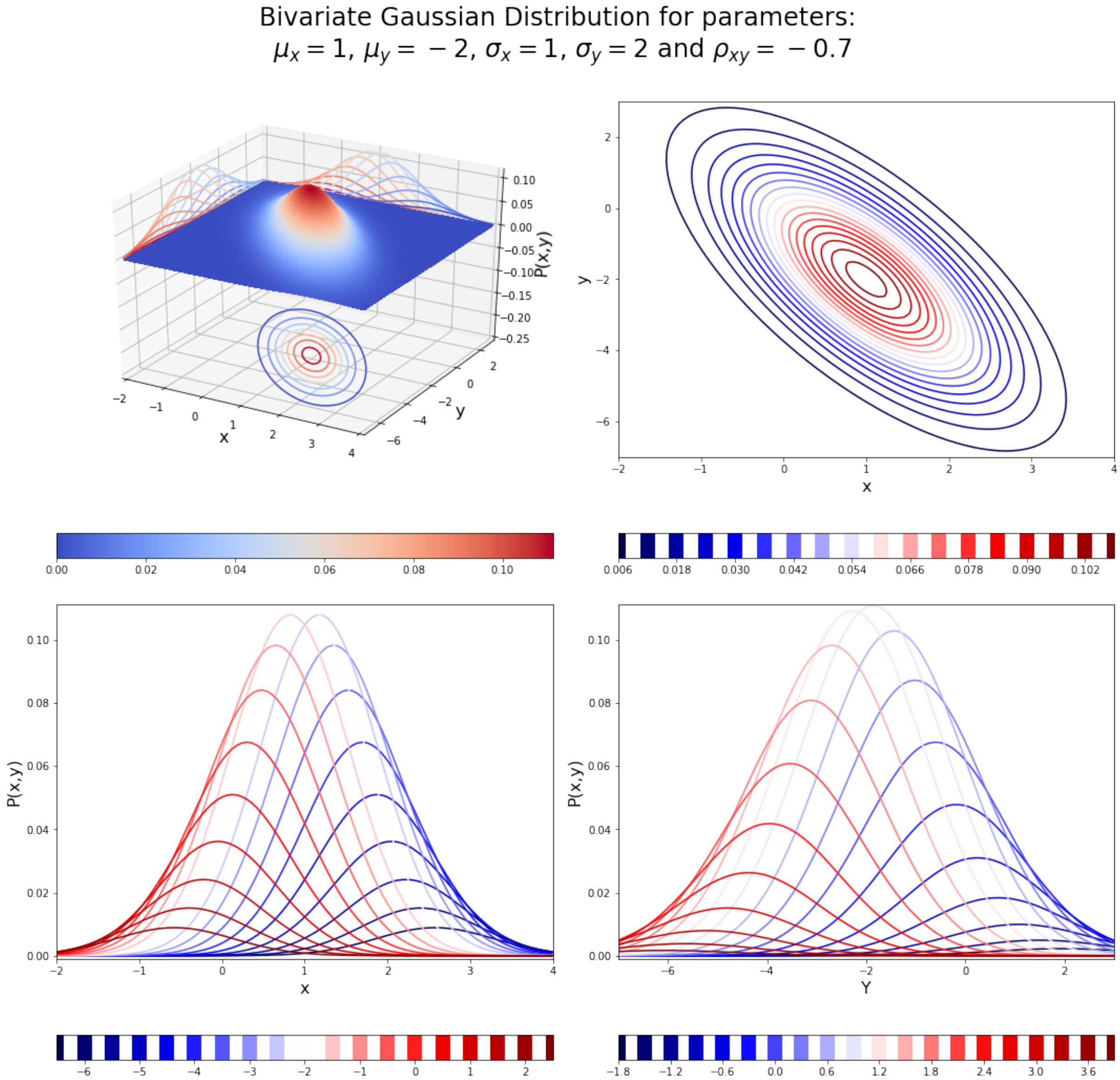


Figure 21: Bivariate Gaussian distribution and its contour plots for $\mu_x = 1$, $\mu_y = -2$, $\sigma_x = 1$, $\sigma_y = 2$, $\rho_{xy} = -0.7$.

3.3 Exercise 4.12

Bivariate Gauss probability distribution

(a) Show that the pdf can be written in the form (I changed the form to account for cases where the means of x and y are not 0)

$$P(x,y) = \frac{\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right]}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

(b) Show that for $z = x - y$, and x, y following the bivariate distribution, the resulting distribution for z is a Gaussian probability distribution with

$$\mu_z = \mu_x - \mu_y$$
$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$$

a) We will find the form of the bivariate Gaussian distribution starting from the joint distribution of a linear combination of independent random variables. Let X_1 and X_2 be normally and independently distributed variables with mean 0 and unit variance. Consider the following random variables:

$$\begin{aligned} Y_1 &= \mu_1 + \sigma_{11}X_1 + \sigma_{12}X_2 \\ Y_2 &= \mu_2 + \sigma_{21}X_1 + \sigma_{22}X_2 \end{aligned} \tag{62}$$

or written in matrix form:

$$\begin{pmatrix} Y_1 - \mu_1 \\ Y_2 - \mu_2 \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \tag{63}$$

where σ_{ij}, μ_i and σ_i are some coefficients. Since they are linear combinations of normally distributed variables, Y_1 and Y_2 are themselves normally distributed, with means μ_1 and μ_2 and variances:

$$\begin{aligned} \sigma_1^2 &= \sigma_{11}^2 + \sigma_{12}^2 \\ \sigma_2^2 &= \sigma_{21}^2 + \sigma_{22}^2 \end{aligned} \tag{64}$$

Let $V = \sigma_{11}\sigma_{21} + \sigma_{12}\sigma_{22}$ be the covariance $cov[Y_1, Y_2]$. The covariance matrix is defined as:

$$V_{ij} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{65}$$

with $\rho = \frac{V}{\sigma_1\sigma_2} = \frac{\sigma_{11}\sigma_{21}+\sigma_{12}\sigma_{22}}{\sigma_1\sigma_2}$ as the covariance coefficient. For variables X_1 and X_2, the joint probability distribution function is:

$$f(X_1, X_2)dX_1dX_2 = \frac{1}{2\pi}\exp\left[-\frac{X_1^2+X_2^2}{2}\right]dX_1dX_2 \tag{66}$$

For the matrix in eq. 63, we use the notation:

$$A = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix} \tag{67}$$

As long as $|A| \neq 0$, the matrix is invertible, and one can solve eq. 63 for X_1 and X_2

$$\begin{pmatrix}X_1\\ X_2\end{pmatrix} = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}^{-1}\begin{pmatrix}Y_1-\mu_1\\ Y_2-\mu_2\end{pmatrix} \tag{68}$$

$$\begin{pmatrix}X_1\\ X_2\end{pmatrix} = \frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}\sigma_{21}}\begin{pmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{12} & \sigma_{11}\end{pmatrix}\begin{pmatrix}Y_1-\mu_1\\ Y_2-\mu_2\end{pmatrix} \tag{69}$$

Using the last equation, the following expression is found for $X_1^2+X_2^2$:

$$X_1^2+X_2^2 = \frac{[\sigma_{22}(Y_1-\mu_1)-\sigma_{12}(Y_2-\mu_2)]^2}{(\sigma_{11}\sigma_{22}-\sigma_{12}\sigma_{21})^2} + \frac{[-\sigma_{21}(Y_1-\mu_1)+\sigma_{11}(Y_2-\mu_2)]^2}{(\sigma_{11}\sigma_{22}-\sigma_{12}\sigma_{21})^2} \tag{70}$$

$$(X_1^2+X_2^2)\,(\sigma_{11}\sigma_{22}-\sigma_{12}\sigma_{21})^2 = \sigma_1^2\sigma_2^2\,\mathscr{C} \tag{71}$$

where we have used $\mathscr{C}$ as a notation for the expression

$$\mathscr{C} = \frac{(Y_1-\mu_1)^2}{\sigma_1^2} + \frac{(Y_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(Y_1-\mu_1)(Y_2-\mu_2)}{\sigma_1\sigma_2} \tag{72}$$

Bringing the fractions in the above expression for $\mathscr{C}$ to a common denominator and noting that $\frac{\sigma_1^2\sigma_2^2}{(\sigma_{11}\sigma_{22}-\sigma_{12}\sigma_{21})^2} = \frac{1}{1-\rho^2}$, brings $X_1^2+X_2^2$ to the form

$$X_1^2+X_2^2 = -\frac{1}{2(1-\rho^2)}\left(\frac{(Y_1-\mu_1)^2}{\sigma_1^2} + \frac{(Y_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(Y_1-\mu_1)(Y_2-\mu_2)}{\sigma_1\sigma_2}\right) \tag{73}$$

We note the expression and move on to solving eq. 63 for X_1 and X_2. Observing that

$\dfrac{\sigma_1^2\sigma_2^2}{(\sigma_{11}\sigma_{22}-\sigma_{12}\sigma_{21})^2}=\dfrac{1}{1-\rho^2}$, leads to the solutions:

$$\begin{cases} X_1 = \dfrac{\sigma_{22}(Y_1-\mu_1)-\sigma_{12}(Y_2-\mu_2)}{\rho^*} \\ \\ X_2 = \dfrac{-\sigma_{21}(Y_1-\mu_1)-\sigma_{11}(Y_2-\mu_2)}{\rho^*} \end{cases} \tag{74}$$

where $\rho^* = \dfrac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{\sigma_{11}\sigma_{22}-\sigma_{12}\sigma_{21}}$. To get dX_1dX_2, one needs to calculate the Jacobian

$$J = \begin{vmatrix} \dfrac{\partial X_1}{\partial Y_1} & \dfrac{\partial X_1}{\partial Y_2} \\ \\ \dfrac{\partial X_2}{\partial Y_1} & \dfrac{\partial X_2}{\partial Y_2} \end{vmatrix} = \begin{vmatrix} \dfrac{\sigma_{22}}{\rho^*} & -\dfrac{\sigma_{12}}{\rho^*} \\ \\ -\dfrac{\sigma_{21}}{\rho^*} & \dfrac{\sigma_{22}}{\rho^*} \end{vmatrix} = \frac{1}{\rho^*}\left(\sigma_2 2\sigma_1 1 - \sigma_2 1\sigma_1 2\right) \tag{75}$$

Using $\sigma_1^2=\sigma_{11}^2+\sigma_{21}^2$, $\sigma_2^2=\sigma_{21}^2+\sigma_{22}^2$, and $\rho=\dfrac{V}{\sigma_1\sigma_2}=\dfrac{\sigma_{11}\sigma_{21}+\sigma_{12}\sigma_{22}}{\sigma_1\sigma_2}$, the final form for the Jacobian is:

$$J = \frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}} \tag{76}$$

leading to the result $dX_1dX_2 = \dfrac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}}\, dY_1dY_2$ which is used to find the solution $P(Y_1,Y_2)dY_1dY_2$:

$$P(Y_1,Y_2)dY_1dY_2 = \frac{1}{2\pi}\exp\left(-\frac{X_1^2+X_2^2}{2}\right)dX_1dX_2 = \tag{77}$$

$$= \frac{\exp\left[-\dfrac{1}{2(1-\rho^2)}\left(\dfrac{(Y_1-\mu_1)^2}{\sigma_1^2}+\dfrac{(Y_2-\mu_2)^2}{\sigma_2^2}-\dfrac{2\rho(Y_1-\mu_1)(Y_2-\mu_2)}{\sigma_1\sigma_2}\right)\right]}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, dY_1dY_2$$

which is the result we were seeking.

b) Consider x and y two random, normally distributed variables. Since they follow the bivariate distribution, they are jointly normal. In this case, any linear combination $ax+by$ is normally distributed. We consider the particular case of the linear combination: $z=x-y$. The variance of z is

$$
\begin{aligned}
Var[z] &= E[z^2] - (E[z])^2 \\
&= E[x^2 + y^2 - 2xy] - (E[x] - E[y])^2 \\
&= E[x^2] - (E[x])^2 + E[y^2] - (E[y])^2 - 2E[xy] - 2E[x]E[y]
\end{aligned}
\tag{78}
$$

Let $\sigma_x^2 = E[x^2] - (E[x])^2$ and $\sigma_y^2 = E[y^2] - (E[y])^2$. The covariance coefficient becomes:

$$
\begin{aligned}
\rho = \frac{E[xy] - E[x]E[y]}{\sigma_x \sigma_y} &\implies 2\rho\sigma_x\sigma_y = 2(E[xy] - E[x]E[y] \\
\implies \sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y
\end{aligned}
\tag{79}
$$

which is the result we were looking for.
For the mean μ_z, the proof requires the evaluation of the expression

$$
f(z) = \int_y \int_x \frac{\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right]}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \delta(z - (x+y))dxdy
\tag{80}
$$

for which finding an analytical solution by hand is tedious. When computed with a symbolic language such as Wolfram Mathematica, yields:

$$
f(z) = \frac{1}{\sqrt{2\pi(\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y)}} \exp\left(-\frac{(z - (\mu_x - \mu_y))^2}{2(\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y)}\right)
\tag{81}
$$

proving $\mu_z = \mu_x - \mu_y$, and confirming that $\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$.

3.4 Exercise 4.13

Convolution of Gaussians: Suppose you have a true distribution which follows a Gaussian distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-x_0)^2}{2\sigma_x^2}} \tag{82}$$

and the measured quantity, y follows a Gaussian distribution around the value x.

$$P(y|x) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-x)^2}{2\sigma_y^2}} \tag{83}$$

What is the predicted distribution for the observed quantity y ?

By law of total probability:

$$P(y) = \int_{-\infty}^{\infty} P(y|x) f(x)\ dx \tag{84}$$

Cosider the function $g(y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{y^2}{2\sigma_y^2}}$ such that $g(y-x) = P(y|x)$. The above expression becomes:

$$\begin{aligned} P(y) &= \int_{-\infty}^{\infty} g(y-x) f(x)\ dx \\ &= f * g(y) \end{aligned} \tag{85}$$

The Convolution Theorem states that $\mathscr{F}^{-1}[\mathscr{F}[f(y)]\mathscr{F}[g(y)]] = f * g(y)$, where $\mathscr{F}$ denotes the Fourier transform and $\mathscr{F}^{-1}$ the inverse Fourier. Therefore

$$\begin{aligned} P(y) &= \mathscr{F}^{-1}[\mathscr{F}[f(y)]\mathscr{F}[g(y)]] \\ &= \mathscr{F}^{-1}\left[\int_{-\infty}^{\infty} f(y) e^{-2\pi i k y} dy \int_{-\infty}^{\infty} g(y) e^{-2\pi i k y} dy\right] \end{aligned} \tag{86}$$

Starting with $\mathscr{F}[f(y)]$ and using the change of variable $y' = y - x_0$:

$$\begin{aligned} \mathscr{F}[f(y)] = \int_{-\infty}^{\infty} f(y) e^{-2\pi i k y}\ dy\ &= \frac{e^{-2\pi i k x_0}}{\sqrt{\pi}\sigma_x} \int_{-\infty}^{\infty} e^{-\frac{y'^2}{2\sigma_x^2}} e^{-2\pi i k y'}\ dy \\ &= \frac{e^{-2\pi i k x_0}}{\sqrt{\pi}\sigma_x} \int_{-\infty}^{\infty} e^{-\frac{y'^2}{2\sigma_x^2}}\ [\cos(2\pi k y') - i \sin(2\pi k y')]\ dy \end{aligned} \tag{87}$$

We note that sin is an odd function, and therefore its integral vanishes. The remaining terms take the form:

$$\begin{aligned}\mathscr{F}[f(y)] &= \frac{e^{-2\pi i k x_0}}{\sqrt{\pi}\sigma_x}\int_{-\infty}^{\infty} e^{-\frac{y'^2}{2\sigma_x^2}} \cos(2\pi k y')\, dy \\ &= 2\frac{e^{-2\pi i k x_0}}{\sqrt{\pi}\sigma_x}\int_{0}^{\infty} e^{-\frac{y'^2}{2\sigma_x^2}} \cos(2\pi k y')\, dy\end{aligned} \tag{88}$$

Where we changed the integrating limits to $[0, \infty)$ and added a factor of 2 to compensate for this. This change is possible due to the fact that the cos is an even function, and the even power of y' in the exponential makes the exponential even as well. Therefore we can split the integrating interval in half, and multiply the expression by a factor of 2. We further make use of the standard form of the integral

$$\int_0^{\infty} e^{at^2} \cos(2xt)dt = \frac{1}{2}\sqrt{\frac{\pi}{a}}\, e^{-\frac{x^2}{a}}$$

to evaluate $\mathscr{F}[f(y)]$ to

$$\mathscr{F}[f(y)] = e^{-2\pi i k x_0} e^{-2\pi^2\sigma_x^2 k^2} \tag{89}$$

Analogously, we find $\mathscr{F}[g(y)]$ to be:

$$\mathscr{F}[g(y)] = e^{-2\pi^2\sigma_y^2 k^2} \tag{90}$$

and $\mathscr{F}^{-1}[\mathscr{F}[f(y)]\mathscr{F}[g(y)]]$:

$$\begin{aligned}\mathscr{F}^{-1}[\mathscr{F}[f(y)]\mathscr{F}[g(y)]] &= \mathscr{F}^{-1}\left[e^{-2\pi i k x_0} e^{-2\pi^2(\sigma_x^2+\sigma_y^2)k^2}\right] \\ \implies P(y) &= \frac{1}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2)} e^{-\frac{(y-x_0)^2}{2(\sigma_x^2+\sigma_y^2)}}}\end{aligned} \tag{91}$$

In conclusion, the distribution for y is a Gaussian with mean x_0 and variance $\sigma = \sigma_x^2 + \sigma_y^2$

3.5 Exercise 4.14

Measurements of a cross section for nuclear reactions yields the following data.

Table 6: Data for exercise 4.14

θ	*Cross section*	*Errors*
30°	11	1.5
45°	13	1.0
90°	17	2.0
120°	17	2.0
150°	14	1.5

The units of cross section are 10^{30} cm^2/steradian. Assume the quoted errors correspond to one Gaussian standard deviation. The assumed model has the form

$$\sigma(\theta) = A + B\cos\theta + C\cos\theta^2$$

(a) Set up the equation for the posterior probability density assuming flat priors for the parameters A, B, C.

(b) What are the values of A, B, C at the mode of the posterior pdf ?

Let y_i denote the measured cross sections. We assume they are Gaussian distributed around the same mean μ, but each with variance σ_i^2. Let D be the set of all data. The probability $P(D|A,B,C)$ is:

$$\begin{aligned} P(D|A,B,C) &= \prod_{i=1}^{5} \mathscr{G}(y_i|\mu,\sigma(\theta_i)) \\ &= \prod_{i=1}^{5} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i-\mu)^2}{2\sigma(\theta_i)}\right] \\ &= \frac{1}{\prod_i[\sigma(\theta_i)]\ 4\pi\sqrt{2\pi}} \exp\left[-\sum_i\left(\frac{(y_i-\mu)^2}{2\sigma(\theta_i)^2}\right)\right] \end{aligned} \tag{92}$$

Bayes Theorem tells us that:

$$
\begin{aligned}
P(A,B,C|D) &= \frac{P(D|A,B,C)}{P(D)} \\
&= \frac{P(D|A,B,C)\ P_0(A,B,C)}{P(D)} \\
&= \frac{P(D|A,B,C)\ P_0(A,B,C)}{\int\int\int(D|A,B,C)P_0(A,B,C)\ dAdBdC}
\end{aligned} \tag{93}
$$

where $P_0(A,B,C)$ is the joint prior probability distribution of A, B and C, which depends on the covariance matrix of the parameters if they are not independent. Since one cannot calculate the covariance without any given relationship between the parameters, we assume A, B and C and independent. Therefore their joint probability $P_0(A,B,C) = P_0(A)P_0(B)P_0(C)$. The form of the posterior, given flat priors is:

$$
\begin{aligned}
P(A,B,C|D) &= \frac{P(D|A,B,C)}{\int_{C_{min}}^{C_{max}}\int_{B_{min}}^{B_{max}}\int_{A_{min}}^{A_{max}} P(D|A,B,C)\ dAdBdC} \\
&= \frac{\frac{1}{\prod_i \sigma(\theta)}\exp\left[-\sum_i \frac{(y_i-\mu)^2}{2\sigma(\theta_i)^2}\right]}{\int_{C_{min}}^{C_{max}}\int_{B_{min}}^{B_{max}}\int_{A_{min}}^{A_{max}} \frac{1}{\prod_i \sigma(\theta)}\exp\left[-\sum_i \frac{(y_i-\mu)^2}{2\sigma(\theta_i)^2}\right] dAdBdC}
\end{aligned} \tag{94}
$$

4 Chapter 5 Exercises

4.1 Exercise 5.1

Follow the steps in the script to fit a Sigmoid function to the following data:

Table 7: Data for exercise 5.1

Energy (E_i)	Trials (N_i)	Successes (r_1)
0.5	100	0
1.0	100	4
1.5	100	22
2.0	100	55
2.5	100	80
3.0	100	97
3.5	100	99
4.0	100	99

(a) Find the posterior probability distribution for the parameters (A, E_0)

(b) Define a suitable test statistic and find the frequentist 68% Confidence Level region for (A, E_0).

b) The sigmoid function to fit to p_i has the following expression:

$$p(E|A, E_0) = \frac{1}{1 + e^{-A(E-E_0)}} \tag{95}$$

The usage of a test statistic greatly reduces the computational costs of frequentist analysis. A test statistic is deemed a "suficient statistic" if complexity of data is reduced without loss of information. For the current analysis, we are dealing with a series of 8 data points $\{E_i^{data},\ N_i^{data},\ r_i^{data}\}$ generated by a binomial process. A useful test statistic for our case is:

$$\xi(\{r_i\}; A, E_0) = \prod_{i=1}^{k} \frac{N_i!}{r_i!(N_i - r_i)!}\, p_i(A, E_0)^{r_i}(1 - p(A, E_0))^{N_i - r_i} \tag{96}$$

The procedure employed in the current analysis is outlined below:

1. A grid of values (A, E_0) was generated in steps of 0.01 for $A \in [2.40, 3.40]$ and $E_0 \in [1.85, 2.10]$. Iteration through all points in the grid was performed.

2. At every point, 8 data points $\{N_i,\ r_i\}$ were generated, one for each E_i^{data}, using the sigmoid function $p(E|A, E_0)$ and the known binomial distribution $P(r_i|N_i, p(E|A, E_0))$.

3. The ξ test statistic described above was calculated for generated set of 8 data points.

4. Steps 2 and 3 were repeated 10000 times for every single point in the grid, and all values observed for ξ were sorted in descending order. The 2 values of ξ for which 68% and 30% of the set was above them were stored.

5. The test statistic ξ^{data} was computed for the observed experimental data set $\{E_i^{data}$, N_i^{data}, $r_i^{data}\}$ and if it was higher than the $\xi_{0.68}$ value determined above, the point (A, E_0) was indicated with a red marker. If $\xi^{data} \geq \xi_{0.30}$, the point was indicated with a green marker. If none of the conditions were fulfilled, the point was indicated with a black marker.

The plot below, in fig. 22, shows the resulting points for which $\xi^{data} \geq \xi_{0.30}$ with green, the points for which $\xi^{data} \geq \xi_{0.68}$ with red, and the rest with black.

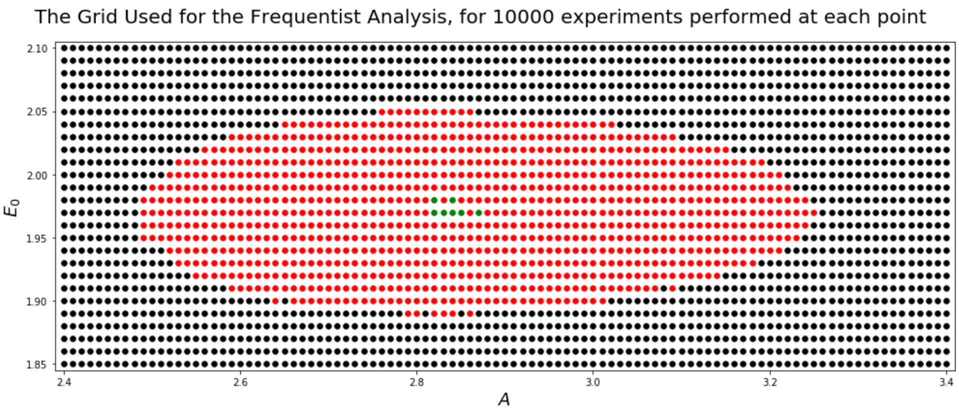


Figure 22: The grid of (A, E_0) generated for the frequentist analysis, with the red markers indicating the 68% confidence level interval for (A, E_0), and the green markers indicating the 30% confidence level points.

a) For the Bayesian Analysis, we start with observing the experimental data. We note that for $E = 2.0$, the value r/N seems to be close to 0.5. We therefore expect E_0 to take a value around 2.0, and use this for our prior. We select a Gaussian prior for E_0, with mean $\mu_{E_0} = 2.0$ and $\sigma_{E_0} = 0.3$. We further observe that as E increases from 2.0 to 2.5, the value of r/N increases by $\approx 35\%$. We deduce that $1/2 \times A/4 \approx 0.35$, and therefore expect A to take values around 2.8. We select a Gaussian as a prior for A with mean $\mu_A = 2.8$ and $\sigma_A = 0.5$.

$$\begin{cases} P_0(A) = \mathscr{G}(A|\mu_A = 2.8, \sigma_A = 0.5) \\ P_0(E_0) = \mathscr{G}(A|\mu_A = 2.0, \sigma_A = 0.3) \end{cases} \tag{97}$$

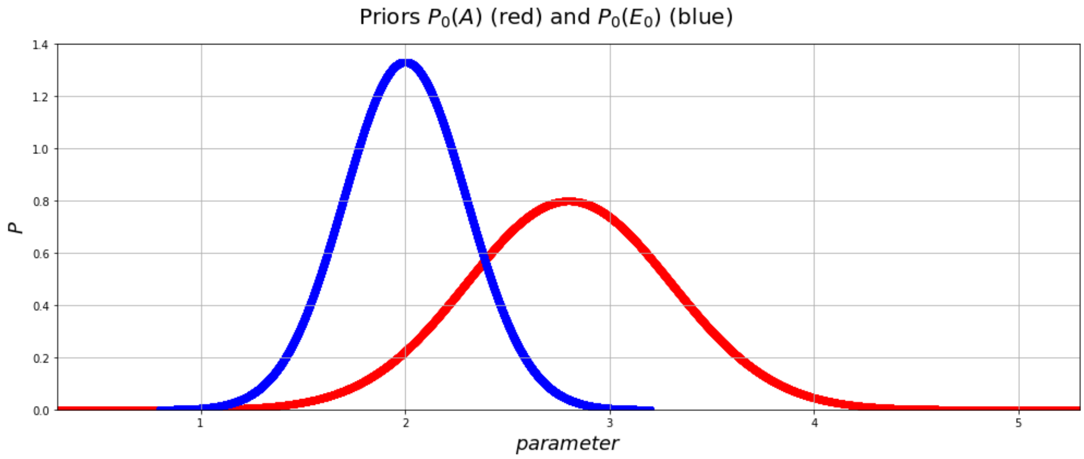


Figure 23: Plots of priors $P_0(A) = \mathscr{G}(A|\mu_A = 2.8, \sigma_A = 0.5)$ *and* $P_0(E_0) = \mathscr{G}(A|\mu_A = 2.0, \sigma_A = 0.3)$

A total of 1001 equally spaced values were generated for $A \in [0.3, 5.3]$ and $E_0 \in [0.8, 3.2]$ using the above priors. The results are shown in fig. 23. The joint probability distribution of the two parameters is:

$$P_0(A, E_0) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\vec{\lambda} - \vec{\mu})^T \Sigma^{-1} (\vec{\lambda} - \vec{\mu})\right] \tag{98}$$

where $\vec{\lambda} = (A, E_0)$ is a vector of our parameters, $\vec{\mu} = (\mu_A, \mu_{E_0})$ is the vector of means, and Σ is the covariance matrix:

$$\Sigma = \begin{pmatrix} cov[A, A] & cov[A, E_0] \\ cov[E_0, A] & cov[E_0, E_0] \end{pmatrix} \tag{99}$$

The covariance matrix was computed numerically using the `NumPy` package function `numpy.cov` and the generated vectors for A and E_0 from their priors. The joint probability distribution prior $P_0(A, E_0)$ is presented below in fig. 24.

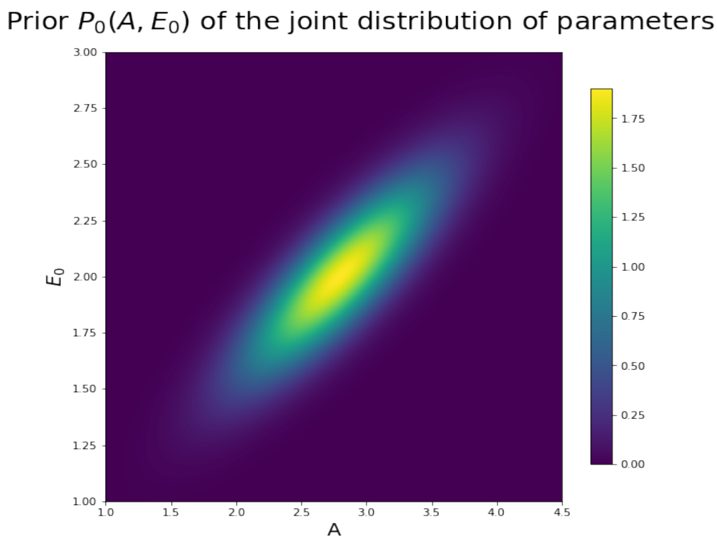


Figure 24: Joint probability distribution $P_0(A, E_0)$*, used as prior for Bayesian analysis.*

The analytical expression for the posterior is:

$$P(A, E_0|\vec{D}) = \frac{P(\vec{D}|A, E_0)P_0(A, E_0)}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(\vec{D}|A, E_0)P_0(A, E_0)\ dAdE_0} \tag{100}$$

where $\vec{D}$ represents the observed data, distributed according to:

$$P(\vec{D}|A, E_0) = \prod_{i=1}^{8} \frac{N_i!}{r_i!(N_i - r_i)!} p^{r_i}(1-p)^{(N_i - r_i)} \tag{101}$$

However, this being tedious to work with, numerical solutions were preferred. It was observed that integrating the prior $P_0(A, E_0)$ over $[0, 10]$ for A and E_0, and subtracting the value from 1, the result was of order of 10^{-11}. Therefore, the prior was considered normalized in the $[0, 10]$ intervals for both A and B, and the $[0, 10]$ interval was considered for the integral in the denominator from the posterior equation. The posterior was computed for a set of 1001 equally spaced values $E_0 \in [1.8, 2.15]$ and 1001 equally spaced values $A \in [2.5, 3.4]$. It is presented below in fig. 25. Its mode is at $A = 2.918$ and $E_0 = 1.974$, which are close values to the means of our priors.

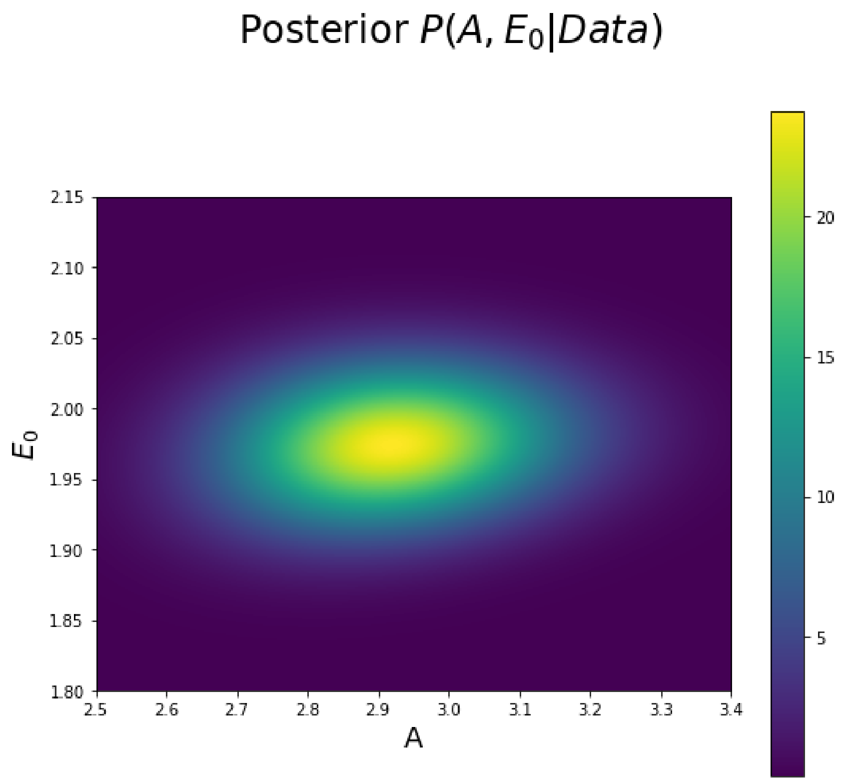


Figure 25: Posterior distribution $P(A, E_0)$.

For comparison purposes, the mode of the posterior and the $A = 2.82$ and $E_0 = 1.97$ point (selected from 30% confidence level) were used to fit the data. The plot is presented in fig. 26. It reveals that both methods of analysis yield parameters for which the model accurately describes the experimental data.

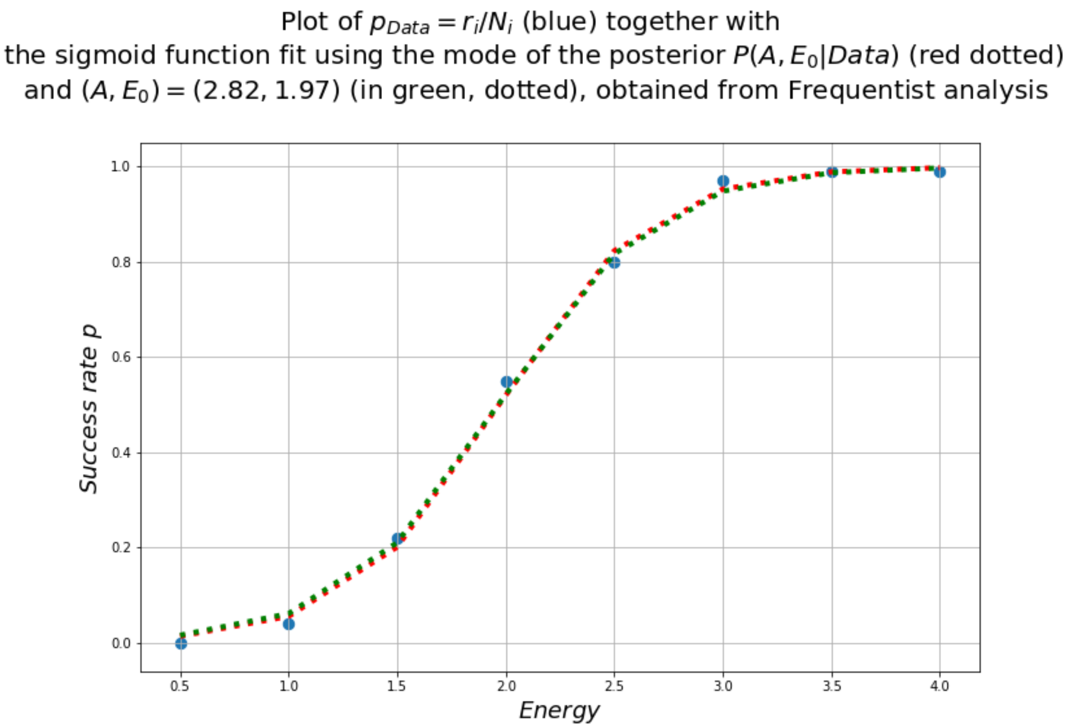


Figure 26: Bayesian analysis fit (red, dotted) and frequentist analysis fit (green, dotted) of the sigmoid function to the data (blue markers).

4.2 Exercise 5.2

Repeat the analysis of the data in the previous problem with the function

$$p(E) = \sin(A(E - E_0)) \tag{102}$$

(a) Find the posterior probability distribution for the parameters (A, E_0)

(b) Find the frequentist 68% CL region for (A, E_0).

(c) Discuss the results

a) For the Bayesian analysis, we perform the same procedure as described in the previous exercise. First, we note that $r/N = 0$ for $E = 0.5$. We therefore set the mean μ_{E_0} of the Gaussian prior to 0.25, since for values larger than 0.5, the boundary conditions are not satisfied anymore. For this value of E_0, A must be between $[0, 0.84]$, such that the argument inside the sin does not yield negative success probabilities p for any energy E_i. From this we conclude that $\mu_A = 0.42$ is a good mean for the Gaussian prior. The priors selected are therefore:

$$\begin{cases} P_0(A) = \mathscr{G}(A|\mu_A = 0.84, \sigma_A = 0.15) \\ P_0(E_0) = \mathscr{G}(A|\mu_A = 0.25, \sigma_A = 0.1) \end{cases} \tag{103}$$

The priors $P_0(A)$ and $P_0(E_0)$ were used to generate probabilities for 1001 equally spaced values of $A \in [0, 0.84]$ and $E_0 \in [0, 0.5]$. They are presented below, in fig. 27. These values are used to compute the covariance matrix of the parameters, such that the joint prior $P_0(A|E_0)$ is determined. The joint prior was normalized for $A \in [0, 0.84]$ and $E_0 \in [0, 0.5]$ by including a multiplication factor, and is presented below in fig. 28.

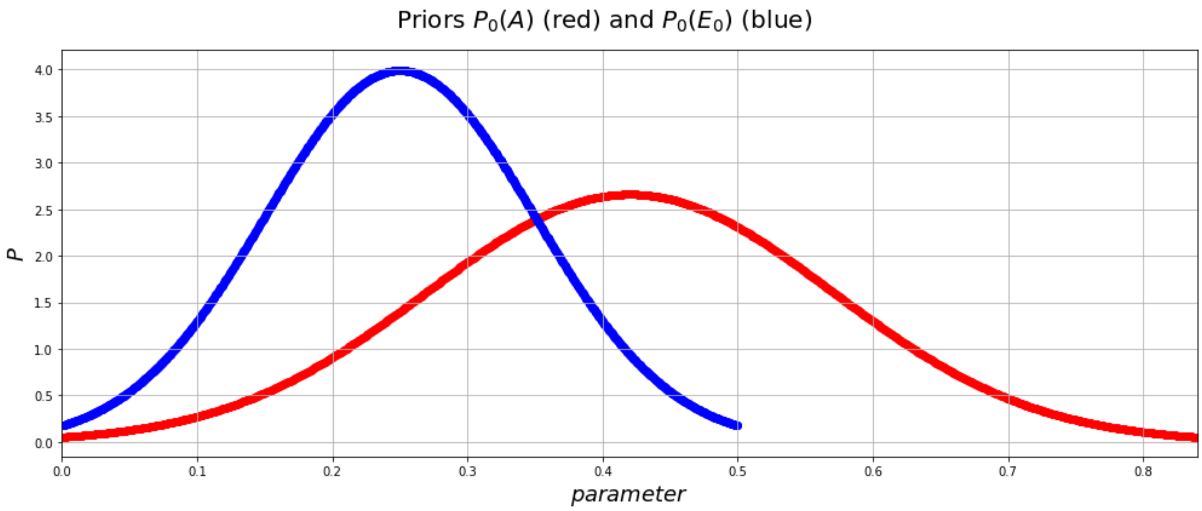


Figure 27: Priors $P_0(A)$ and $P_0(E_0)$ for $A \in [0, 0.84]$ and $E_0 \in [0, 0.5]$

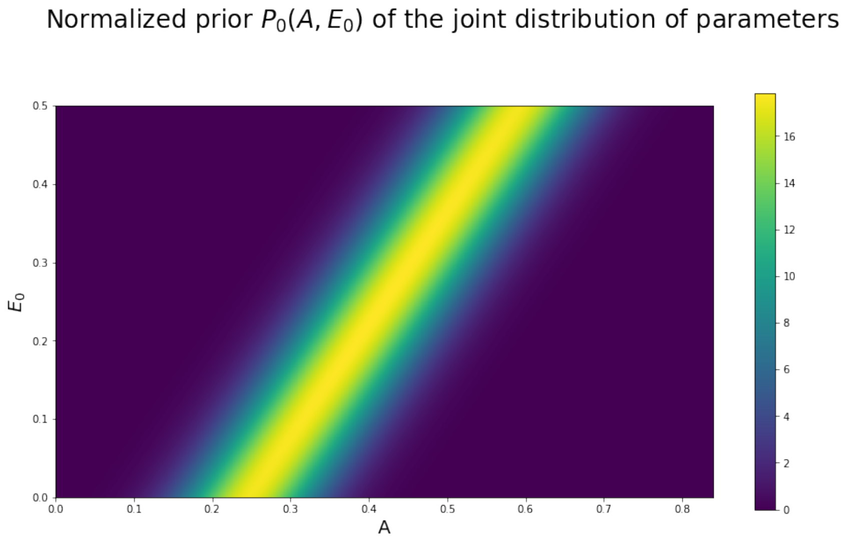


Figure 28: Normalized joint prior $P_0(A|E_0)$ for $A \in [0, 0.84]$ and $E_0 \in [0, 0.5]$.

The posterior $P(A, E_0|\vec{D})$ was computed numerically, and is presented below in fig. 29.

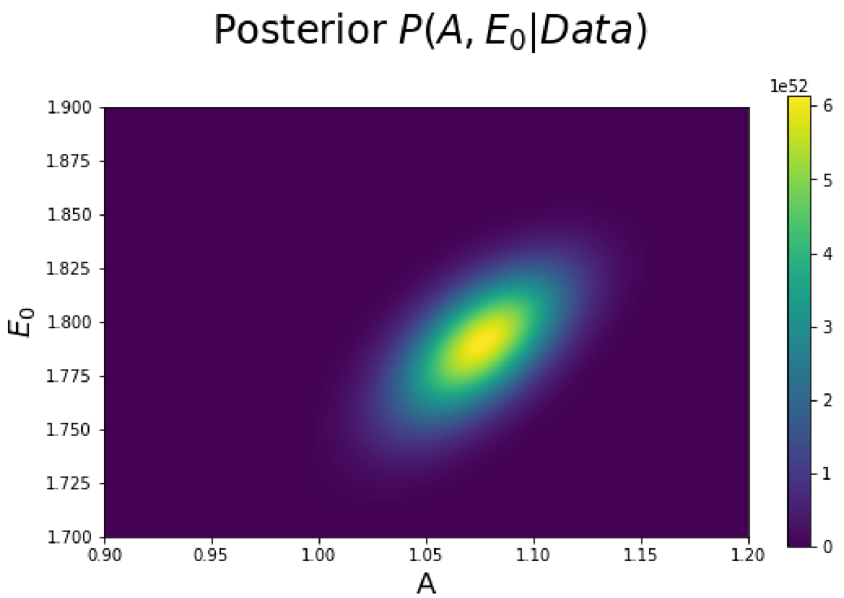


Figure 29: Posterior $P(A, E_0|\vec{D})$.

b) We now turn to a frequentist analysis to fit the given function to the data. The same test statistic and procedure outlined in the previous exercise was used. One key difference were the boundary conditions set for $A(E-E_0)$. Since p is a success probability, it cannot take negative values. Therefore we set the boundary conditions that $0 \leq A(E - E_0) \leq \pi$. Moreover, since E_0 is an energy value, we expect it to take only positive values $E_0 \geq 0$. To create the grid of (A, E_0) points, A was iterated from -1 to 1 for 201 equally spaced values, and E_0 was iterated from 0 to 10 for 201 equally spaced values. For each point in the grid, 1000 values of the ξ test statistic were generated, and the $\xi_{0.68}$ and $\xi_{0.90}$ values were recorded. The plot presented in fig. 30 shows that ξ^{data} fits in none of the intervals.

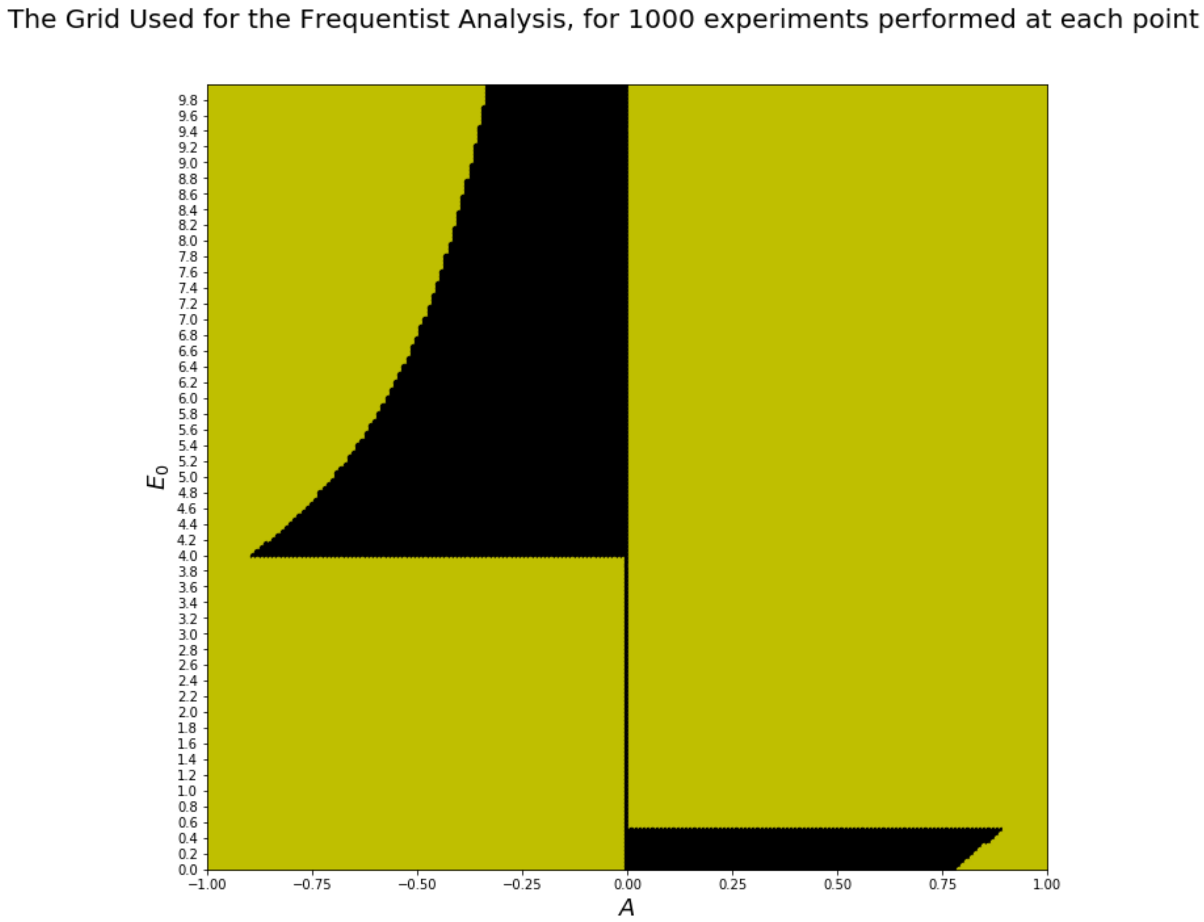


Figure 30: Frequentist analysis grid. Yellow points indicate that the $0 \leq A(E-E_0) \leq \pi$ *boundary conditions are not met*

Extending the range of A in this case makes little sense, since the boundary conditions would not be satisfied. Moreover, extending the range of E_0 to 25 shows again, no point for which ξ^{data} fits in either the 68% or the 90% interval (fig. 31). This is indicative of a flawed model to fit to the data.

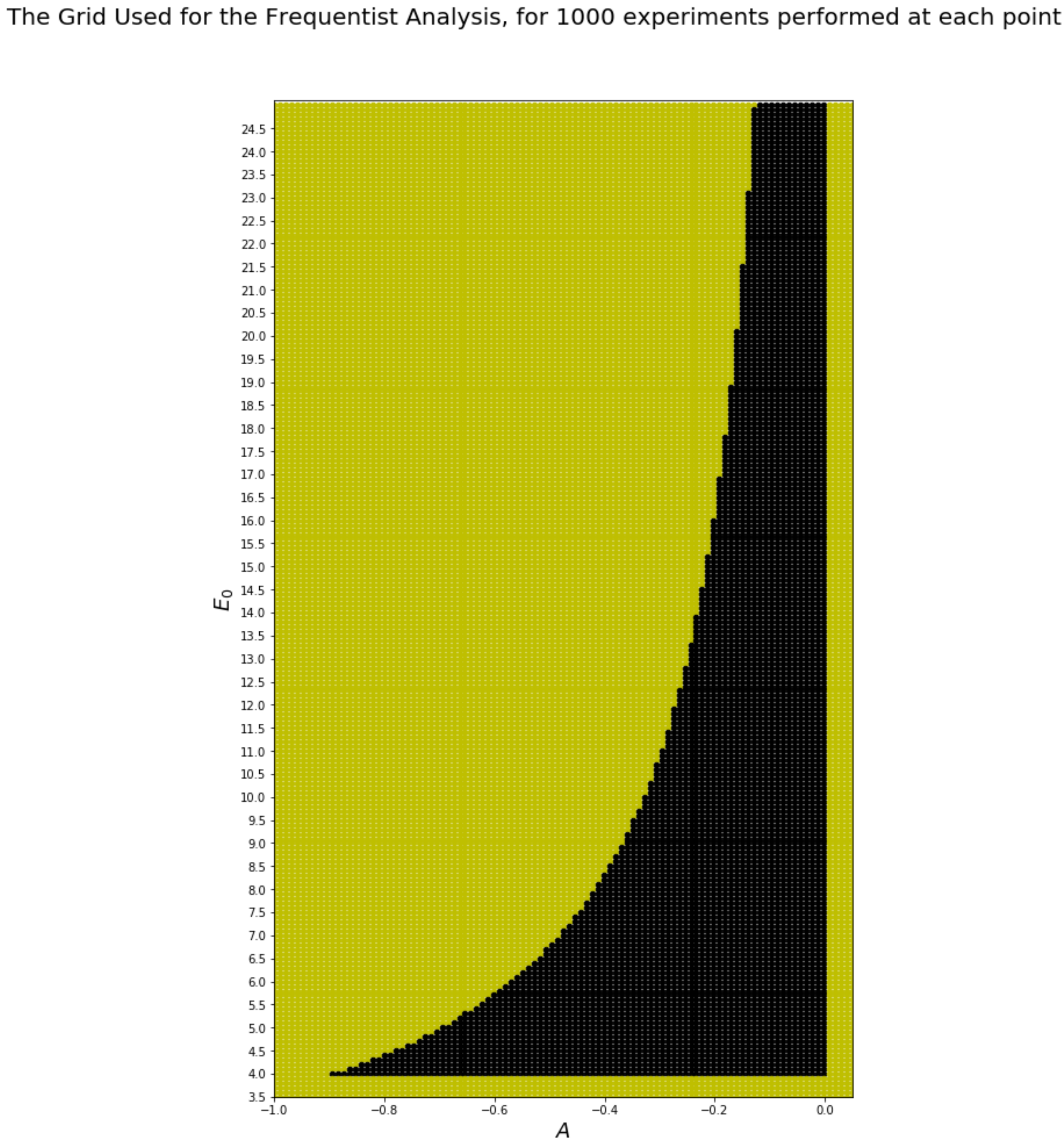


Figure 31: Frequentist analysis extended grid. Yellow points indicate that the $0 \leq A(E - E_0) \leq \pi$ *boundary conditions are not met*

c) In this case, we note that for the same data, the Frequentist and Bayesian analysis have yielded completely different results. While Frequentist analysis clearly shows that the model is a poor choice to fit the data, the Bayesian analysis does return a posterior distribution. Not surprising, since we expect Bayes Theorem to always return a prior, even when it makes little sense, such as it is the case now. In this sense, the Frequentist method is more reliable for cases when the model is wrong. This problem is a good example for the attention which needs to be given to results derived from Bayes Theorem. One needs a careful selection of a prior and a good understanding of the proccess generating the data to ensure that the results from Bayesian analysis are not wrongfully interpreted!

4.3 Exercise 5.3

Derive the mean, variance, and mode for the χ^2 distribution for one data point.

To derive an expression for the χ^2 distribution, we take $E_i = 2.5$. For this value, $N_i = 100$ and $r = 80$. We have seen that the binomial distribution can be approximated to a Gaussian for N and r large enough. We consider this to be the case.

$$P_0(r|N,P) \approx \frac{1}{\sqrt{\pi Np(1-p)}} \exp\left(-\frac{1}{2}\frac{(r-Np)^2}{Np(1-p)}\right) \tag{104}$$

where the mean $\mu = Np$ and standard deviation $\sigma = \sqrt{Np(1-p)}$. Since $p = p(E)$ and, we note that both μ and σ are functions of E. Considering our given model in exercise 1, where $p(E) = \frac{1}{1+\exp(-A(E-E_0))}$:

$$\begin{cases} \mu = \dfrac{N}{1+\exp(-A(E-E_0))} = f(E|A,E_0) \\ \\ \sigma = \dfrac{\sqrt{N\exp[-A(E-E_0)]}}{1+\exp[-A(E-E_0)]} = g(E|A,E_0) \end{cases} \tag{105}$$

The definition of χ^2 for our single data point is:

$$\chi^2 = \frac{r_i - f(E_i|A,E_0)}{g(E_i|A,E_0)} \tag{106}$$

Considering A and E_0 at the mode of the posterior probability distribution from exercise 1, $A = 2.918$ and $E_0 = 1.974$. For these values and $E_i = 2.5$, the value of χ^2 is 1.131. But we are more interested in the probability distribution $P(\chi^2)$:

$$P(\chi^2)\left|\frac{d\chi^2}{dr}\right| = 2P(r)$$

$$\left|\frac{d\chi^2}{dr}\right| = \frac{2(r - f(E|A,E_0))}{g(E|A,E_0)^2}$$

$$\left|\frac{d\chi^2}{dr}\right| = \frac{2\sqrt{\chi^2}}{g(E|A,E_0)} \tag{107}$$

$$\Longrightarrow P(\chi^2) = \frac{2g(E|A,E_0)}{2\sqrt{2\pi\chi^2}g(E|A,E_0)}e^{-\frac{\chi^2}{2}}$$

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

where we have used the Gaussian approximation for $P(r)$. The derived equation for the distribution of χ^2 has mean 1, variance 2 and mode at $\chi^{2*} = 0$. We note that they are all independent of the data or parameters A, E_0.

4.4 Exercise 5.8

Analyze the following data set assuming that the data can be modeled using a Gauss probability distribution where all data have the same uncertainty given by $\sigma = 4$. Try the two models:

1. quadratic, representing background only:

$$f(x|A, B, C) = A + Bx + Cx^2 \tag{108}$$

2. quadratic + Breit-Wigner representing background + signal:

$$f(x|A, B, C, x_0, \Gamma) = A + Bx + Cx^2 + \frac{D}{(x - x_0)^2 + \Gamma^2} \tag{109}$$

(a) Perform a chi-squared minimization fit, and find the best values of the parameters as well as the covariance matrix for the parameters. What is the p-value of the fits?

(b) Perform a Bayesian fit assuming flat priors for the parameters. Find the best values of the parameters as well as uncertainties based on the marginalized probability distributions. What is the Bayes Factor for the two models?

Table 8: Data for exercise 5.8

x	y	x	y
0.10	11.3	0.55	90.3
0.15	19.9	0.60	72.2
0.20	24.9	0.65	89.9
0.25	31.1	0.70	91.0
0.30	37.2	0.75	102.0
0.35	36.0	0.80	109.7
0.40	59.1	0.85	116.0
0.45	77.2	0.90	126.6
0.50	96.0	0.95	139.8

<u>**Model 1:**</u>
a) The definition of the chi^2 for a fitting function $f(x|\lambda)$ for Gaussian distributed data y_i is:

$$\chi^2 = \sum_i \frac{(y_i - f(x_i|\vec{\lambda}))^2}{\sigma_i^2} \tag{110}$$

where $\vec{\lambda}$ is the vector of parameters of the function. In our case, the parameters are A, B and C for the quadratic model, and the standard deviations σ_i of all measurements is $\sigma = 4$. The procedure is based on the proportionality relation between

$$P(A,B,C|Data) \propto e^{-\chi^2(A,B,C)} \tag{111}$$

which implies that minimizing $\chi^2(A,B,C)$ will maximize the posterior probability, and therefore return the mode of the posterior. We start by imposing the condition that $\left.\frac{\partial\chi^2}{\partial\lambda_k}\right|_{\lambda_k} = 0 \;\forall\; k$ Evaluating the partial derivative:

$$\frac{\partial\chi^2}{\partial\lambda_k} = -\sum_{i=1}^{18}\left[\frac{(y_i - f(x_i|\vec{\lambda}))^2}{\sigma^2}\frac{\partial f(x_i|\vec{\lambda})}{\partial\lambda_k}\right] \tag{112}$$

We make use of the notation $C_{ik} = \frac{\partial f(x_i|\vec{\lambda})}{\partial\lambda_k}$ and observe that $C_{i1} = 1$, $C_{i2} = x_i$ and $C_{i3} = x_i^2$. Remembering that $A = \lambda_1$, $B = \lambda_2$, $C = \lambda_3$, one can rewrite $\left.\frac{\partial\chi^2}{\partial\lambda_k}\right|_{\lambda_k} = 0$ as

$$\sum_{i=1}^{18}\frac{y_i}{\sigma}C_{ik} = \sum_{i=1}^{18}\sum_{l=1}^{3}\frac{\lambda_l C_{il} C_{ik}}{\sigma^2} \tag{113}$$

or in matrix form:

$$\vec{Y} = M\cdot\vec{\lambda} \tag{114}$$

where $Y_k = \sum_{i=1}^{18}\frac{C_{ik}}{\sigma^2}\,y_i$ and $M_{kl} = \sum_{i=1}^{18}\frac{C_{ik}C_{il}}{\sigma^2}$. If the determinant of M $|M| \neq 0$, the matrix is invertible, and eq. 114 has solution

$$\hat{\vec{\lambda}} = M^{-1}\,\vec{Y} \tag{115}$$

where the inverse M^{-1} is the covariance matrix of the parameters. For the given data and model 1, M, M^{-1} and $\vec{Y}$ were computed numerically:

$$\begin{cases} M = \begin{pmatrix} 1.125 & 0.591 & 0.386 \\ 0.591 & 0.386 & 0.282 \\ 0.386 & 0.282 & 0.220 \end{pmatrix} \\ \\ M^{-1} = \begin{pmatrix} 15.280 & -61.114 & 51.600 \\ -61.114 & 286.274 & -260.062 \\ 51.600 & -260.062 & 247.678 \end{pmatrix} \\ \\ \vec{Y} = (83.138, 54.484, 39.771) \end{cases} \tag{116}$$

The resulting values for the parameters at the mode of the posterior are $A = -7.268$, $B = 173.467$, $C = -28.878$. These values were used to fit the data in fig. 32. It can be observed that our current model does a poor job at fitting the data, especially in the middle region of the x-interval.

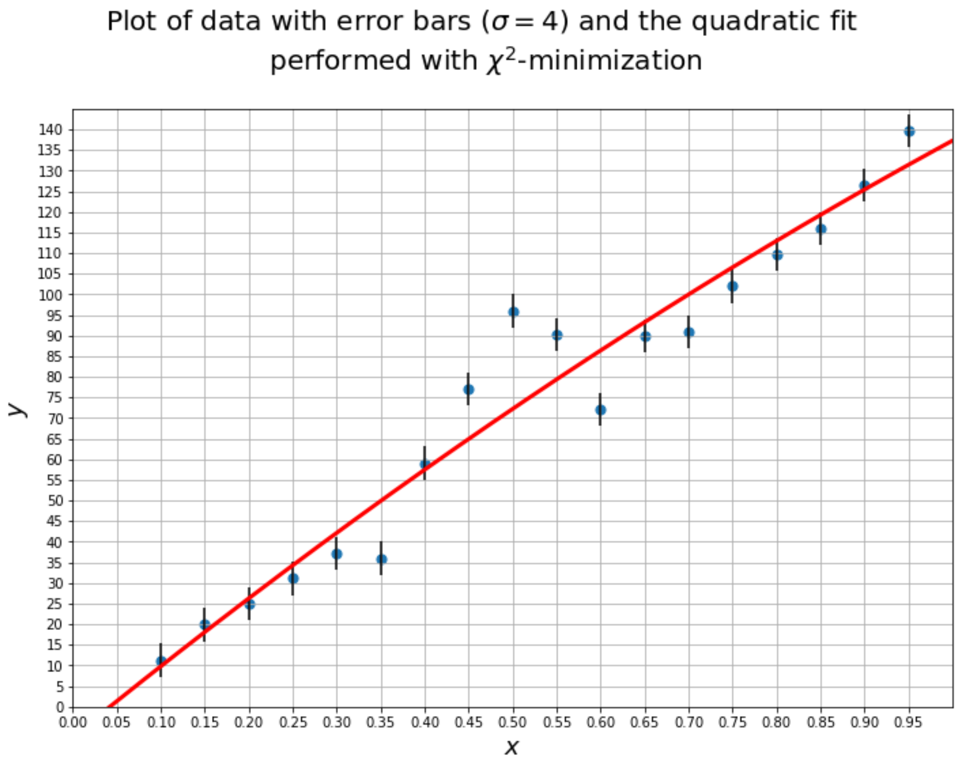


Figure 32: Observed data together with the quadratic model fit for $A = -7.268$, $B = 173.467$, $C = -28.878$.

The χ^2 distribution for n sets of data is given by:

$$P(\chi^2|n) = \frac{(\chi^2)^{n/2-1}}{2^{n/2}\Gamma(\frac{n}{2})}\, e^{-\chi^2/2} \tag{117}$$

where $\Gamma(x)$ is the *Gamma*-function, taking the value $(x-1)!$ when x is an integer, which is the case at hand. To compute the p-value of the fits, one must calculate $\chi^2_{A,B,C}$ for the mode of the posterior and evaluate the integral:

$$\int_{\chi^2_{A,B,C}}^{\infty} P(\chi^2|n) d\chi^2 \tag{118}$$

This was performed numerically and resulted in a p-value of $5.109 \cdot 10^{-12}$. This is an extremely small value, indicative of a poor fit to the data.

b) To perform Bayesian analysis, one starts from the Gaussian probability distribution for our 18 data points $\vec{D}$:

$$P(\vec{D}|\vec{\lambda}) = \prod_{i=1}^{18} \left(\frac{1}{\sqrt{2\pi}\sigma}\right) \exp\left[-\frac{1}{2}\sum_{i=1}^{18} \frac{(y_i - f(x_i|\vec{\lambda}))^2}{\sigma^2}\right] \tag{119}$$

For the case of flat priors for A, B and C, the posterior $P(A, B, C|Data) \propto e^{-\chi^2(A,B,C)}$ and therefore finding its mode is the same procedure as the one performed above for χ^2 minimization. The mode of the posterior is at $A = -7.268$, $B = 173.467$, $C = -28.878$. Furthermore, since M^{-1} is the covariance matrix, its diagonals contain the variances of each parameter. Uncertainties for the parameters can be found by considering their standard deviations $\sigma_A = 3.909$, $\sigma_B = 16.920$, $\sigma_C = 15.738$.

<u>Model 2:</u>

$$f(x|A, B, C, x_0, \Gamma) = A + Bx + Cx^2 + \frac{D}{(x - x_0)^2 + \Gamma^2} \tag{120}$$

a) For the χ^2 minimization procedure, as outlined above, one needs to determine the coefficients $C_{ik} = \frac{\partial f}{\partial \lambda_k}$ by finding the derivatives of $f(x|A, B, C, x_0, \Gamma)$ with respect to all parameters. This is needed for computing the terms in eq. 114 and solving for $\vec{\lambda}$. The required derivatives are:

$$\begin{cases} C_{i1} = \frac{\partial f}{\partial A} = 0 \\ C_{i2} = \frac{\partial f}{\partial B} = x_i \\ C_{i3} = \frac{\partial f}{\partial C} = x_i^2 \\ C_{i4} = \frac{\partial f}{\partial x_0} = \frac{2D(x-x_0)}{((x-x_0)^2+\Gamma^2)^2} \\ C_{i5} = \frac{\partial f}{\partial \Gamma} = \frac{-2D\Gamma}{((x-x_0)^2+\Gamma^2)^2} \end{cases} \tag{121}$$

The next step in the procedure is finding an analytical form for the elements of $\vec{Y}$ and M:

$$\begin{cases} Y_k = \sum_{i=1}^{18} \frac{C_{ik}}{\sigma^2} y_i \\ M_{kl} = \sum_{i=1}^{18} \frac{C_{ik}C_{il}}{\sigma^2} \end{cases} \tag{122}$$

Replacing these in eq. 114 and rearranging the terms will result in a system of 5 equations with 5 unknown variable (the parameter in $\vec{\lambda}$) which would yield the mode values of the parameters of the posterior distribution. Since solving this analytically is very tedious, numerical solutions are preferred.

5 Extra Exercises

5.1 Exercise 1

The family of Bernoulli distributions have the probability density $P(x|p) = p^x(1-p)^{1-x}$.

(a) Calculate the Fisher Information $I(p) = -E\left[\frac{\partial^2 \ln P(x|p)}{\partial p^2}\right]$

(b) What is the maximum likelihood estimator for p?

(c) What is the expected distribution for $\hat{p} - p_0$?

a) To determine the Fisher information $I(p)$, we start from the expression

$$\begin{aligned}
&\ln P(x|p) = x \ln p + (1-x)\ln(1-p) \\
\implies &\frac{\partial \ln P(x|p)}{\partial p} = \frac{x}{p} - \frac{1-x}{1-p} \\
\implies &\frac{\partial^2 \ln P(x|p)}{\partial p^2} = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}
\end{aligned} \tag{123}$$

Using $E[x] = p$, the Fisher information expression becomes:

$$\begin{aligned}
I(p) &= \frac{E[x]}{p^2} + \frac{1-E[x]}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} \\
&= \frac{1}{p(1-p)}
\end{aligned} \tag{124}$$

b) The likelihood for some x_i Bernoulli variables is defined as:

$$\begin{aligned}
\mathscr{L}(p) &= \prod_{i=1}^{n} P(x_i|p) \\
&= p^{\sum x_i}(1-p)^{n-\sum x_i}
\end{aligned} \tag{125}$$

For the mode of the likelihood, consider $\ln \mathscr{L}(p)$ instead. Since ln is an increasing function, maximizing $\ln \mathscr{L}(p)$ will give the same result as maximizing $\mathscr{L}(p)$. The advantage of working with the ln function in the expression

$$
\begin{aligned}
\ln \mathscr{L}(p) &= \ln p \sum x_i + (n - \sum x_i) \ln p \\
\implies \frac{\partial \ln \mathscr{L}(p)}{\partial p} &= \frac{\sum x_i}{p} - \frac{(n - \sum x_i)}{1-p}
\end{aligned} \tag{126}
$$

Setting the condition that the derivative is 0 results in:

$$
\begin{aligned}
(1-p)\sum x_i - (n - \sum x_i)p = 0 \\
\implies \sum x_i - np = 0 \\
\implies \hat{p} = \frac{\sum x_i}{n}
\end{aligned} \tag{127}
$$

c) The expected distribution for $\hat{p} - p_0$ is $P(\hat{p} - p_0) = \mathscr{G}(0, \frac{1}{nI(p)})$. Replacing the found expressiong for the Fisher information:

$$
\begin{aligned}
P(\hat{p} - p_0) &= \mathscr{G}(0, \frac{1}{nI(p)}) \\
&= \mathscr{G}(0, \frac{p(1-p)}{n})
\end{aligned} \tag{128}
$$

i.e. it is a Gaussian with mean 0 and standard deviation $\frac{p(1-p)}{n}$.

5.2 Exercise 2

The family of exponential distributions have pdf $P(x|\lambda) = \lambda e^{-\lambda x}$, $x > 0$.

(a) Generate $n = 2, 10, 100$ values of x using $x = -\ln U$ where U is a uniformly distributed random number between $(0, 1)$. Find the MLE estimator from your generated data. Repeat this for 1000 experiments and plot the distribution of the maximum likelihood estimator, $\hat{\lambda}$ (note that the true value in this case is $\lambda_0 = 1$).

(b) Compare the distributions you found for the MLE to the expectation from the Law of Large Numbers and CLT (see lecture notes) and discuss.

The plots below (fig. 33) show the distribution of the 1000 MLE values of the data generated for $n = 2$, 10, 100. The Law of Large Numbers (LLN) applied in this case tells us that as $n \to \infty$, $L_n(\lambda) \to L_(\lambda)$. This implies that the mode $\hat{\lambda}$ will approach the true value λ_0. This can be easily observed empirically from the plots: the larger n is, the closer the mode $\hat{\lambda}$ is to the true value of 1.

The CLT in this case tells us that as $n \to \infty$, the distribution $P(\hat{\lambda} - \lambda_0) \to \mathscr{G}(0, \frac{1}{nI(\lambda)})$, where $I(\lambda)$ is the Fisher information. This implies that for large n, not only can the distribution be approximated to a Gaussian, but it also becomes narrower. This can be observed by analyzing the plots for different n. As the mode of $P(\hat{\lambda} - \lambda_0)$ approaches 0, the mode of $P(\hat{\lambda}$ approaches $\lambda_0 = 1$, which is again obvious from the plots in fig. 33.

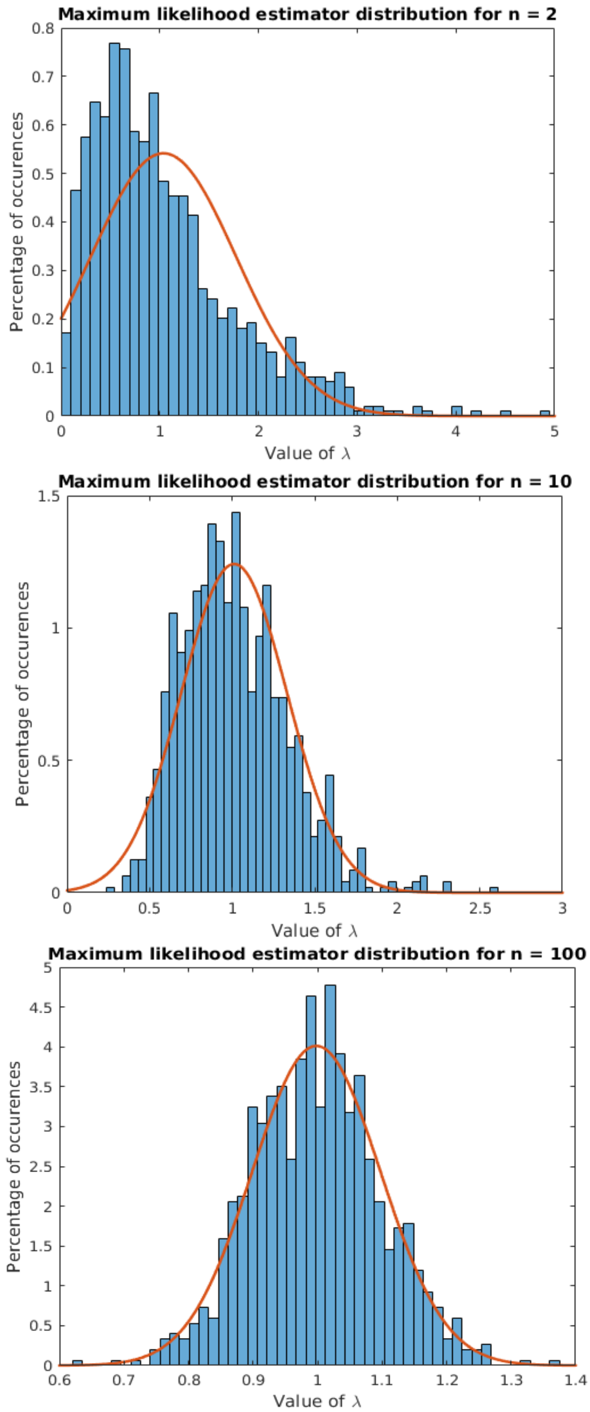


Figure 33: Normalized histograms (50 bins) of 1000 *values of the MLE for* $n = 2$*,* 10 *and* 100 *numbers generated from the exponential distribution. The Gaussian distributions expected to approximated the histograms, as per CLT and LLN, are plotted in orange.*