Report on the lecture *Data Analysis*, WS17/18

Philipp Holl

February 15, 2018

Abstract

The lecture *Data Analysis* by Allen Caldwell covers the statistical analysis of measured data from a frequentist's and a Bayesian point of view. This report contains my solutions to the exercises which accompany the lecture.

Contents

1 Chapter 1: Probabilistic reasoning

Notation for chapter 1:

An event or discrete random variable is denoted by an upper case letter and its inverse with a bar on top.

The probability for a certain outcome ω or value of a random variable $X = x$ is denoted $P(\omega)$ and $P(X = x) \equiv P(x)$, respecively.

The joint probability of A and B is $P(A, B)$ and the conditional probability is $P(A|B)$.

The function $\mathbb{1}$[condition] is 1 if the condition is fulfilled, 0 otherwise.

Exercise 1: Janes Children

You meet Jane on the street. She tells you she has two children, and has pictures of them in her pocket. She pulls out one picture, and shows it to you. It is a girl. What is the probability that the second child is also a girl ? Variation: Jane takes out both pictures, looks at them, and is required to show you a picture of a girl if she has one. What is now the probability that the second child is also a girl ?

Solution:

Jane has two children 1 and 2 which can be either girls G or boys $\bar{G}$. The prior probability for a child to be a boy or a girl are equal.

$$P(G) = P(\bar{G}) = \frac{1}{2}$$

Jane shows us that one of her children, which we label as 1, is a girl, $P(G_1) = 1$. We can express the probability for child 2 to also be a girl as

$$P(G_2) = P(G_2|G_1)P(G_1) + P(G_2|\bar{G}_1)P(\bar{G}_1) = P(G_2|G_1)$$

Assuming that all children of Jane are independent and identically distributed (i.i.d.), we see that

$$P(G_2|G_1) = P(G_2) = P(G) = \frac{1}{2}$$

Variation: Jane only shows us a picture of a girl, G_1, if she has one and we are interested in the other child 2. We call the event that a picture is shown P. If Jane does not show us a picture, she has no girl by definition, so $P(G_2|\bar{P}) = 0$. Otherwise we want to know the probability of child 2 to be a girl

$$P(G_2|P) = \frac{P(P|G_2) \cdot P(G_2)}{P(P)} = \frac{1 \cdot P(G_2|G_1)P(G_1)}{\sum_{A,B \in \{G,\bar{G}\}} \mathbb{1}[A = G \text{ or } B = G]P(A, B)} = \frac{1/2 \cdot 1/2}{3 \cdot 1/4} = \frac{1}{3}$$

where we have used the above result that $P(G_1) = P(G_2|G_1) = P(G) = 1/2$.

Exercise 2: More definitions of data probability

Go back to section 1.2.3 and come up with more possible definitions for the probability of the data.

Solution:

Tossing a coin 10 times yields a sample space Ω with 2^{10} possible outcomes. Let the event space

$\mathcal{F}$ be the power set of Ω. We can then define events $A \in \mathcal{F}$ and assign probabilities to them $P : \mathcal{F} \to [0, 1]$. The only requirements are $P(A) \leq 0 \; \forall A \in \mathcal{F}$, $P(\Omega) = 1$ and for disjoint events $A_1, A_2, \ldots$: $P(\bigcup_i A_i) = \sum_i P(A_i)$.

We are usually only interested in some aspects of a random experiment. We therefore define (discrete) random variable $X : \Omega \to \mathbb{R}$, defining $P(X = x) \equiv P(\{\omega \in \Omega : X(\omega) = x\})$.

In the concrete case we could for example only measure the first toss with a random variable.

$$X_1 \equiv \mathbb{1}[\text{"first toss shows heads"}] \qquad P(X_1) = 50\%$$

Or the probability that there are at least six tails in a row

$$X_2 \equiv \mathbb{1}[\text{"six tails in a row"}] \qquad P(X_2) = (1/2)^6 \cdot 5 \approx 7,8\%$$

Exercise 3: Energy resolution of a particle detector

Your particle detector measures energies with a resolution of 10 %. You measure an energy, call it E. What probabilities would you assign to possible true values of the energy ? What can your conclusion depend on ?

Solution:

When an event with true energy T occurs, our detector outputs a value E that has a distribution $P(E|T)$ around T with a standard deviation of 10%.

$$\sigma = \sqrt{\int_0^\infty dE \; (E - \bar{E})^2 P(E|T)} = 0.1$$

The true value T can be expressed using Bayes' theorem.

$$P(T|E) = \frac{P(E|T)P(T)}{P(E)}$$

While $P(E|T)$ might be approximately known (e.g. a normal distribution), $P(T)$, the prior distribution of true energies, is at this point unknown. Depending on the experiment, there may only be specific energies allowed or a certain range of energies favored over others. This influences the likely value of $P(T|E)$. Without this knowledge, one could use a uniform distribution $P(T) = \text{const.}$ and arrive at $P(T|E) = P(E|T)$ with the same 10% uncertainty.

Exercise 4: Mongolian swamp fever

Mongolian swamp fever is such a rare disease that a doctor only expects to meet it once every 10000 patients. It always produces spots and acute lethargy in a patient; usually (I.e., 60 % of cases) they suffer from a raging thirst, and occasionally (20 % of cases) from violent sneezes. These symptoms can arise from other causes: specifically, of patients that do not have the disease: 3 % have spots, 10 % are lethargic, 2 % are thirsty and 5 % complain of sneezing. These four probabilities are independent. What is your probability of having Mogolian swamp fever if you go to the doctor with all or with any three out of four of these symptoms ? (From R.Barlow)

Solution:

A patient going to the doctor might have the Mongolian swamp fever F or not $\bar{F}$. The disease is

very rare $P(F) = \frac{1}{10000}$. The four kind of recorded symptoms with given probabilities are listed below.

Symptom S	$P(S\|F)$	$P(S\|\bar{F})$
Spots P	1	3%
Accute lethargy L	1	10%
Raging thirst T	60%	2%
Violent sneezes V	20%	5%

The probability that a patient with all of these symptoms has the swamp fever is given by

$$P(F|P,L,T,V) = \frac{P(P,L,T,V|F)\cdot P(F)}{P(P,L,T,V)} = \frac{P(P|F)P(L|F)P(T|F)P(V|F)\cdot P(F)}{\sum_{f\in\{F,\bar{F}\}}\prod_{S\in\{P,L,T,V\}}P(S|f)P(f)}$$

where we have assumed that all symptoms are independent of each other, both with and without fever. The numerator, which also appears in the denominator, yields $\frac{0.6\cdot 0.2}{10000} = 1.2\cdot 10^{-5}$. The probability of all symptoms without fever is $\frac{9999}{10000}\cdot 0.03\cdot 0.1\cdot 0.02\cdot 0.05 = 3\cdot 10^{-6}$. Inserting this results in

$$P(F|P,L,T,V) = \frac{1.2\cdot 10^{-5}}{1.2\cdot 10^{-5} + 3\cdot 10^{-6}} = 80.0\%$$

If the symptoms P or L are missing, there is no chance of Mongolian swamp fever. Assuming these two and T are present, the probability of the disease is

$$P(F|P,L,T,\bar{V}) = \frac{P(P,L,T,\bar{V}|F)\cdot P(F)}{P(P,L,T,\bar{V}|F) + P(P,L,T,\bar{V}|\bar{F})} = \frac{4.8\cdot 10^{-5}}{4.8\cdot 10^{-5} + 5.7\cdot 10^{-5}} = 45.7\%$$

Should violent sneezes accompany P and L, the probability is

$$P(F|P,L,\bar{T},V) = \frac{8\cdot 10^{-6}}{8\cdot 10^{-6} + 1.47\cdot 10^{-4}} = 5.2\%$$

2 Chapter 2: Binomial and Multinomial Distribution

Exercise 8: Confidence intervals

For the following function

$$P(x) = xe^{-x} \qquad 0 \leq x \leq \infty$$

(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 mode and 68 % smallest interval

Solution:

a)

The mean $\mathbb{E}[x]$ is given by

$$\mathbb{E}[x] = \int_0^\infty dx\ x\ P(x) = \int_0^\infty dx\ x^2\ e^{-x} = 2$$

The standard deviation is defined as

$$\sigma^2 = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \int_0^\infty dx\ x^2\ P(x) - \mathbb{E}[x]^2 = 6 - 4 = 2$$

$$\sigma = \sqrt{2} \approx 1,41$$

The total probability content within the interval $[\mathbb{E}[x] - \sigma, \mathbb{E}[x] + \sigma] = [2 - \sqrt{2}, 2 + \sqrt{2}]$ is

$$\int_{2-\sqrt{2}}^{2+\sqrt{2}} dx\ P(x) = \frac{6\sinh(\sqrt{2}) - 2\sqrt{2}\cosh(\sqrt{2})}{e^2} \approx 73,8\%$$

b)

The median m satisfies the condition

$$\int_0^m dx\ P(x) = \int_m^\infty dx\ P(x)$$

Inserting $P(x) = x\ e^{-x}$ yields

$$(m+1)e^{-m} = \frac{1}{2}$$

which has exactly one solution for $x \geq 0$ given by the analytic continuation of the product log function

$$m = -W_{-1}\left(\frac{-1}{2e}\right) - 1 \approx 1.68$$

The 68% central interval $[r_1, r_2]$ is constructed such $(1 - 68\%)/2 = 15.87\%$ probability content is located to the left of r_1 and to the right of r_2.

$$\int_0^{r_1} dx\ P(x) = 15.87\% \Rightarrow r_1 \approx 0.71$$

$$\int_{r_2}^\infty dx\ P(x) = 15.87\% \Rightarrow r_2 \approx 3.30$$

c)

The mode x_m is the point for which $P(x)$ reaches its maximum. It fulfills

$$\frac{dP}{dx}(x_m) = 0$$

Solving the equation yields $x_m = 1$. The smallest interval (highest posterior density interval) $[s_1, s_2]$ is constructed around the mode so that it fulfills two conditions:

(i) It contains at least 68% of the total probability. For the given continuous function $P(x) = x\, e^{-x}$, this translates to

$$\int_{s_1}^{s_2} dx\; x\; e^{-x} \stackrel{!}{=} 68,27\%$$

(ii) For the second condition, there are a couple of equivalent formulations. The lecture stated $P(s_1) = P(s_2)$. Here we use the fact that it is the smallest interval fulfilling condition (i).

$$s_2 - s_1 \text{ minimal} \qquad \Rightarrow \qquad \frac{d(s_2(s_1) - s_1)}{ds_1} = 0$$

Solving the first equation for s_2 yields

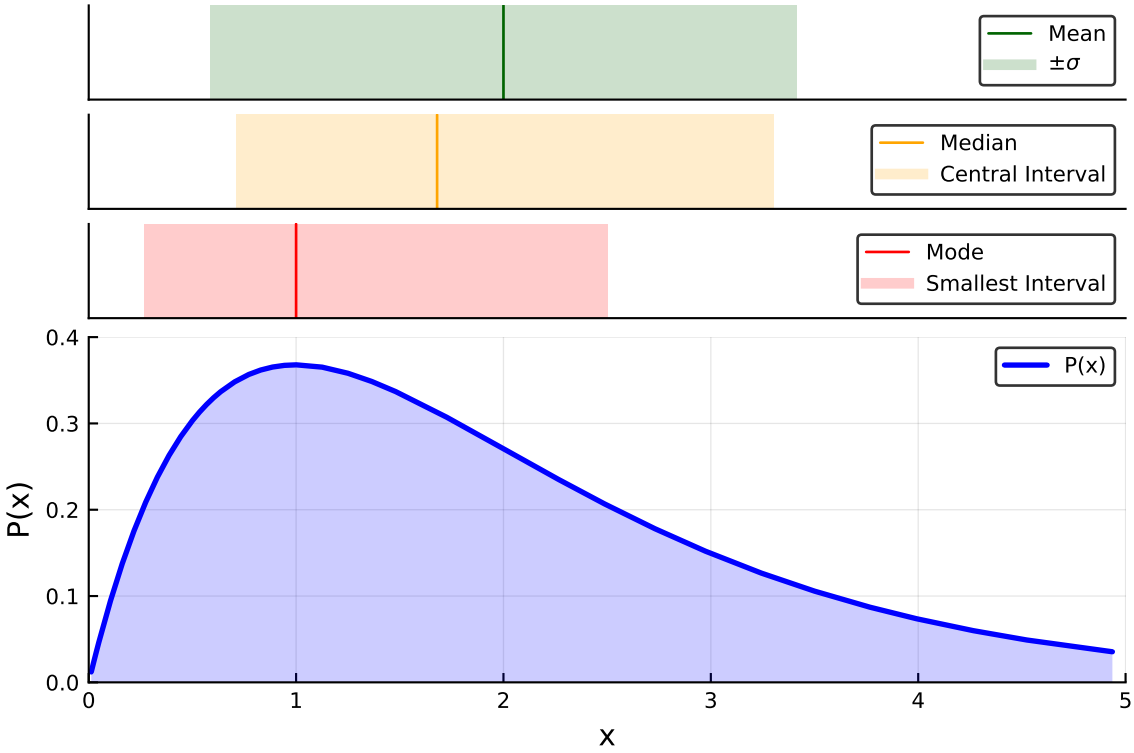
$$s_2(s_1) = -W_{-1}\left(68,27\% - e^{-s_1-1}(s_1+1)\right) - 1$$

where W again refers to the product log function. The second condition condition fixes both s_1 and s_2.

$$s_2 - s_1 \text{ minimal} \qquad \Rightarrow \qquad \frac{d}{ds_1} W_{-1}\left(68,27\% - e^{-s_1-1}(s_1+1)\right) \stackrel{!}{=} -1$$

This equation can be solved numerically, resulting in $s_1 \approx 0,27$ and $s_2 \approx 2,50$.

The following image shows the probability density function $P(x)$ as well as the three intervals.



Exercise 10: Bayesian analysis of binomials

Consider the data in the table:

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

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.

Solution:

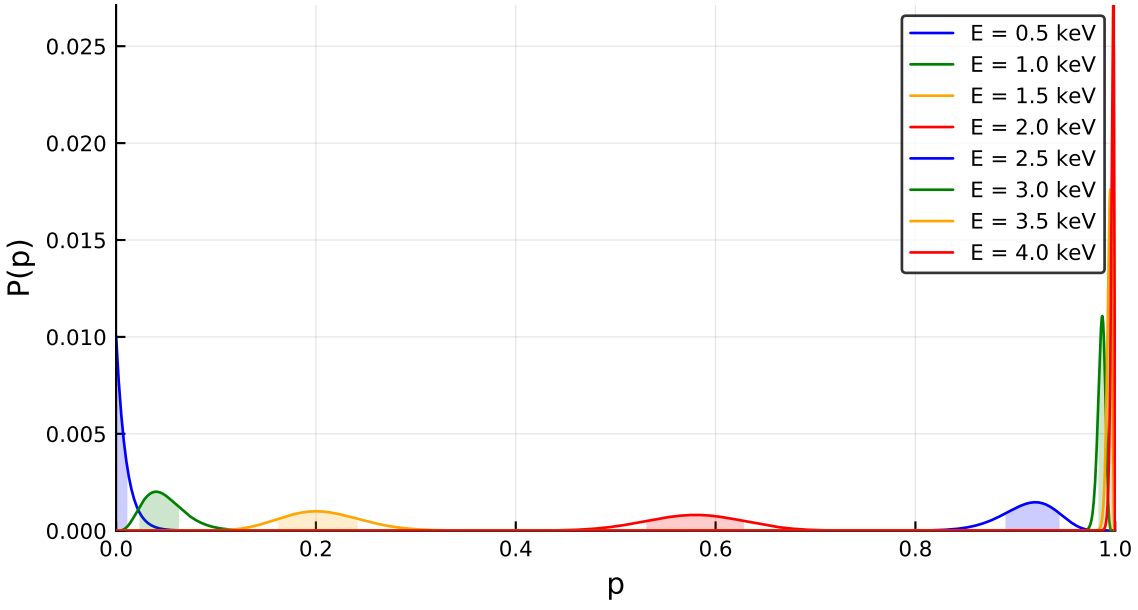
The number of trials N is fixed for each energy E and the number of successes r is observed. $P(r|N,p)$ therefore is a binomial distribution. Using Bayes' theorem, the posterior probability for p can be written as

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

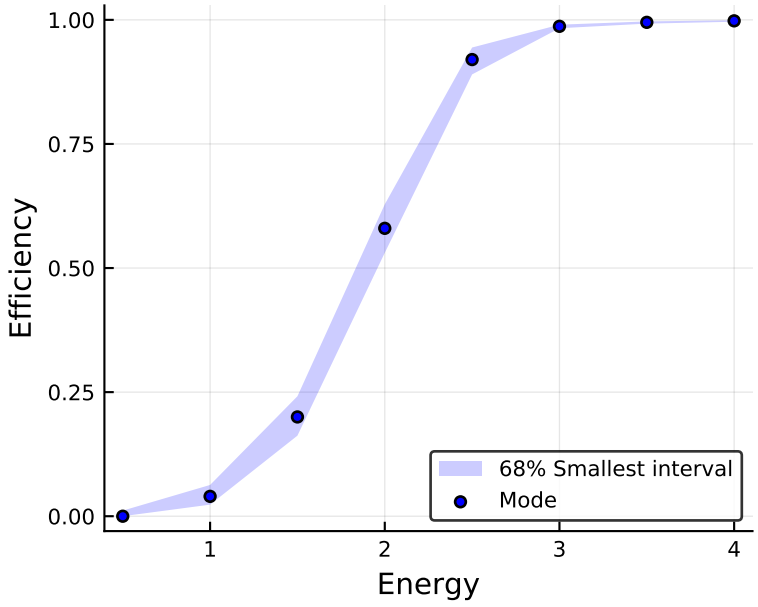
where we have inserted the flat prior $P(p) = \mathbb{1}[0 \leq x \leq 1]$ and used the fact that $P(r|N)$ does not depend on p. The mode of the posterior can then easily be calculated to be

$$\text{mode}[P(p|N,r)] = \frac{r}{N}$$

The 68% smallest interval (highest posterior density interval) can be computed numerically. The resulting posterior densities and their corresponding 68% intervals are plotted below.



Plotting the modes and intervals against the energy results in the plot below. The table next to it shows the expected efficiencies for p with their respective 68% intervals.



Energy	Efficiency
0.5	$0.0^{+0.011}_{-0.0}$
1.0	$0.04^{+0.023}_{-0.017}$
1.5	$0.2^{+0.042}_{-0.038}$
2.0	$0.58^{+0.048}_{-0.049}$
2.5	$0.92^{+0.024}_{-0.03}$
3.0	$0.987^{+0.003}_{-0.004}$
3.5	$0.995^{+0.002}_{-0.003}$
4.0	$0.998^{+0.001}_{-0.002}$

Exercise 11: Frequentist analysis of binomials

Analyze the data in the table (exercise 10) 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.

Solution:

We keep the notation from the previous exercise. The 90% central interval $\mathcal{O}^C_{0.90}$ can be computed for every p by finding the maximum r_1 and minimum r_2 such that

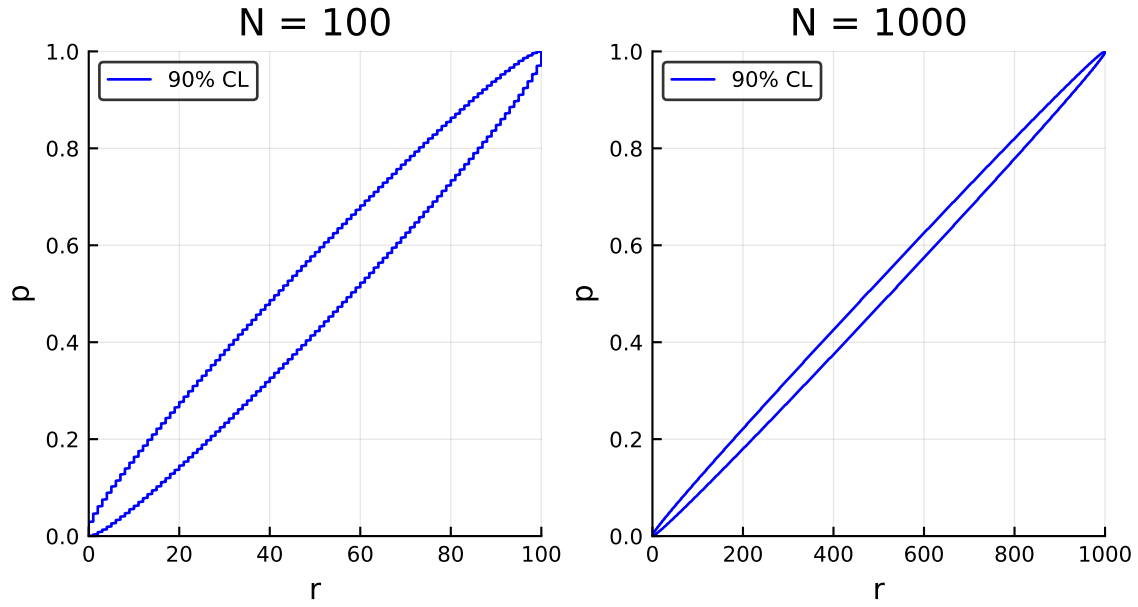
$$\sum_{r=0}^{r_1} P(r|N,p) \leq 5\% \qquad \sum_{r=r_2}^{N} P(r|N,p) \leq 5\%$$

For a given measurement r_D, the 90% CL interval for p is then constructed as

$$I_{0.90} = \{p \in [0,1] : r_D \in \mathcal{O}^C_{0.90}(p)\}$$

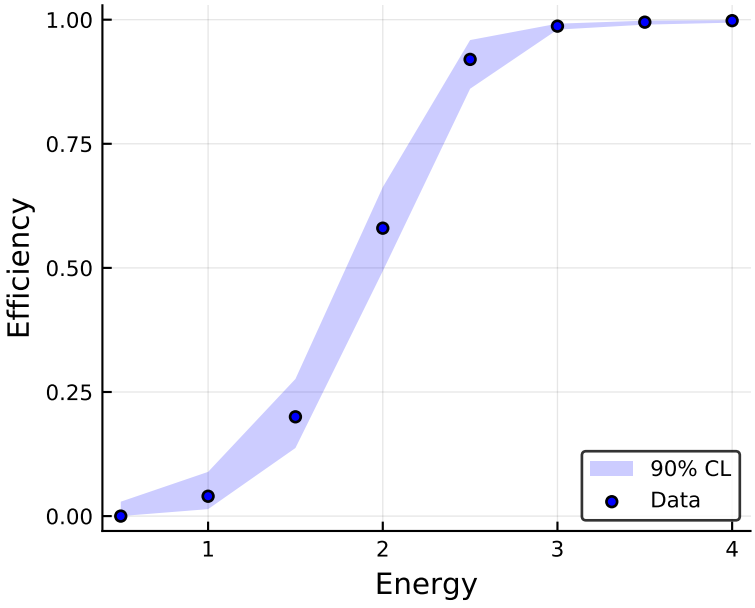
For the case of binomials, this set is always a compact interval. Its supremum and infimum will be referred to as $p^{\min}_{0.90}$ and $p^{\max}_{0.90}$, respectively.

For any given N, like the occuring cases of $N = 100$ and $N = 1000$, the Neyman 90% confidence intervals can be computed before the experiment is done. Performing this on a computer results in the plot shown below.



Given a measured value r_D, the confidence interval for p simply is the range from the lower curve to the upper curve at that r_D.

A numerical analysis of the given data results in the 90% confidence limits shown below where *Data* refers to the fraction $\frac{r}{N}$ at every energy.



Energy	$p^{\text{min}}_{0.90}$	$p^{\text{max}}_{0.90}$
0.5	0.0	0.029
1.0	0.014	0.089
1.5	0.137	0.276
2.0	0.493	0.663
2.5	0.861	0.959
3.0	0.98	0.992
3.5	0.99	0.998
4.0	0.994	0.999

Exercise 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

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

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

Solution:

We prove the relation by induction. The relation for $n = 1$ reads

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

and was proven in the lecture. One iteration updates our probability according to the rule

$$\begin{aligned} P_{n\to n+1}(p|N,r) &= \frac{P(r|N,p)\ P_n(p|N,r)}{\int dp'\ P(r|N,p')\ P_n(p|N,r)} \\ &= \frac{\binom{N}{r}p^r(1-p)^{N-r}\cdot\gamma\ p^{nr}(1-p)^{n(N-r)}}{\int dp'\binom{N}{r}p'^r(1-p')^{N-r}\cdot\gamma\ p'^{nr}(1-p')^{n(N-r)}} \\ &= \frac{p^{(n+1)r}(1-p)^{(n+1)(N-r)}}{\int dp' p'^{(n+1)r}(1-p')^{(n+1)(N-r)}} \end{aligned}$$

where γ is a constant factor depending only on N, n and r. The denominator in the last expression is a beta function and can be integrated exactly

$$\int dp'... = \beta((n+1)r+1,(n+1)(N-r)+1) = \frac{((n+1)r)!\ ((n+1)(N-r))!}{((n+1)N+1)!}$$

Putting all together, we see that the relation holds.

$$P_{n\to n+1}(p|N,r) = \frac{((n+1)N+1)!}{((n+1)r)!\ ((n+1)(N-r))!}p^{(n+1)r}(1-p)^{(n+1)(N-r)} \equiv P_{n+1}(p|N,r)$$

The expectation value of the posterior is given by

$$\begin{aligned} \mathbb{E}[p] &= \int_0^1 dp\ p\ P_n(p|N,r) \\ &= \frac{(nN+1)!}{(nr)!(nN-nr)!}\int dp p^{nr+1}(1-p)^{n(N-r)} \\ &= \frac{(nN+1)!}{(nr)!(nN-nr)!}\frac{(nr+1)!(nN-nr)!}{(nN+2)!} \\ &= \frac{nr+1}{nN+2} = \frac{r+\frac{1}{n}}{N+\frac{2}{n}} \to^{n\to\infty,N\neq 0} \frac{r}{N} \end{aligned}$$

which is exactly what we expected since the result for P_1 was shown in the lecture and $P_n(p|N,r) = P_1(nN,nr)$. Making use of this relation, the variance can be written as

$$\mathrm{Var}[p] = \mathbb{E}[p^2] - \mathbb{E}[p]^2 = \frac{(nr+1)(nN-nr+1)}{(nN+2)^2(nN+3)} = \frac{1}{n}\frac{(r+\frac{1}{n})(N-r+\frac{1}{n})}{(N+\frac{2}{n})^2(N+\frac{3}{n})} \to^{n\to\infty,N\neq 0} 0$$

This shows that reusing the data decreases the estimated uncertainty of the result, i.e. one is overconfident in the measurement.

3 Chapter 3: Poisson distribution

Exercise 4: Symmetric exponential distribution

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

Solution:

a) The given function $f(x)$ can be regarded as a probability distribution $P(x)$ for x since the integral $\int_{-\infty}^{\infty} dx\ f(x) = 1$. Evidently the function $f(x)$ is symmetric around 0. The expectation value can therefore be evaluated trivially as

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} dx\ x\ f(x) = 0$$

because the integrand is antisymmetric.

The variance is

$$\sigma^2 = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \int_{-\infty}^{\infty} dx\ x^2 f(x) = 2\int_{0}^{\infty} dx\ x^2\ f(x) = \int_{0}^{\infty} dx\ x^2 e^{-x} = 2$$

The standard deviation therefore is $\sigma = \sqrt{2} \approx 1.41$.

b)

The full width half maximum (FWHM) is the interval for which $f(x) > \frac{1}{2} f_{\max} = \frac{1}{4}$. Solving this condition for $x > 0$ yields

$$f(x) = \frac{1}{2}e^{-x} \stackrel{!}{=} \frac{1}{4} \qquad \Rightarrow \qquad e^{-x} = \frac{1}{2} \qquad \Rightarrow \qquad x = \ln 2 \approx 0,69$$

Making use of symmetry tells us that $x = -\ln 2$ is the other interval border. The FWHM interval therefore is $[-\ln 2, \ln 2]$.

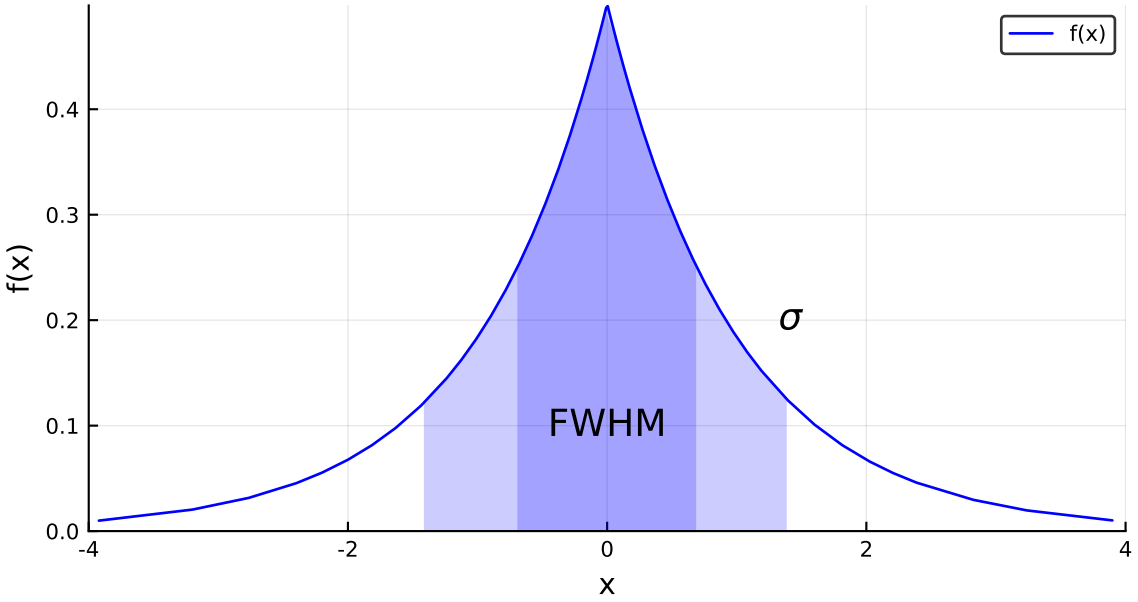
Unlike with the normal distribution, here the FWHM interval is smaller than the $[-\sigma, +\sigma]$ interval.

c)

To determine the probability content contained within the $[-\sigma, +\sigma]$ interval, we split the integral as before

$$\int_{-\sigma}^{\sigma} dx\ f(x) = 2\int_{0}^{\sigma} dx\ \frac{1}{2}e^{-x} = \int_{0}^{\sqrt{2}} dx\ e^{-x} = 1 - e^{-\sqrt{2}} \approx 0,76$$

The distribution with its FWHM and σ interval is plotted below.



Exercise 7: Parameter Inference for the Poisson distribution

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 ν? 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?

Solution:

a)

The Poisson distribution is given by the formula

$$P(n|\nu) = \frac{1}{n!} e^{-\nu} \nu^n$$

where n is the number of observed events, $n_D = 9$ and ν denotes the unknown, true expectation.

Using Bayes' theorem and a flat prior for ν, we can write the probability distribution for ν as

$$P(\nu|n) = \frac{P(n|\nu)P(\nu)}{P(n)} = \frac{1}{n!} e^{-\nu} \nu^n$$

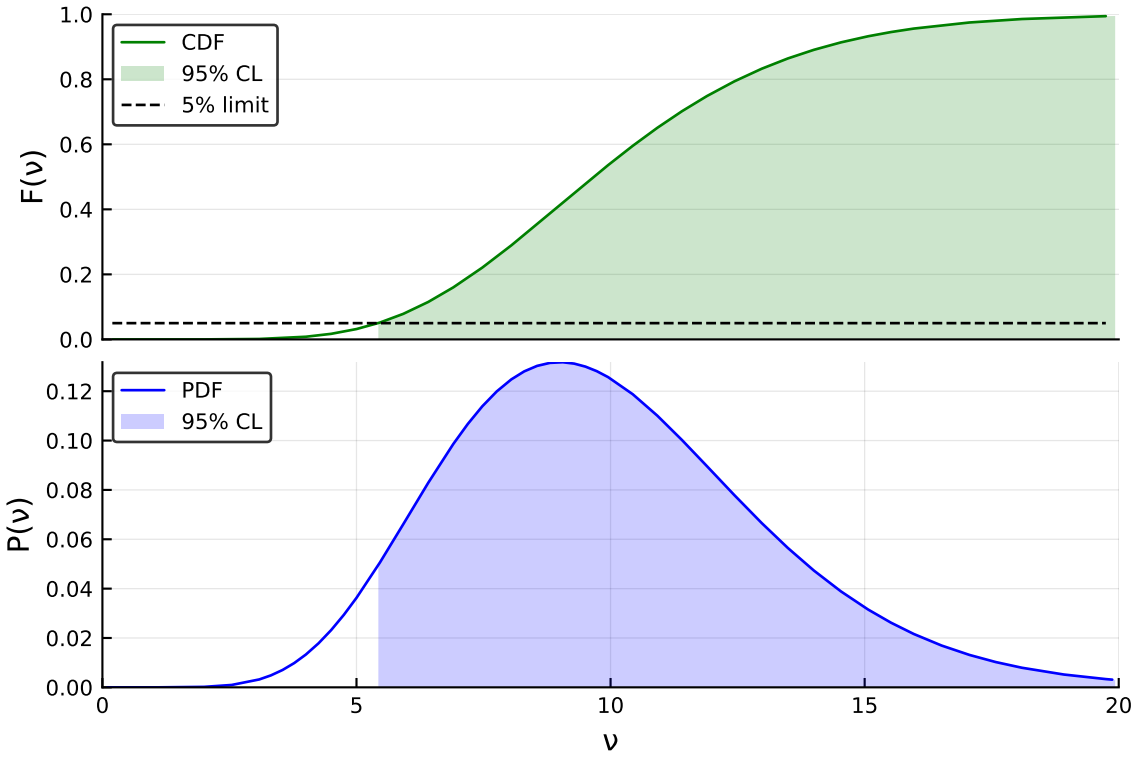
The 95% lower limit l on the posterior can be defined through the condition

$$0.95 \stackrel{!}{=} \int_l^\infty d\nu \; P(\nu) = \frac{1}{9!} \int_l^\infty d\nu \; e^{-\nu} \nu^9 = e^{-\nu} \sum_{i=0}^{9} \frac{\nu^i}{i!} = \frac{\Gamma(10, l)}{9!}$$

where Γ is the incomplete gamma function.

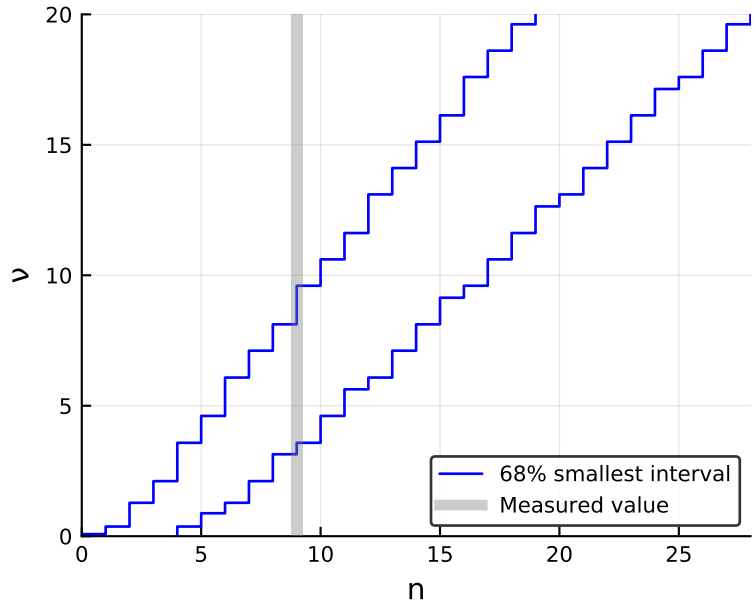
This equation has one solution for $l > 0$ which, evaluated numerically, is $l \approx 5,43$. Therefore the resulting statement is that, assuming a flat prior, the value of ν is larger than 5,43 with 95% probability.

The probability density function (PDF) and cumulative distribution function (CDF) are plotted below. The filled area shows the 95% probability mass above the lower limit. In the top graph, the 5% lower limit is shown as a dashed line.



b)

For the Neyman construction we compute the 68% smallest (maximum density) intervals for all possible values for ν on the computer. The result is plotted below.



The Neyman confidence interval is the range of ν for which the measured value $n_D = 9$ lies within the 68% smallest interval. In the given case this is the interval $[6.4, 13.3]$ which corresponds to the maximum and minimum in the diagram at the measured value.

With this information we can state that for true values of ν between 6.4 and 13.3, the measured value of $n_D = 9$ would lie within the 68% smallest interval.

Exercise 8: Parameter Inference for the Poisson distribution with background

Repeat the previous exercise, assuming you had a known background of 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

Solution:

As in exercise 7, we measure $n_D = 9$ events. This time, the assumed model is the sum of signal with unknown mean ν and background with mean $\lambda = 3.2$.

$$P(n|\nu, \lambda) = \text{Pois}(n|\mu) = \frac{e^{-\mu}\mu^n}{n!}$$

where $\mu = \nu + \lambda$ is the total event expectation.

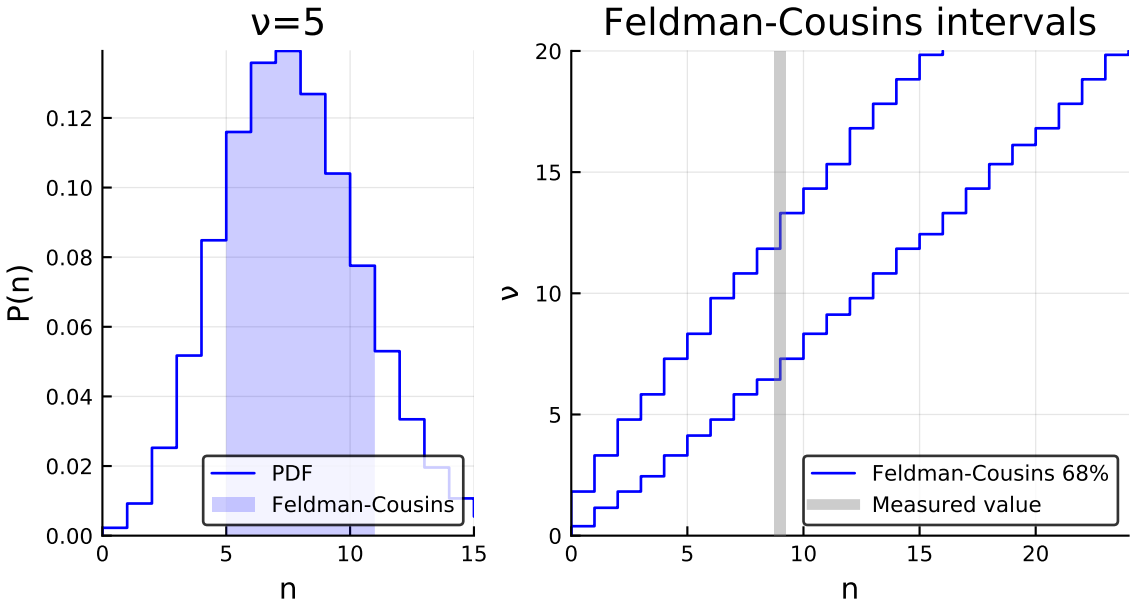
a)

The 68% Feldman-Cousins confidence level interval is similar to the Neyman interval in that the range on the unknown parameter ν is set by adding up 68% confidence intervals in n that contain the measured n_D. However, these intervals are constructed using the metric

$$r = \frac{P(n|\mu)}{P(n|\hat{\mu})}$$

i.e. for a given μ, those values of n are included in the interval for which r is highest until the limit of 68% is reached. Here, $\hat{\mu}$ is the model value that maximizes $P(n|\mu)$ under the constraints set upon μ. In the concrete case, the mode of the Poisson distribution dictates $\hat{\mu} = n$ with the additional constraint that $\hat{\mu} \geq \lambda = 3.2$ because $\nu \geq 0$.

Constructing these intervals on the computer results in the plot shown below.

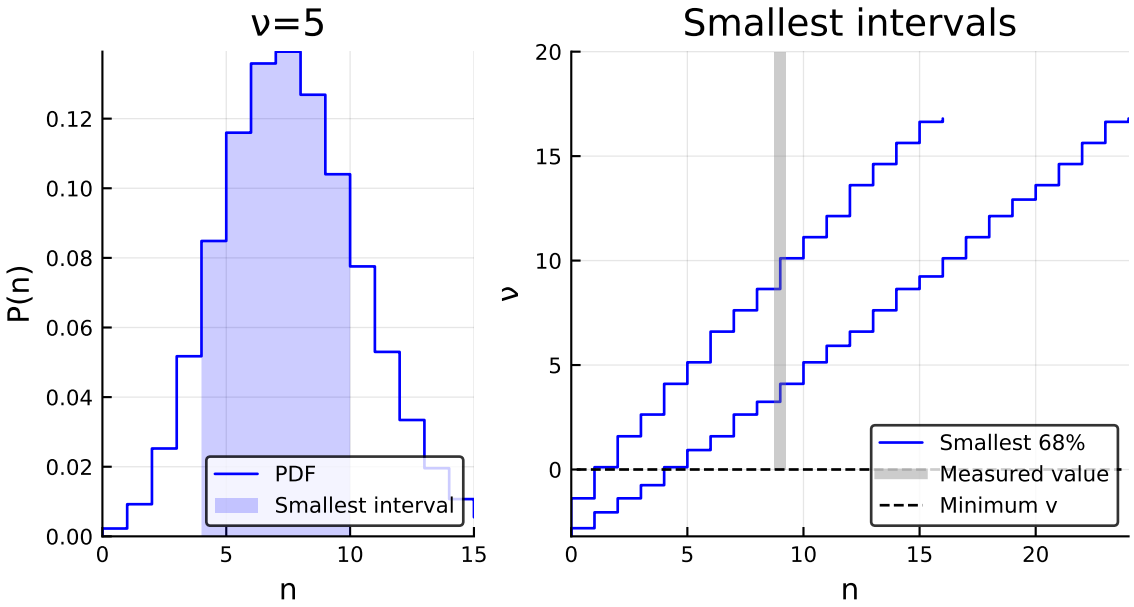


The left plot shows the probability density function (PDF) and the corresponding Fledman-Cousins 68% confidence level interval for $\nu = 5$. It can be seen quite clearly that this interval construction differs from the usual definitions such as central or smallest interval. The right plot shows the intervals in n for a large range of possible ν. For $n = n_D = 9$, the Feldman-Cousins 68% confidence level interval reaches from 3.14 to 9.59.

b)

For the Neyman 68% confidence level interval we can choose one of the other interval definitions. Here, we use the smallest (highest density) interval.

We construct the usual Neyman intervals for $\mu = \nu + \lambda$ on the computer. The results are shown below.



Compared to the Feldman-Cousins interval, the smallest interval has more mass at lower n. This leads to intervals on ν that have slightly higher values. Unlike Feldman-Cousins', The Neyman construction can produce empty or negative intervals at low values of n_D. For $n_D = 9$, the Neyman interval spans from 3.24 to 10.1.

c)

For the Bayesian analysis, we use a flat prior $P(\nu)$. Then the posterior probability distributions for ν is

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

and the corresponding cumulative distribution function is

$$F(\nu|n,\lambda) = 1 - \frac{e^{-\nu}\sum_{i=0}^{n}\frac{(\lambda+\nu)^i}{i!}}{\sum_{i=0}^{n}\frac{\lambda^i}{i!}}$$

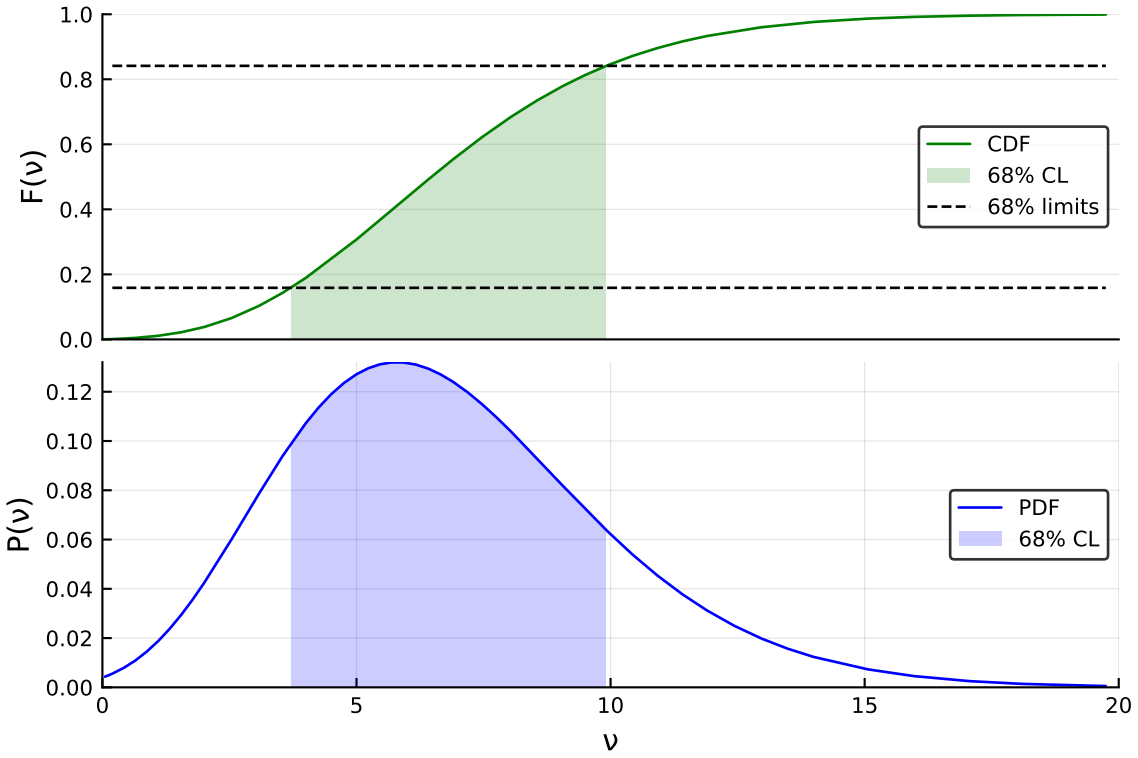
The mode of the distribution occurs at $\nu* = \max(0, n-\lambda) = 5.8$.

We find the $(1-\alpha) = 68\%$ central interval of the posterior by demanding

$$F(\mu_1|n,\lambda) \stackrel{!}{=} \frac{\alpha}{2} \qquad F(\mu_2|n,\lambda) \stackrel{!}{=} 1 - \frac{\alpha}{2}$$

Solving these two conditions numerically yields $\mu_1 = 3.71$ and $\mu_2 = 9.91$.

The posterior probability density function (PDF) and its cumulative distribution function (CDF) are shown below. The CDF plot also includes the $\alpha/2$ and $1-\alpha/2$ limits.



Exercise 13: Binned probability density

In this problem, we look at the relationship etween 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

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

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 $n_j = 1$ events. 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.

Solution:

We are given the probability density $f(x|\lambda)$ with

$$\int_a^b dx\ f(x|\lambda) = 1 \qquad f(x|\lambda) = 0\ \forall x \notin [a, b]$$

and we observe n events with values x_i which are distributed according to f.

Assuming the events are independent, we can write the likelihood of the data

$$\mathcal{L}(\lambda) = P(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\lambda)$$

Now we split the interval $[a, b]$ into K bins Δ_j, $j \in 1...K$ with central values x_j and measured event counts n_j. The probability content in each bin evidently is

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

and the binned likelihood becomes

$$\mathcal{L}_B(\lambda) = \prod_{j=1}^{K} \text{Pois}(n_j|\nu_j) = \prod_{j=1}^{K} \frac{e^{-\nu_j}\nu_j^{n_j}}{n_j!}$$

where we have assumed a Poisson distribution for each bin.

Taking $K \to \infty$, i.e. making the bins infinitesimally small, will only leave the possible contents $n_j = 0$ and $n_j = 1$ because x is continuous and no two events will have the exact same value. Also the expectation for one bin can be rewritten

$$\lim_{K\to\infty} \nu_j = f(x_j|\lambda)\Delta$$

where Δ is the width of the interval.

Now we can expand the likelihood

$$\lim_{K\to\infty} \mathcal{L}_B(\lambda) = \lim_{K\to\infty} \prod_{j=1}^{K} \frac{e^{-\nu_j}\nu_j^{n_j}}{n_j!} = \lim_{K\to\infty} \prod_{j=1}^{K} e^{-\nu_j} \prod_{\{j:n_j=0\}} \frac{\nu_j^0}{0!} \prod_{\{j:n_j=1\}} \frac{\nu_j^1}{1!}$$

The second product gives 1 and inserting ν_j into the third product yields

$$\prod_{\{j:n_j=1\}} \nu_j = \prod_{i=1}^{n} f(x_i|\lambda)\Delta = \mathcal{L}(\lambda)\Delta$$

because only those n_j contribute where one value $x_i \in \Delta_j$ was measured.

The first product can be pulled into the exponential function as a sum

$$\lim_{K\to\infty} \prod_{j=1}^{K} e^{-\nu_j} = \lim_{K\to\infty} \exp\left(-\sum_{j=1}^{K} \nu_j\right) = \exp\left(-\int_a^b dx\ f(x|\lambda)\right) = \frac{1}{e}$$

where the sum of all intervals equals the complete interval $[a, b]$. Combining the results of the three products leads to

$$\lim_{K\to\infty} \mathcal{L}_B(\lambda) = \frac{\Delta}{e}\mathcal{L}(\lambda)$$

We see that the binning process approaches the original unbinned likelihood (up to a known constant) for infinitesimally small bins if each bin can be described by a Poisson process.

Exercise 16: Thinned Poisson process

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

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

Solution:

Intuitively, this result is obvious. We could just count the events n for which $X_n = 1$, assume a Poisson distribution and immediately arrive at the last formula.

To show this result mathematically we first find the individual distributions of N and $X|N$. We know that N is distributed according to a Poisson distribution with mean ν

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

For a given N, $X = \sum_{n=1}^{N} X_n$ measures the number of successes r is a binomial distribution with N trials. This can be seen by splitting up the sum

$$X = \sum_{n=1}^{N} X_n = \sum_{n=1}^{r} 1 + \sum_{n=1}^{N-r} 0 = r$$

where we count $X_n = 1$ as a success. The distribution of X is therefore a binomial

$$P(X|N) = \binom{N}{X} p^X (1-p)^{N-X}$$

Now we can determine the probability distribution for X by marginalizing out N

$$\begin{aligned}
P(X) &= \sum_{N=X}^{\infty} P(X|N)P(N) \\
&= \frac{e^{-\nu}p^X}{X!} \sum_{N=X}^{\infty} \frac{(1-p)^{N-X}\nu^N}{(N-X)!} \\
&= \frac{e^{-\nu}p^X\nu^X}{X!} \sum_{H=0}^{\infty} \frac{((1-p)\nu)^H}{H!} \\
&= \frac{e^{-\nu}p^X\nu^X}{X!}\nu^X e^{(1-p)\nu} \\
&= \frac{e^{(p\nu)}(p\nu)^X}{X!} \\
&= \mathrm{Pois}(X|p\nu)
\end{aligned}$$

where we have substituted $H = N - X$ in the third step. The sum is then the series expansion of the exponential $e^{(1-p)\nu}$. The original sum goes from X to ∞ because for $N \leq X$ the binomial distribution vanishes.

As expected X is also Poisson-distributed with mean $p\nu$ instead of ν. This property of the Poisson distribution allows us to simply count the end result and not worry too much about events that were lost due to detector inefficiencies or similar. The result is still a Poisson distribution.

4 Chapter 4: Gaussian probability distribution function

Exercise $\sqrt{-1}$: Central Limit Theorem for high-frequency distributions

This exercise is the result of a conversation between Allen and me about the applicability of the CLT in the presence of high-frequencies.

In the lecture, the Central Limit Theorem (CLT) was derived for probability distributions $P(x)$ for which all moments are finite. In this exercise we will look at the convergence behavior of probability densities which contain large high-frequency components.

(a) We take n samples from a centered and normalized probability density $P(x)$ which contains large angular frequency components ω. Show that the mean of the samples $\bar{x} \equiv \frac{1}{n}\sum x_i$ contains these components shifted to $n \cdot \omega$. Find a criterion of $P(x)$ which determines if these frequencies vanish for large n. In the case where they don't, show that the limiting normal distribution is only approximated in the limit $n \to \infty$ under the metric $|f(x) - g(x)|$.

(b) Consider the probability density $P(x) = \frac{\delta(x-a)+\delta(x+a)}{2}$ with $a > 0$. Show that the CLT is applicable but a normal distribution is not approximated for any finite number of samples. Show that the mean of n samples is distributed according to a discrete binomial distribution.

Solution:

(a)

We are given a probability distribution with mean $\mu = 0$ for which all moments exist. The CLT is therefore applicable. The characteristic function of the mean of n samples $\bar{x} \equiv \frac{1}{n}\sum x_i$ can be calculated by multiplying the characteristic functions of $P(x_i) = P(x)$.

$$\phi_{\bar{x}}(\omega) = \prod_{i=1}^{n} \phi_{x_i/n}(\omega)$$

The characteristic function for one x_i is

$$\phi_{x_i/n}(\omega) = \int_{-\infty}^{\infty} dx\ e^{\frac{i\omega x}{n}} P(x)$$

This looks very similar to a Fourier transform, in fact if we take

$$\phi_{x_i/n}(n\omega) = \int_{-\infty}^{\infty} dx\ e^{i\omega x} P(x)$$

we get exactly the FT of $P(x)$. A strong frequency component ω of $P(x)$ is now seen at $n\omega$. Evidently, the frequency gets shifted. To determine the amplitude of this frequency, we calculate the combined characteristic function which becomes

$$\phi_{\bar{x}}(\omega) = \left[\int_{-\infty}^{\infty} dx\ e^{\frac{i\omega x}{n}} P(x)\right]^n$$

For large n, all frequency components $|\phi_{\bar{x}}(n\omega)| < 1$ tend towards 0.

Note also, that $|\phi| \leq 1$ for all normalized probability densities $P(x) \geq 0$. This fact becomes obvious when considering an example. Let $P(x) = c + A\sin(\omega_s x)$ on the interval $[-1, 1]$. If we choose ω_s such that the sine term vanishes at the boundaries, it does not contribute to the overall probability content and we get $c = \frac{1}{2}$ from normalization. To enforce $P(x) \geq 0$, we must choose $A \leq \frac{1}{2}$. In the extremal case of $A = \frac{1}{2}$, we get a frequency component of

$$|\phi(\omega_s)| = |\int_{-1}^{1} dx\ e^{i\omega_s x} \frac{1}{2}\sin(\omega_s x)| = 1.$$

This restriction enforces, that the amplitude does not diverge for $n \to \infty$. But what about the limiting case of $\phi(\omega) = 1$? Here, the amplitude of the Fourier component stays constant, even for very large n. Viewed by the metric $|f(x) - g(x)|$, the distribution $P(\bar{x})$ does not approximate a Gaussian for any finite n. The high-frequency wiggles that are superimposed, get shifted to higher and higher frequencies and in the limiting case $n = \infty$ have been thrown out to infinity and vanish. But this only happens for frequencies that are as strong as they can possibly be. There is a very intuitive explanation for this behavior which will become clear in the next section.

(b)

We are given the probability density $P(x) = \frac{\delta(x-a)+\delta(x+a)}{2}$. This density corresponds to getting either outcome a or $-a$ with 50% probability. All moments of the function exist and can be calculated to be

$$M_n[x] = \int dx\ x^n \frac{1}{2}(\delta(x-a) + \delta(x+a)) = \frac{1}{2}(a^n + (-a)^n).$$

For the mean and the standard deviation, this gives us $\mu = 0$ and $\sigma = a$. Therefore, the CLT is applicable. We now take n samples from that probability and calculate the mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. To get the distribution of the mean, we use characteristic functions as before. For one x_i

$$\phi_{x_i/n}(\omega) = \int_{-\infty}^{\infty} dx\ e^{i\frac{\omega x}{n}} \frac{1}{2}(\delta(x-a) + \delta(x+a)) = \frac{1}{2}(e^{i\frac{\omega a}{n}} + e^{-i\frac{\omega a}{n}}) = \cos\left(\frac{\omega a}{n}\right)$$

and for $\bar{x}$

$$\phi_{\bar{x}}(\omega) = \cos\left(\frac{\omega a}{n}\right)^n.$$

The approximation for large n in the CLT gives us

$$\phi_{x_i/n}(\omega) = 1 - \frac{\omega^2 a^2}{2n^2}$$

$$\phi_{\bar{x}}(\omega) = e^{-\frac{\omega^2 a^2}{2n}}$$

instead. In the limit $n \to \infty$, these results are the same as all maximum of the cosine, except for the one at zero, are pushed out to infinity.

To get the distribution $P(\bar{x})$, we need to invert the Fourier transform.

$$P(\bar{x}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\ e^{-i\omega x}\phi_{\bar{x}}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\ e^{-i\omega x}\frac{1}{2^n}\left(e^{i\frac{\omega a}{n}} + e^{-i\frac{\omega a}{n}}\right)^n$$

Here, we have written the cosine term as the sum of two exponentials. This allows us to take the sum to the power of n. The result of that operation is a large sum of exponential products where each term appears with a frequency given by combinatorics. The terms containing $e^{\pm i\frac{\omega a}{n}}$ n times appear only once while the mixed terms appear $\binom{n}{k}$ times, where k is the number of $e^{i\frac{\omega a}{n}}$ in that term. Therefore

$$P(\bar{x}) = \frac{1}{2\pi\ 2^n}\int_{-\infty}^{\infty} d\omega\ e^{-i\omega x}\left(\sum_{k=0}^{n}\binom{n}{k}e^{i(2k-n)\frac{\omega a}{n}}\right) = \frac{1}{2\pi\ 2^n}\int_{-\infty}^{\infty} d\omega\left(\sum_{k=0}^{n}\binom{n}{k}e^{i\omega(\frac{(2k-n)a}{n}-x)}\right)$$

Each term in the sum now produces a delta distribution $2\pi\delta\left(x - \frac{(2k-n)a}{n}\right)$ when integrated. The distribution of $\bar{x}$ therefore is

$$P(\bar{x}) = \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}\ \delta\left(x - \frac{(2k-n)a}{n}\right)$$

which is a binomial distribution with success rate 50% as expected. This distribution is made up of $n+1$ delta spikes, each representing one possible value of the mean. For any finite n, the distribution $P(x) = 0$ at almost all x, only in the limit $n \to \infty$ does it become a true binomial which by then has turned into a normal distribution.

Exercise 8: Applicability of the central limit theorem

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.

Solution:

(a)

We take n samples from the exponential distribution $p(x) = \lambda e^{-\lambda x}$ and calculate the mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The CLT states that for $n \to \infty$, $\bar{x}$ is normally distributed if all moments of $p(x)$ are finite. This is the case as all arising integrals of the form

$$M_n[x] = \int_0^\infty dx\ x^n p(x) = \lambda \int_0^\infty dx\ x^n e^{-\lambda x}$$

converge. The limiting normal distribution will have the same mean as $p(x)$ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$.

Mean value of $p(x)$ is

$$\mu \equiv \mathbb{E}[x] = \int_0^\infty dx\ x\ p(x) = \int_0^\infty dx\ x\lambda e^{-\lambda x} \tag{1}$$

$$=^{\text{p.i.}} \left[-x\ e^{-\lambda x}\right]_0^\infty - \int_0^\infty dx\ \ -e^{-\lambda x} = \lim_{x\to\infty}(x\ e^{-\lambda x}) - \left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda}. \tag{2}$$

Similarly, the variance can be calculated to be

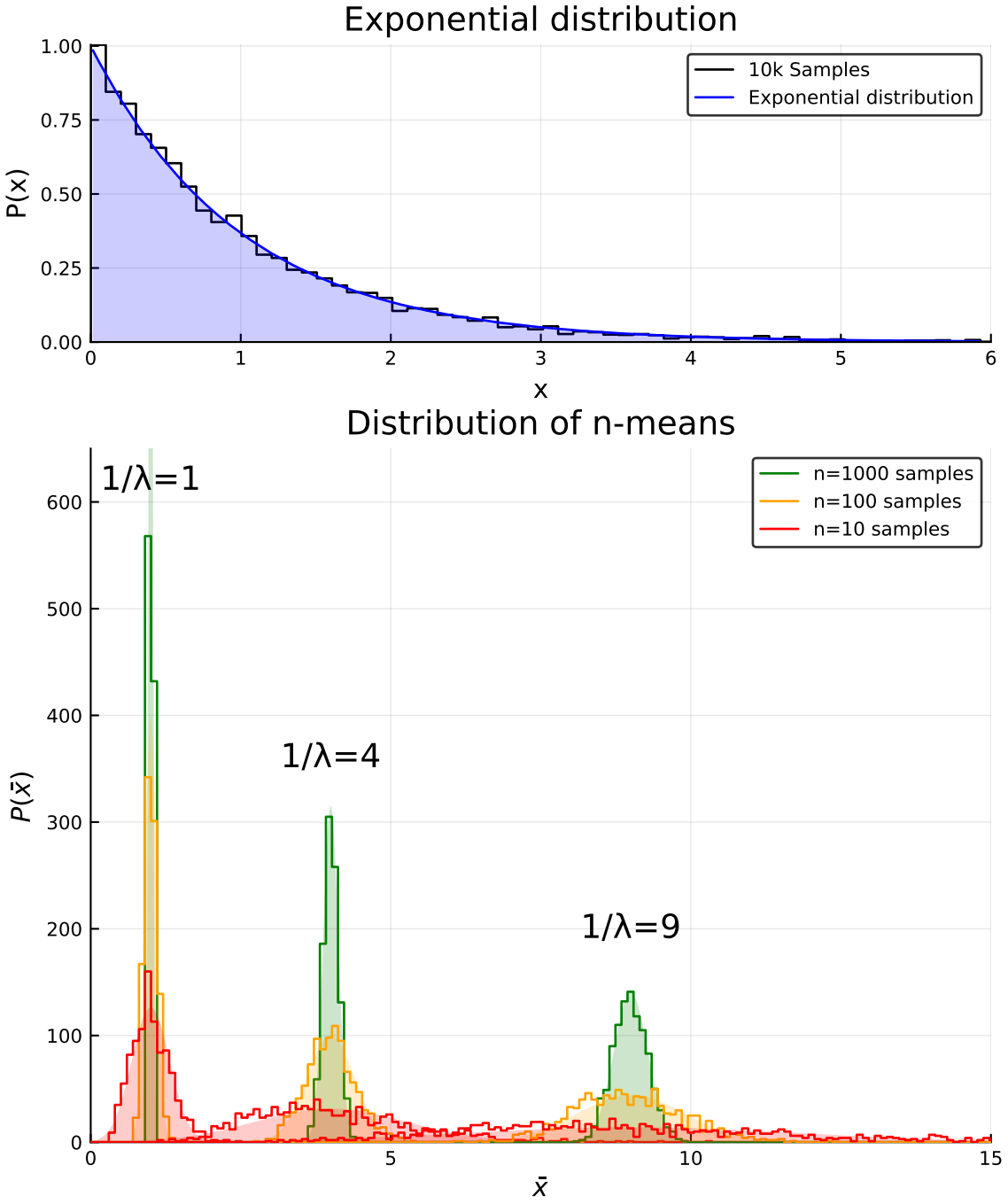
$$V[x] = \sigma^2 = \mathbb{E}[x^2] - E[x]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

and standard deviation is $\sigma = \frac{1}{\lambda}$.

Therefore the distribution of the mean of n samples approximately is a normal distribution $\mathcal{N}\left(x|\mu = \frac{1}{\lambda}, \sigma = \frac{1}{\lambda\sqrt{n}}\right)$. Performing the experiment many times will approximately reproduce this distribution.

Performing the experiment has been simulated on the computer. For the values of $\lambda \in \{1, \frac{1}{4}, \frac{1}{9}\}$ and $n \in \{10, 100, 1000\}$.

The upper plot shows the PDF of the exponential distribution and a random experiment with $n = 10000$. Below, the distribution of mean values of n samples are shown. The experiment was performed 1000 for each configuration of n and λ. The solid lines indicate the measured data while the transparent fills are the normal distributions predicted from the CLT.



(b)

For the CLT to hold, all moments must have a finite value. For the Cauchy distribution, the moments are

$$M_n[x] = \int_{-\infty}^{\infty} dx\ x^n \frac{1}{\pi\gamma} \frac{\gamma^2}{(x - x_0)^2 + \gamma^2} \to^{n>0} \pm\infty$$

All moments $n > 0$ diverge because the tails don't fall to zero quickly enough. Even the expectation

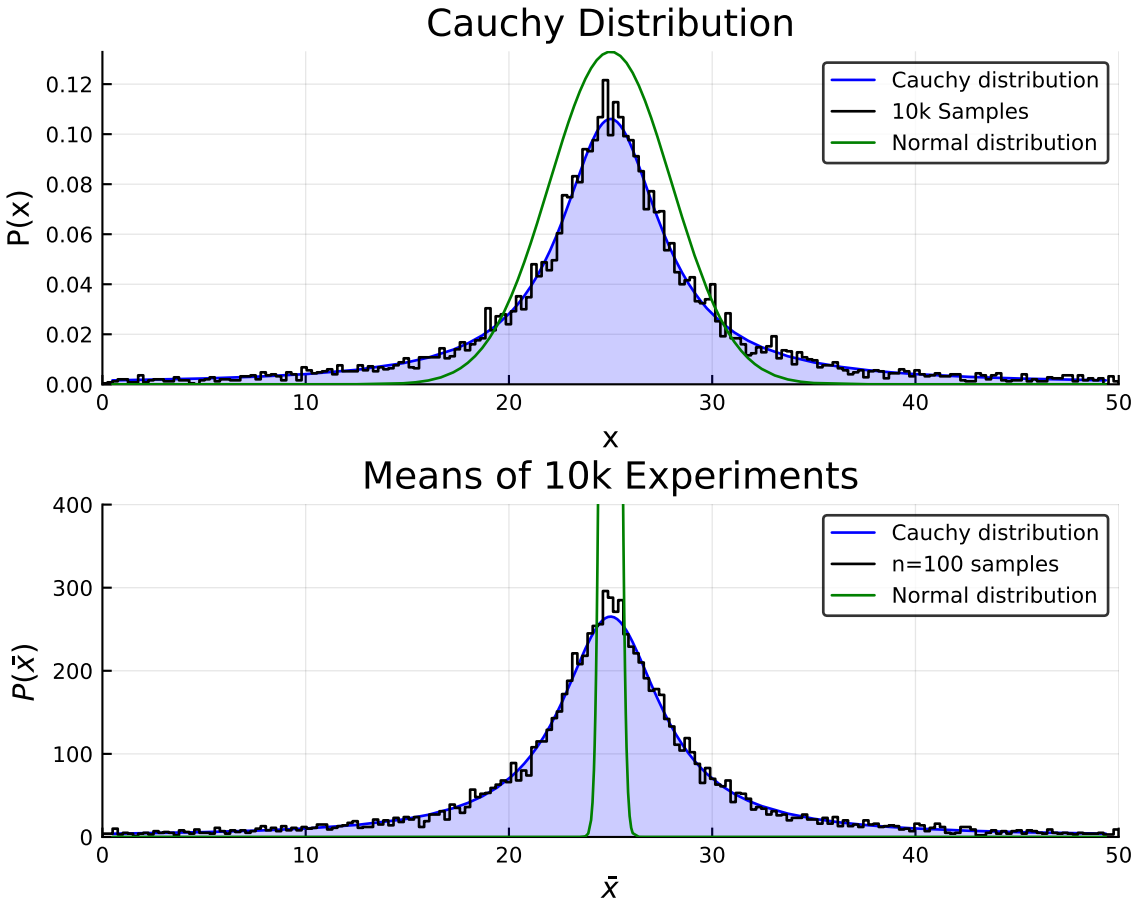
value

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} dx\; x \frac{1}{\gamma} \frac{\gamma^2}{(x - x_0)^2 + \gamma^2} = \gamma \int dx\; \frac{x}{x^2 + \gamma^2}$$

cannot be evaluated using the standard integral definition, despite the function being symmetric around x_0.

The Cauchy distribution with $x_0 = 25$ and $\gamma = 3$ is plotted below, in the upper diagram. The black curve shows the distribution of 10000 values sampled from the distribution. The green curve is a normal distribution $\mathcal{N}(x|\mu = 25, \sigma = 3)$.

The lower graph shows the distribution of the mean values of $n = 100$ samples each. The result is again the same Cauchy function. The green curve shows the expected result if we had sampled from a normal distribution instead. Unlike with the normal distribution, taking more samples does not decrease the variance of mean values with the Cauchy distribution.



Exercise 11: Contours of the bivariate normal distribution

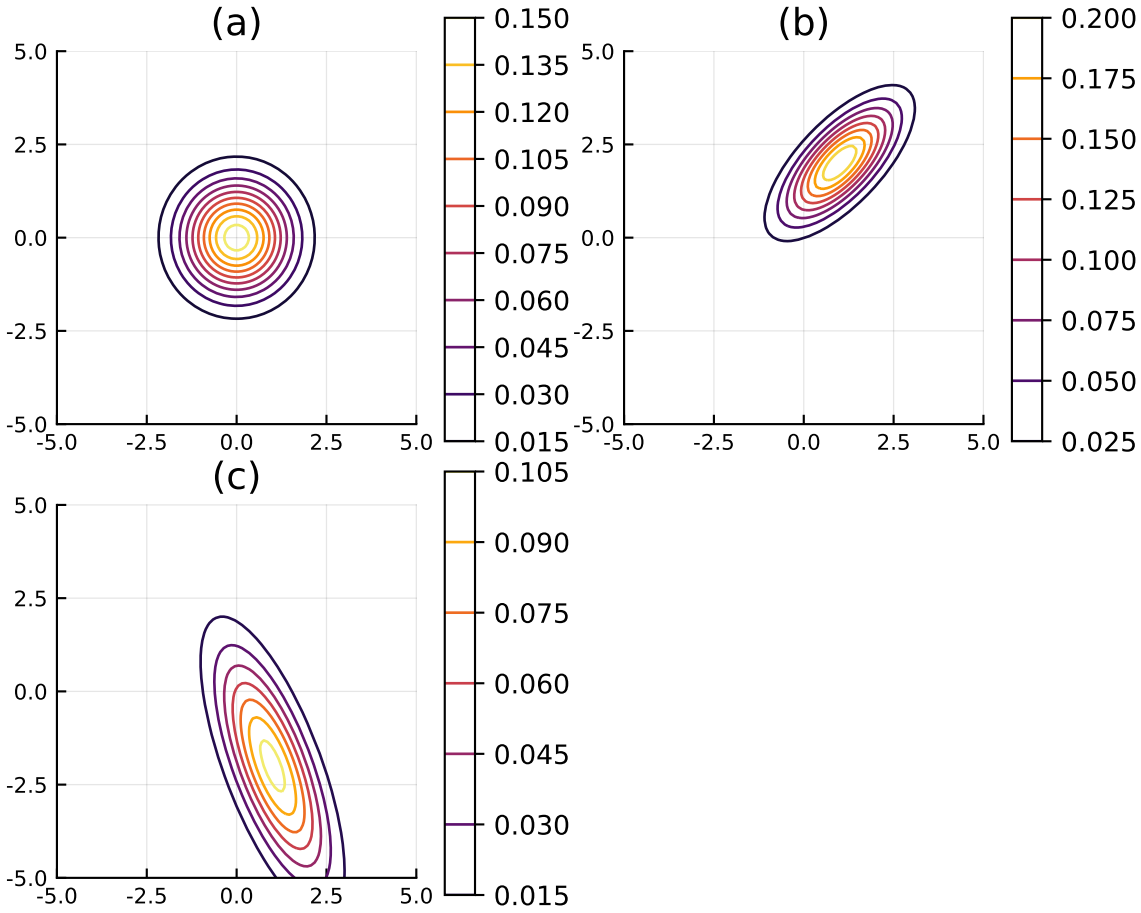
With a plotting program, draw contours of the bivariate Gauss function for the following parameters
(a) $\mu_x = 0, \mu_y = 0, \sigma_x = 1, \sigma_y = 1, \rho_{xy} = 0$
(b) $\mu_x = 1, \mu_y = 2, \sigma_x = 1, \sigma_y = 1, \rho_{xy} = 0.7$
(c) $\mu_x = 1, \mu_y = -2, \sigma_x = 1, \sigma_y = 2, \rho_{xy} = -0.7$

Solution:
The density function of the bivariate normal distribution is

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

The mean value of the distribution is located at $\vec{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$ and the covariance matrix is $\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\ \sigma_x\sigma_y \\ \rho\ \sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$.

With this information, we can plot the contour lines on the computer. The result is shown below. In the contour diagrams, x is plotted to the right and y upwards.



Exercise 12: Bivariate normal distribution

(a) Show that the PDF of the bivariate normal distribution can be written in the form

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

(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

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

Solution:

(a)

The correlation coefficient $\rho \in [-1, 1]$ of two variables x and y is defined as

$$\rho_{xy} = \frac{\mathrm{Cov}[x,y]}{\sigma_x\sigma_y}$$

where σ_x, σ_y denote the standard deviations of the distributions of x and y. The general form of a multivariate normal distribution is defined by the mean $\vec{\mu} = (\mu_1, \mu_2, ...)$ and the covariance matrix $\Sigma_{i,j} = \mathrm{Cov}[x_i, x_j]$ and has the form

$$\mathcal{N}(\vec{x}|\vec{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right).$$

For the bivariate case, Σ takes the form

$$\Sigma = \begin{pmatrix} \mathrm{Cov}[x,x] & \mathrm{Cov}[x,y] \\ \mathrm{Cov}[y,x] & \mathrm{Cov}[y,y] \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

and its inverse is

$$\Sigma^{-1} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)} \begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}$$

where $\sigma_x^2\sigma_y^2(1-\rho^2)$ is the determinant of Σ. Inserting this into the multivariate normal distribution yields

$$\begin{aligned}\mathcal{N}(x,y|\mu_x,\mu_y,\sigma_x,\sigma_y,\rho) = \frac{1}{\sqrt{2\pi\sigma_x^2\sigma_y^2(1-\rho^2)}} \exp\Bigg(-\frac{1}{2}(x-\mu_x \quad y-\mu_y)\frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)} \\ \begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}\Bigg) \\ = \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_x\sigma_y} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right)\end{aligned}$$

which reduces to the given formula for $\mu_x = \mu_y = 0$.

(b)

The distribution of $z = x - y$ can be expanded as

$$P(z) = \int dx\ dy\ \delta(z-x+y)\ P(x,y)$$

where $P(x,y)$ is the joint distribution for x and y and takes the form of a bivariate normal distribution, like in exercise (a). Inserting $P(x,y)$, we get

$$P(z) = \int dx \int dy\ \delta(z-x+y)\ \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho xy}{\sigma_x\sigma_y}\right)\right)$$

We can now perform the trivial integration over dy by substituting $y = x - z$ and thereby eliminating the δ distribution.

$$\begin{aligned}P(z) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \int dx \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2} + \frac{(x-z)^2}{\sigma_y^2} - \frac{2\rho x(x-z)}{\sigma_x\sigma_y}\right)\right) \\ = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \int dx \exp\left(-\frac{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}{2(1-\rho^2)\sigma_x^2\sigma_y^2}\left(x^2 + 2x\frac{z\sigma_x(\rho\sigma_y-\sigma_x)}{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}\right) - \frac{z^2}{2(1-\rho^2)\sigma_y^2}\right) \\ = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}} \exp\left(-\frac{1}{2(1-\rho^2)\sigma_x^2\sigma_y^2}\frac{z^2\sigma_x^2(\rho\sigma_y-\sigma_x)^2}{\sigma_y^2+\sigma_x^2-2\rho\sigma_x\sigma_y} - \frac{z^2}{2(1-\rho^2)\sigma_y^2}\right)\end{aligned}$$

In the last step, the x integration has been performed where $\int_{-\infty}^{\infty} dx\ \exp(a(x-\mu)^2) = \frac{\sqrt{\pi}}{\sqrt{a}}$. Reordering the argument of the exponential brings us to the expected form

$$P(z) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}} \exp\left(-\frac{1}{2}\frac{z^2}{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}\right)$$

Exercise 13: Convolution of Gaussians

Suppose you have a true distribution which follows a Gaussian distribution:

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

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

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

What is the predicted distribution for the observed quantity y?

Solution:

Using the law of total probability, we can express the probability distribution of y as

$$P(y) = \int dx\ P(y|x)\ P(x)$$

Inserting the the known distributions

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

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

yields

$$P(y) = \frac{1}{\sqrt{2\pi}^2\sigma_x\sigma_y} \int_{-\infty}^{\infty} dx\ \exp\left[-\frac{1}{2}\left(\left(\frac{x-x_0}{\sigma_x}\right)^2 + \left(\frac{y-x}{\sigma_y}\right)^2\right)\right]$$

Completing the quadrature for x allows us to perform the x integration

$$P(y) = \frac{1}{\sqrt{2\pi}^2\sigma_x\sigma_y} \int_{-\infty}^{\infty} dx\ \exp\left[-\frac{1}{2}\left(\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}\left(x - \frac{\frac{x_0}{\sigma_x^2}+\frac{y}{\sigma_y^2}}{\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}}\right)^2 + \frac{x_0^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} + \frac{(\frac{x_0}{\sigma_x^2}+\frac{y}{\sigma_y^2})^2}{\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}}\right)\right]$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_x^2+\sigma_y^2}} \exp\left[-\frac{1}{2}\left(\frac{y-x_0}{\sqrt{\sigma_x^2+\sigma_y^2}}\right)^2\right]$$

where the integral over x produced the factor $\sqrt{2\pi}\frac{\sigma_x\sigma_y}{\sqrt{\sigma_x^2+\sigma_y^2}}$.

We see that the convolution of two normal distributions is again a normal distribution.

Exercise 14: Bayesian analysis of measured cross section data

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

Angle θ	Cross secion	Error
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?

Solution:

(a)

We are given five measurements of the reaction cross section $\sigma(\theta)$ and want to infer the model parameters A, B, C using a Bayesian analysis. We assume an error term that follows a normal distribution with standard deviation $\varepsilon(\theta)$. The assumed model then is

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

where $r(\theta) \sim \mathcal{N}(0, \varepsilon(\theta))$ represents the random noise term.

Using Bayes' theroem, we can express the probability of our model as

$$P(A, B, C|\{\sigma_i\}, \{\theta_i\}, \{\varepsilon_i\}) = \frac{P(\{\sigma_i\}|A, B, C, \{\theta_i\}, \{\varepsilon_i\})\ P(A, B, C)}{\int dA\ dB\ dC\ P(\{\sigma_i\}|A, B, C, \{\theta_i\}, \{\varepsilon_i\})\ P(A, B, C)}$$

We assume flat priors for A, B, C so the priors $P(A, B, C)$ drop out of the equation. To do this in a mathematically correct way, we take finite interval boundaries $A^{\pm}, B^{\pm}, C^{\pm}$ and let them go to infinity after integration.

Using our model above and that our data are i.i.d, we can express the likelihood as

$$P(\{\sigma_i\}|A, B, C, \{\theta_i\}, \{\varepsilon_i\}) = \prod_i P(\sigma_i|A, B, C, \theta_i, \varepsilon_i) = \prod_i \mathcal{N}(\sigma_i|A + B\cos(\theta_i) + C\cos(\theta_i^2), \varepsilon_i)$$

For normalization reasons, the evidence must be

$$\int dA\ dB\ dC\ \prod_i \mathcal{N}(\sigma_i|A, B, C, \theta_i, \varepsilon_i)\ P(A, B, C) = P(A, B, C).$$

This argument only holds as long as there are at least as many data points as model variables. Else, the integral diverges to $+\infty$ because all normal distributions have been integrated to a positive finite value before the last integration is performed.

The posterior probability density can now be written as

$$P(A, B, C|\{\sigma_i\}, \{\theta_i\}, \{\varepsilon_i\}) = \prod_i \mathcal{N}(\sigma_i|A + B\cos(\theta_i) + C\cos(\theta_i^2), \varepsilon_i)$$

(b)

To determine the mode of the distribution, we find the maximum by varying the values of A, B, C

$$\max_{A,B,C} P(A, B, C|\{\sigma_i\}, \{\theta_i\}, \{\varepsilon_i\}) = \max_{A,B,C} \prod_i \frac{1}{\sqrt{2\pi}\varepsilon_i} \exp\left[-\frac{1}{2}\left(\frac{\sigma_i - (A + B\cos(\theta_i) + C\cos(\theta_i^2))}{\varepsilon_i}\right)^2\right]$$

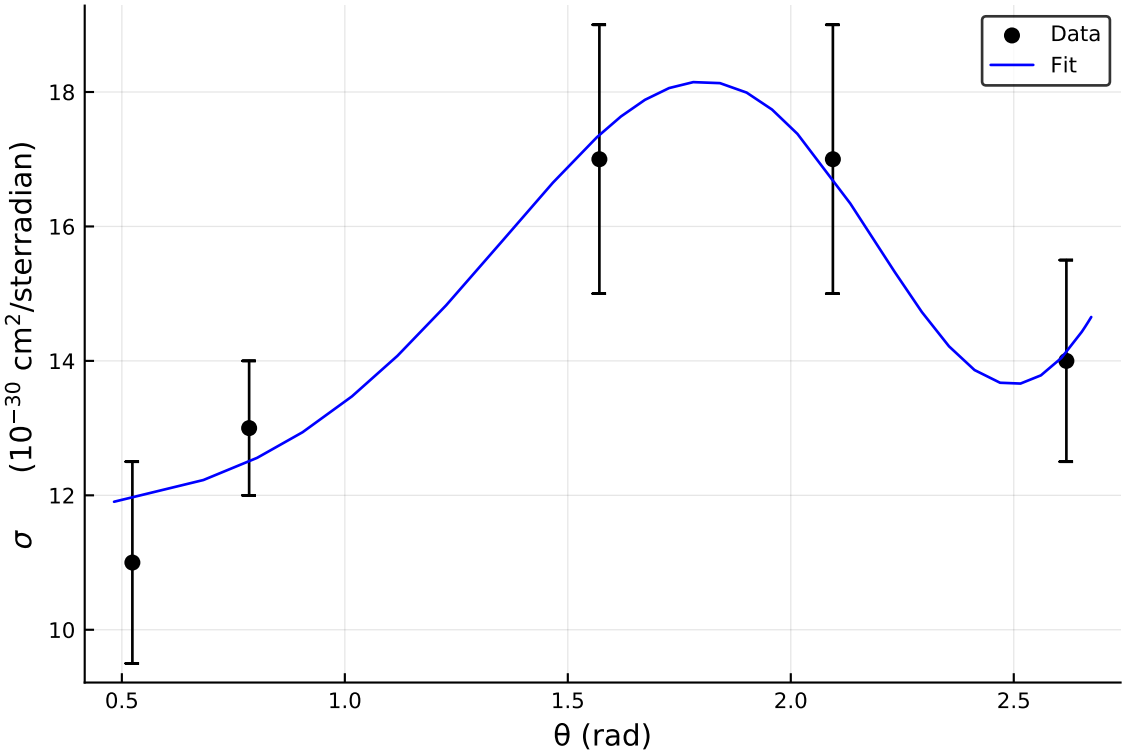
Since ε_i is constant, it suffices to find a maximum of the product of exponentials and we can pull the product into the exponential function.

$$\max_{A,B,C} \text{ const} \cdot \exp\left[-\frac{1}{2}\sum_i \left(\frac{\sigma_i - (A + B\cos(\theta_i) + C\cos(\theta_i^2))}{\varepsilon_i}\right)^2\right]$$

Because the argument of the exponential funcion is always negative, it sufices to find a minimum of the sum

$$\min_{A,B,C} \sum_i \left(\frac{\sigma_i - (A + B\cos(\theta_i) + C\cos(\theta_i^2))}{\varepsilon_i}\right)^2$$

This method is also called *least squared error* and is implemented in many computer algebra systems. A numerical evaluation found the best values for the model to be $A = 15.4, B = -1.08, C = -2.57$. The data and fitted model are plotted below.



5 Chapter 5: Model-fitting and model selection

Exercise 1: Bayesian and Frequentist inference

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

Energy	Trials	Successes
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).

Solution:

(a)

We fit the Sigmoid function $\sigma(E|A, E_0) = \frac{1}{1+e^{-A(E-E_0)}}$ to the given data $\mathbb{D} = \{(E_i, n_i, r_i)\}_{i=1}^{8}$. The likelihood of our data given the model is

$$P(\mathbb{D}|A, E_0) = \prod_{i=1}^{8} \binom{n_i}{r_i} \sigma(E_i|A, E_0)^{r_i} (1 - \sigma(E_i|A, E_0))^{n_i - r_i}.$$

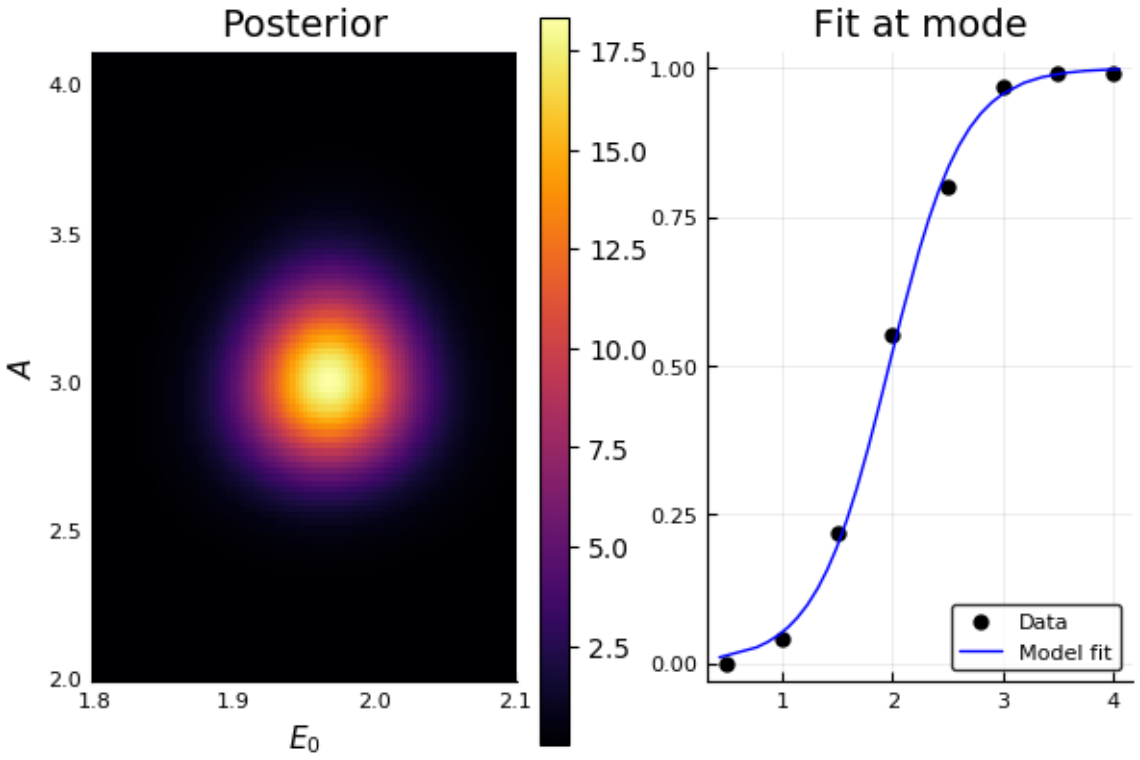
To get the posterior probability density, we use Bayes' theorem and write

$$P(A, E_0|\mathbb{D}) = \frac{P(\mathbb{D}|A, E_0)\ P(A, E_0)}{P(\mathbb{D})}$$

where $P(A, E_0)$ is our prior information about the model parameters. As no prior information is given, we choose constant priors $P(A, E_0) = \text{const}$ and arrive at

$$P(A, E_0|\mathbb{D}) = \frac{P(\mathbb{D}|A, E_0)}{\int dA \int dE_0 P(\mathbb{D}|A, E_0)}$$

The denominator is a normalization constant, ensuring that the overall probability is conserved. This posterior density can be evaluated numerically on a grid spanning all possible A and E_0. The posterior probability is plotted in the E_0, A plane below. Its mode is located at $E_0 = 1.97$ and $A = 3.0$. Also the parameters E_0 and A approximately form a normal distribution in the parameter space, showing no signs of correlation.



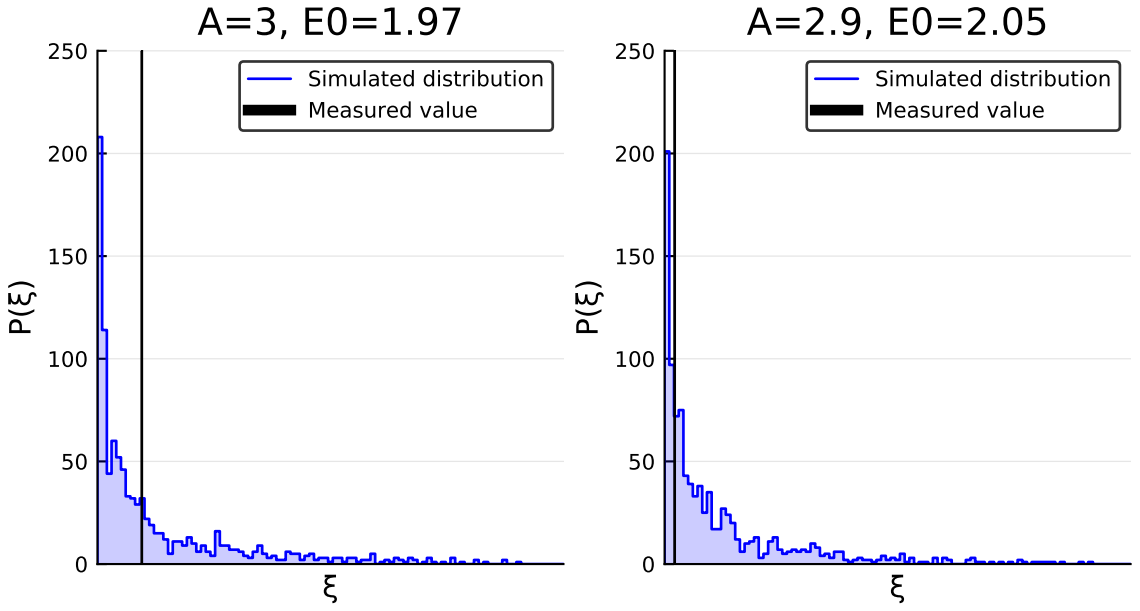
(b)

Following the procedure from the lecture, we choose the likelihood of the data given a certain set of model parameters as our test statistic.

$$\xi(\mathbb{D}, A, E_0) = P(\mathbb{D}|A, E_0)$$

For each point in (A, E_0) space, we can determine the probability distribution $P(\xi|A, E_0)$ by sampling from random experiments, simulated on a computer. We then define the 68% CL region as all A, E_0 for which $\xi(\mathbb{D}, A, E_0)$ lies within the 68% probable interval.

This was simulated on a computer with 1000 experiments for each parameter configuration. The resulting distributions in ξ for two parameter choices are shown below.

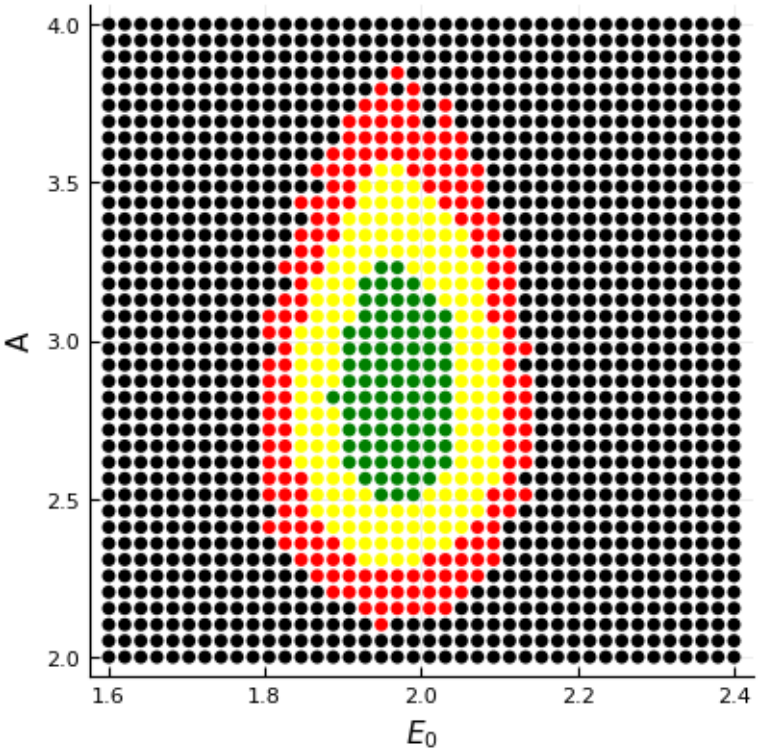


For the chosen test statistic, a higher value of ξ means a more likely outcome. Therefore the 68%

confidence level interval in $P(\xi)$ is defined so the right tail integrates to 68%.

$$\int_{\xi_{0.68}}^{\infty} d\xi \; P(\xi) = 0.68$$

A numerical simulation of 1000 experiments for each (A, E_0) yields the 68% CL region plotted below. The green points lie within the 68% CL region, the yellow within 95% and the red within 99.7%.



Exercise 2: Bayesian inference

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

$$\epsilon(E) = \sin(A(E - E_0))$$

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

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

(c) Discuss the results.

Solution:

(a)

Like in the previous exercise, we do a Bayesian model fit, this time for the function $\epsilon(E|A, E_0)$. The likelihood is then given as

$$P(\mathbb{D}|A, E_0) = \prod_{i=1}^{8} \binom{n_i}{r_i} \epsilon(E_i|A, E_0)^{r_i} (1 - \epsilon(E_i|A, E_0))^{n_i - r_i}.$$

and the posterior is

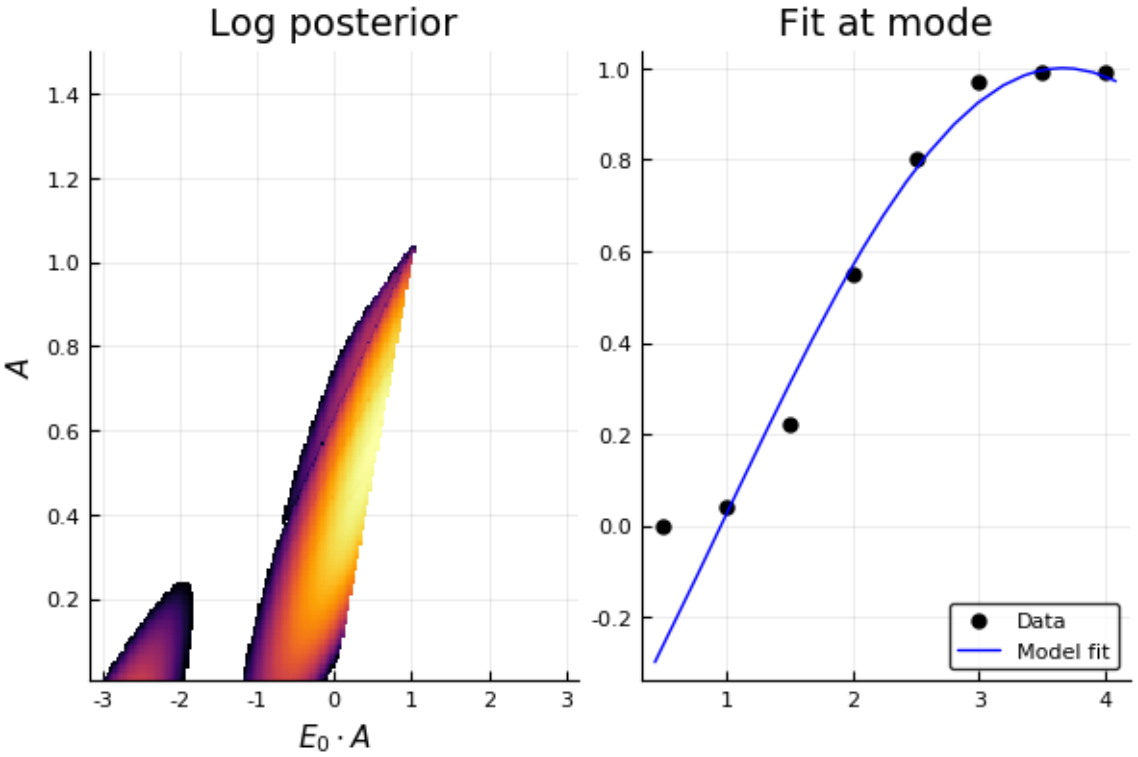
$$P(A, E_0|\mathbb{D}) = \frac{P(\mathbb{D}|A, E_0) \; P(A, E_0)}{\int dA \int dE_0 P(\mathbb{D}|A, E_0) \; P(A, E_0)}$$

Unlike with the Sigmoid function in the previous exercise, $\epsilon(E)$ is periodic in E_0 and invariant under a flip $A- > -A$ and a shift. The posterior probability density will therefore exhibit the same symmetries and contain infinitely many modes which represent the same configuration.

To avoid this redundancy, we restrict the possible values of A and E_0 to only contain one interval. This can be implemented through the priors $P(A, E_0)$, which we set to be zero outside the region of interest. First, we exclude the region for which $A < 0$ because there is an identical configuration for positive A. Next, we observe that the periodicity in E_0 has a length of $\frac{2\pi}{|A|}$. Expressed differently, the parameter $\tilde{E}_0 \equiv E_0 \cdot A$ is 2π-periodic. This is a very convenient property as it allows us to fit a rectangular grid in $(A, \tilde{E}_0)$ space. Putting this together, we choose a flat prior in the region $(A, \tilde{E}_0) \in [0, \infty) \times [-\pi, \pi)$. It follows, that the prior in the original variable E_0 is not flat anymore.

$$P(E_0) = P(\tilde{E}_0)|\frac{d\tilde{E}_0}{dE_0}| = A\ P(\tilde{E}_0)$$

A numerical evaluation of the posterior on the computer yields the diagram below.

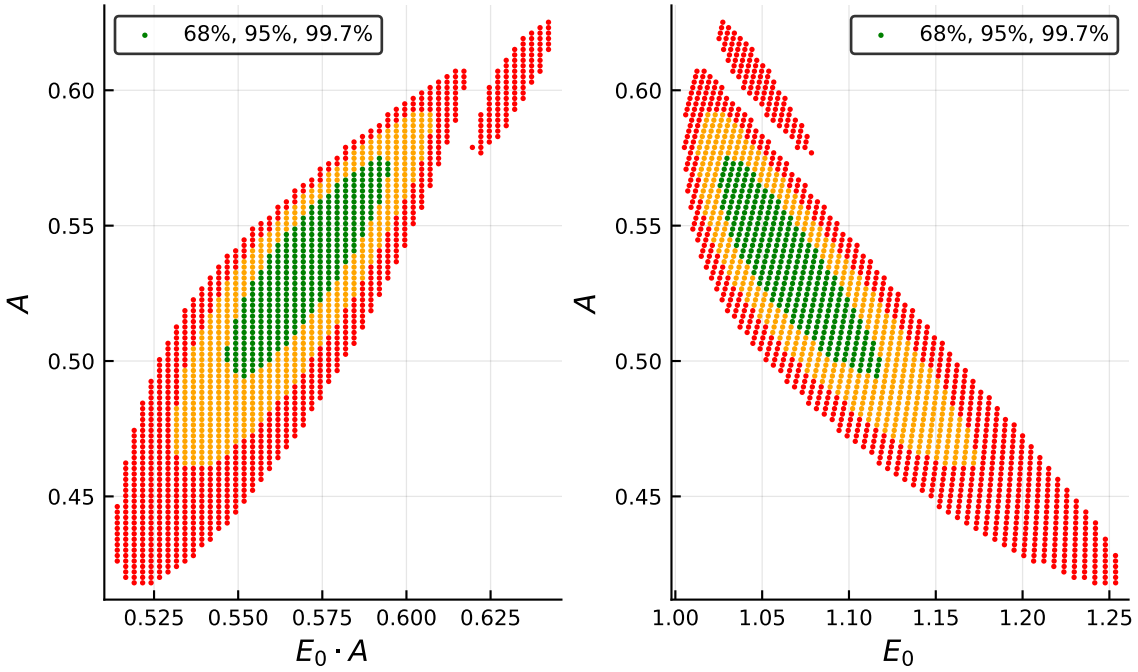


The mode of the distribution is located at $A = 0.58$ and $E_0 = 0.95$.

(b)

To find the 68% Confidence Level region, we add up the individual probabilities of the posterior, starting with the largest, until the sum surpasses 0.68.

This is done for the grid points from assignment (a) on the computer. In addition to the 68% set (green), the 95% (orange) and 99.7% (red) CL regions are also plotted. The left plot shows the evenly spaced points on the $(A, \tilde{E}_0)$ grid while the right plot is the transformed version showing the parameter E_0.



(c)

While the best-fit result may look somewhat reasonable from a purely mathematical standpoint, there are a couple of issues when you consider the physical background. The data we fit represents the efficiency of a detector which has some turn-on energy E_0. Above E_0, the efficiency is high, below E_0, the efficiency is zero. The region around E_0 should interpolate between these states in a continuous, differentiable fashion. The Sigmoid function from exercise 1 does that very well, but the sin function falls short in several ways.

1. The slope of the function does not decrease for small energies. It is therefore impossible to approach zero efficiency in a differentiable fashion.
2. The function is periodic, so at higher energies the efficiency decreases. This effect does not represent the behavior of real detectors.
3. The function can produce negative values and does so even in the region where data was taken. Negative efficiencies are not possible in a physical context.

All in all the sine function is most likely a bad model for any kind of application based on the efficiency data.

Exercise 3: Properties of χ^2 for one data point

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

Solution:

The χ^2 distribution is defined as the sum of squared residuals of all data points

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

where (x_i, y_i) are the data points with weights $w_i \in \mathbb{R}^+$, f is the model function with model parameters λ.

For one data point (x, y), χ^2 reduces to

$$\chi^2 = \frac{(y - f(x|\lambda))^2}{w^2}.$$

The observed y follows some distribution $P(y|x)$. Assuming our model is correct, we can write $y(x) = f(x|\lambda) + r$ where r is a random noise term. Commonly the distribution of r is assumed to be a normal distribution with mean zero and variance σ^2. This assumption is somewhat justified by the CLT in case that every y is the sum or mean of many individual measurements.

In the general case, we can now express the distribution for χ^2 as

$$P(\chi^2) = P(y(\chi^2))|\frac{dy}{d\chi^2}|$$

Inverting the definition of χ^2 yields

$$y(\chi^2) = w\sqrt{\chi^2} + f(x|\lambda)$$

so that

$$P(\chi^2) = P(y(\chi^2))\frac{w}{\sqrt{\chi^2}}.$$

At this point, we have to know the distribution $P(y)$ to continue. We will assume a normal distribution for $r \sim \mathcal{N}(0, \sigma)$ which is equivalent to assuming $y \sim \mathcal{N}(f(x|\lambda), \sigma)$. Then,

$$P(\chi^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{y - f(x|\lambda)}{\sigma}\right)^2\right]\frac{w}{\sqrt{\chi^2}} = \frac{1}{\sqrt{2\pi\chi^2}}e^{-\frac{\chi^2}{2}}.$$

The last step follows from the choice $w = \sigma$ for the weight parameters. Evidently the mode of $P(\chi^2)$ is located at $\chi^2 = 0$ because $\frac{d}{d\chi^2}(\chi^2) < 0$.

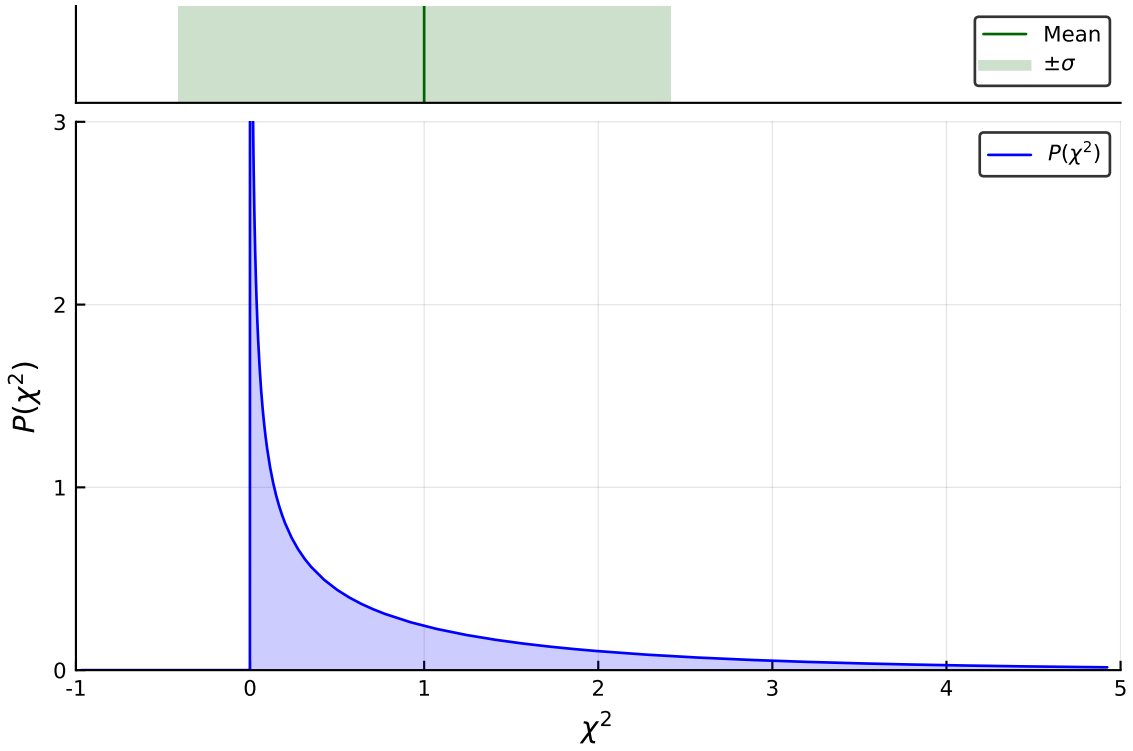
For the mean value, we calculate the integral

$$\mathbb{E}[\chi^2] = \int_0^\infty d\chi^2\ \chi^2\ P(\chi^2) = \frac{1}{\sqrt{2\pi}}\int_0^\infty d\chi^2\ \chi^{2\,\frac{1}{2}}e^{-\frac{\chi^2}{2}} = 1$$

which is a standard integral using the Gamma function $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Similarly, the variance of the distribution is

$$\mathbb{V}[\chi^2] = \int_0^\infty d\chi^2\ \chi^{2^2}\ P(\chi^2) - \mathbb{E}[\chi^2] = \frac{1}{\sqrt{2\pi}}\int_0^\infty d\chi^2\ \chi^{2\,\frac{3}{2}}e^{-\frac{\chi^2}{2}} - \mathbb{E}[\chi^2] = 3 - 1 = 2.$$

The probability density distribution of χ^2 as well as the mean and standard deviation $\sigma = \sqrt{\mathbb{V}[\chi^2]}$ is plotted below.



Exercise 8: Bayesian and Frequentist model comparison

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:

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

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

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

(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?

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

Solution:

We are given data $\mathbb{D}$ consisting of $N = 18$ data points and two models, one background-only model $M_1$and another, M_2, modeling background+signal. Both models give a prediction $f(x|\vec{\lambda})$ where $\vec{\lambda}$ refers to the model parameters, A, B, C for M_1 and A, B, C, D, x_0, Γ for M_2.

The likelihood of the data is assumed to follow a normal distribution with $\sigma = 4$ in each data point

$$P(\mathbb{D}|\vec{\lambda}) = \prod_{i=1}^{18} P(y_i|x_i, \vec{\lambda}) = \prod_{i=1}^{18} \mathcal{N}\left(y_i|f(x_i|\vec{\lambda}), 4\right)$$

(a)

For the χ^2 minimization fit, we use an existing numerics library to find the best fitting values as well as the covariance matrix.

For M_1, the optimal values are $A = -7.27$, $B = 173.5$ and $C = -28.9$. The covariance matrix is

$$\begin{pmatrix} 94 & -377 & 318 \\ -377 & 1765 & -1604 \\ 318 & -1604 & 1527 \end{pmatrix}$$

The resulting value of χ^2 for these model parameters is

$$\chi^2 = \sum_{i=1}^{18} \left(\frac{y_i - f(x_i|\vec{\lambda})}{\sigma} \right)^2 = 92.5.$$

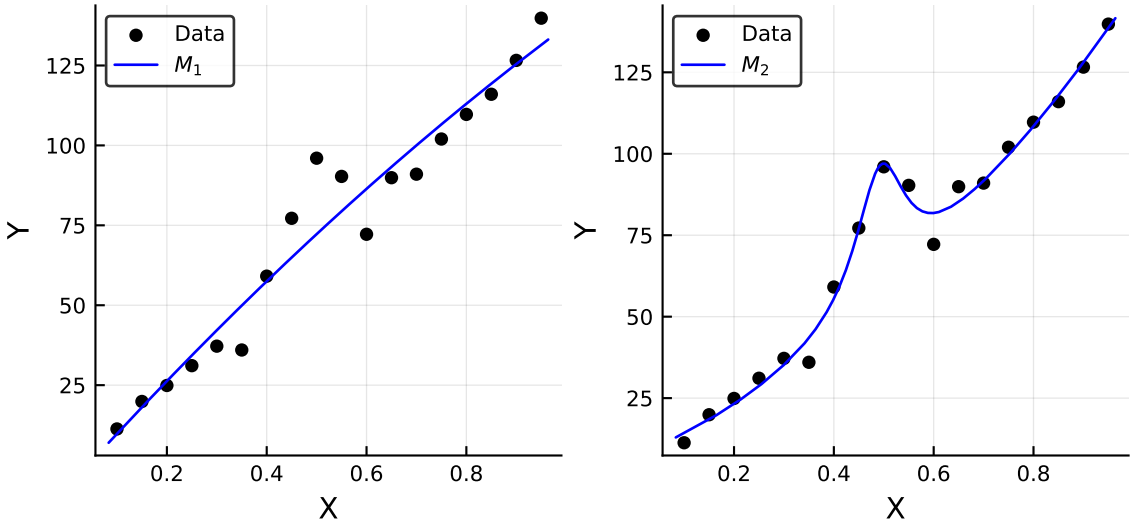
For M_2, the results are $A = 6.71$, $B = 56.8$, $C = 85.5$, $D = 0.155$, $x_0 = 0.493$ and $\Gamma = 0.062$. The

covariance matrix is

$$\begin{pmatrix} 25.5836 & -136.0465 & 122.5431 & 0.1667 & 0.0062 & 0.0258 \\ -136.0465 & 930.0140 & -884.9216 & -1.7123 & -0.0331 & -0.2781 \\ 122.5431 & -884.9216 & 858.0021 & 1.6792 & 0.0267 & 0.2727 \\ 0.1667 & -1.7123 & 1.6792 & 0.0061 & 0.0001 & 0.0011 \\ 0.0062 & -0.0331 & 0.0267 & 0.0001 & 0.0001 & 0.0000 \\ 0.0258 & -0.2781 & 0.2727 & 0.0011 & 0.0000 & 0.0002 \end{pmatrix}$$

The error of the fit is here $\chi^2 = 15.0$.

The best fit curves of both models are plotted below.



We can get the p-value of our fits by integrating $P(\chi^2)$ from the found χ^2_{fit} to infinity.

$$p = \int_{\chi^2_{\text{fit}}}^{\infty} d\chi^2 \; P(\chi^2)$$

The distribution of $P(\chi^2)$ depends both on the number of data points N and the number of parameters n. Specifically, it depends on their difference $\hat{N}$, the number of degrees of freedom.

$$P(\chi^2) = \frac{(\chi^2)^{\frac{\hat{N}}{2}-1}}{2^{\frac{\hat{N}}{2}}\,\Gamma(\frac{\hat{N}}{2})} e^{-\frac{\chi^2}{2}}$$

and has a mean of $\mathbb{E}[\chi^2] = \hat{N}$ which is 15 for M_1 and 12 for M_2. The resulting p-values then are $3 \cdot 10^{-13}$ for M_1 and 0.24 for M_2. For the correct model, the p-value is uniformly distributed between zero and one while an incorrect model will produce a small p-value. With this in mind, the p-value of M_1 makes this model seem rather unlikely while M_2 is the preferred model for the data.

(b)

For a Bayesian fit, we require the probability density for the model parameters $\vec{\lambda}$, the posterior probability density. We get the posterior by utilizing Bayes' theorem

$$P(\vec{\lambda}|\mathbb{D}) = \frac{P(\mathbb{D}|\vec{\lambda}) \; P(\vec{\lambda})}{\int d\vec{\lambda} \; P(\mathbb{D}|\vec{\lambda}) \; P(\vec{\lambda})}$$

By choosing flat priors $P(\vec{\lambda}) = \text{const}$, the equation reduces to

$$P(\vec{\lambda}|\mathbb{D}) = \frac{P(\mathbb{D}|\vec{\lambda})}{\int d\vec{\lambda} \; P(\mathbb{D}|\vec{\lambda})} = \frac{\prod_{i=1}^{18} \mathcal{N}\left(y_i | f(x_i|\vec{\lambda}), 4\right)}{\int d\vec{\lambda} \; \prod_{i=1}^{18} \mathcal{N}\left(y_i | f(x_i|\vec{\lambda}), 4\right)}$$

where the denominator is a normalization constant. To get uncertainties on the parameter values, we can evaluate the marginalized probability distributions for the individual model parameters

$$P(\lambda_i|\mathbb{D}) = \int_{j\neq i} d\lambda_j \; P(\vec{\lambda}|\mathbb{D})$$

and quote the 68% CL smallest interval on the marginalized distribution.

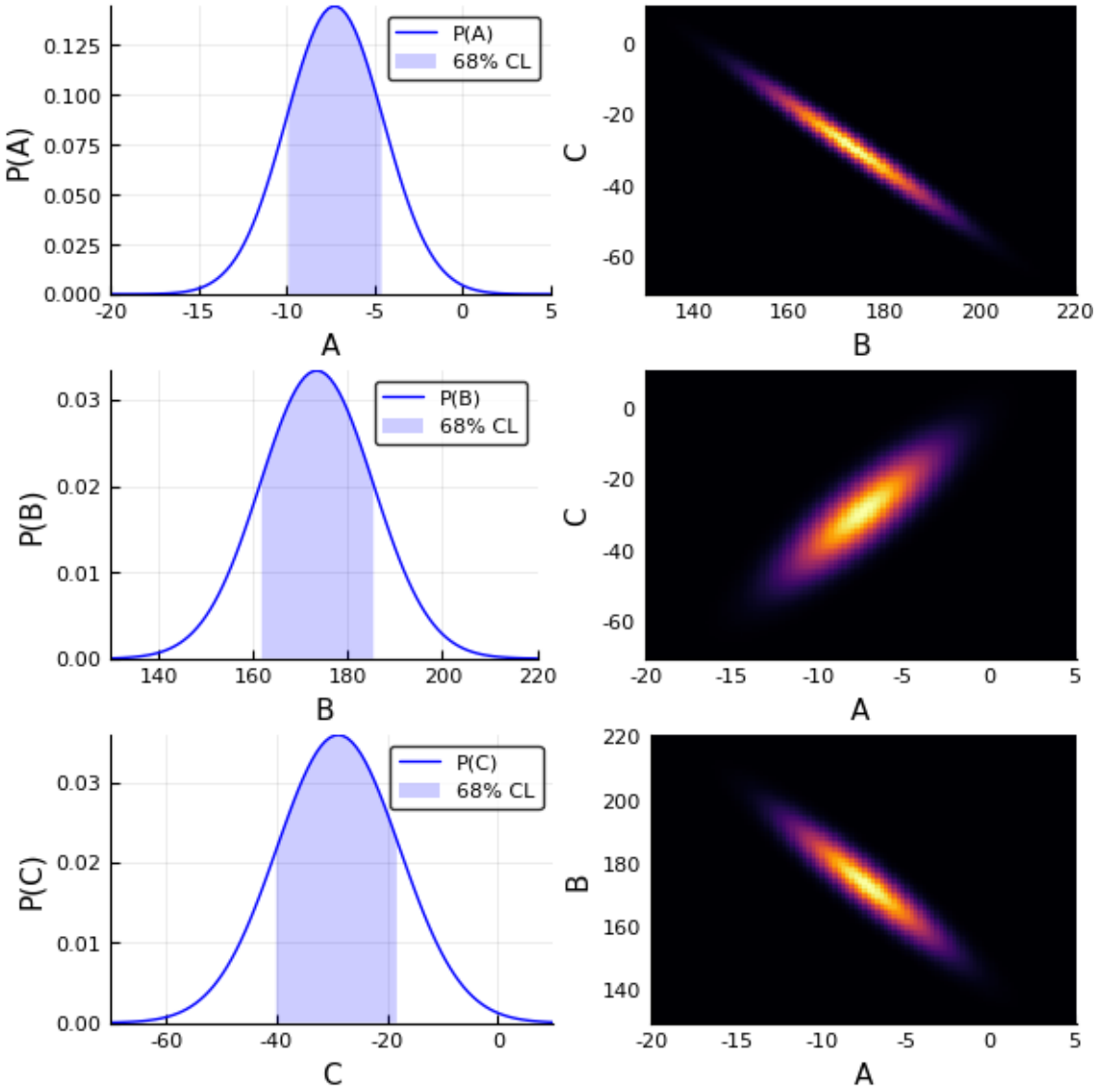
The posterior can be evaluated numerically, e.g. on a equidistant grid. This approach is limited by the fact that every parameter adds a dimension to the search space which is three-dimensional for M_1 and six-dimensional for M_2. Also the grid must be chosen large enough so that varying one parameter λ_i does not move the contributing region of the other parameters $\lambda_{j\neq i}$ out of bounds.

For finding the mode, we can also make use of the fact that the denominator is constant, so

$$P(\vec{\lambda}|\mathbb{D}) \sim \prod_{i=1}^{18} \mathcal{N}\left(y_i|f(x_i|\vec{\lambda}), 4\right) \sim \prod_{i=1}^{18} \exp\left[-\frac{1}{2}\left(\frac{y_i - f(x_i|\vec{\lambda})}{4}\right)^2\right] \sim \exp\left[-\sum_{i=1}^{18}\left(\frac{y_i - f(x_i|\vec{\lambda})}{4}\right)^2\right]$$

Like in exercise 14 of chapter 4, we see that to get the mode of the posterior, we simply need to find the $\vec{\lambda}^*$ for which χ^2 is minimal. This has already been done in problem (a) of this exercise, so the modes are equal.

For M_1, the full posterior was calculated on a three-dimensional grid with 100 points in each parameter around the mode. The marginalized distributions are plotted below. The mode is located at $A = -7.27$, $B = 173.5$, $C = -28.9$ and the most likely values for the individual parameters are $A = -7.4 \pm 2.8$, $B = 173 \pm 12$, $C = -29 \pm 11$ (68% CL).



This brute-force numerical approach would lead to very long computation times in six dimensions. Taking 100 steps in each parameter leads to a trillion grid points that need to be evaluated.

Instead, we will approximate the posterior density $P(\lambda_i|\mathbb{D})$ to second order, assuming the likelihood is unimodal in every parameter. Specifically, we assume that the likelihood falls off like an asymmetric Gauss curve. This behavior is motivated by the fact that the distribution of χ^2 also approximates a normal distribution.

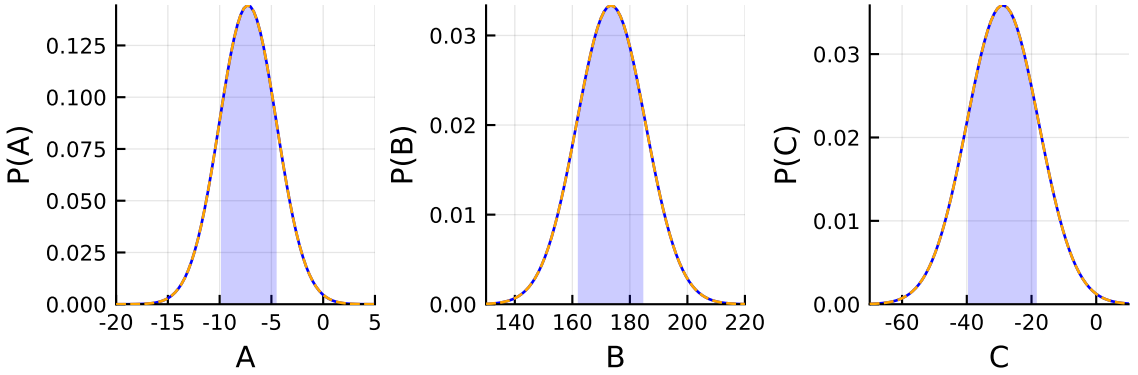
For a specific value of the parameter λ_i to be investigated, the largest contribution to $P(\lambda_i|\mathbb{D})$ comes from the point where all other parameters are at the mode of the so-constrained parameter space, $\lambda^*_{j\neq i}$. For the second order, we take a sample Δ above and Δ below that mode in every dimension and fit a Gauss curve, leading to a relative probability contribution of

$$P \sim \sigma = \frac{\Delta}{\sqrt{-2\ln(\frac{A_\Delta}{A_c})}}$$

for each half where A_C is the posterior at the constrained mode $\lambda^*_{j\neq i}$ and A_Δ at $\lambda^*_{j\neq i} + \Delta\ \hat{e}_d$.

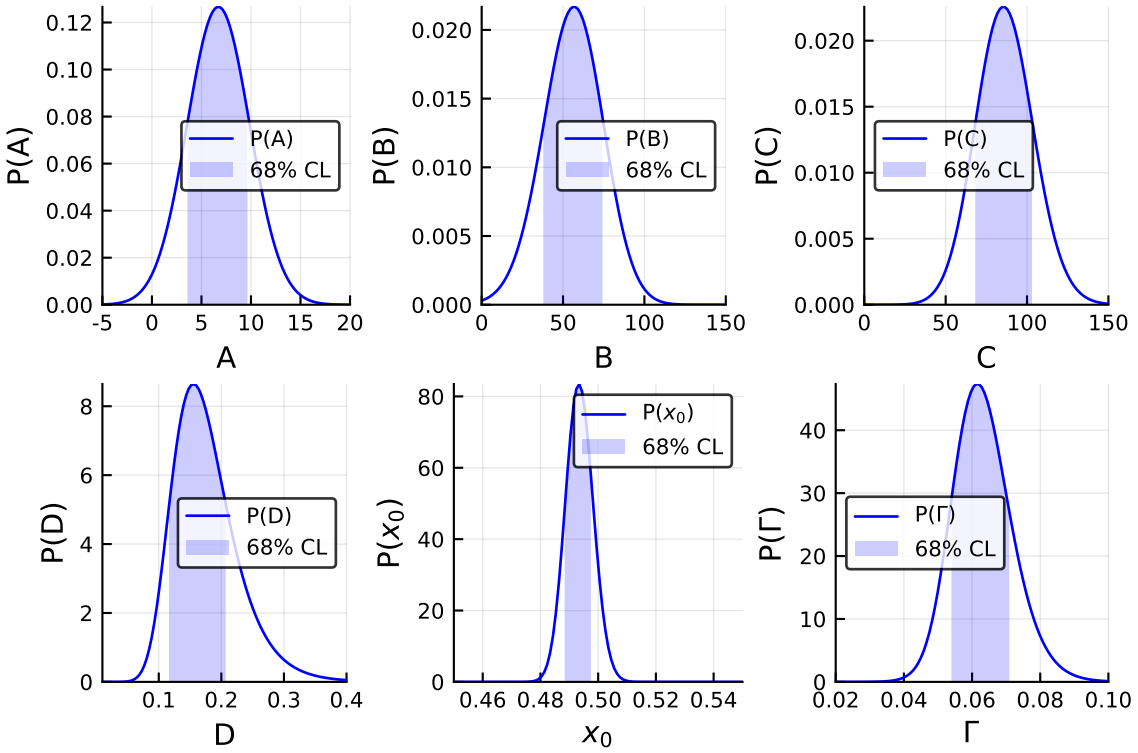
This approximation reduces the number of computations from $\mathcal{O}(n^6)$ to $\mathcal{O}(n)$ but with a higher constant factor because a least squares fit is performed for each point.

For the three-dimensional fit of M_1, the below plot shows the comparison between the grid-based approach (blue line) and the approximation (dashed orange line). As can be seen, the curves are indistinguishable when the grid-based approach has a fine enough resolution.



Reducing the amount of grid points in each dimension from 100 to 40 already leads to an error of up to 10% in the uncertainties. The approximation method does not suffer from this problem as the minimization algorithm can quickly find the constrained mode with a much higher precision than the step size of the grid. Therefore the approximation method produces more accurate results at a lower cost when small step sizes in each dimension are not feasible.

The marginalizations for M_2 have been evaluated on a computer using the fast method outlined above. The computation of all marginalizations finished within one second of computation time for 100 points along each axis. The distributions are plotted below.



The mode is located at $A = 6.71$, $B = 56.8$, $C = 85.5$, $D = 0.155$,$x_0 = 0.493$, $\Gamma = 0.0616$. The most likely values for the individual parameters are $A = 6.6 \pm 3.0$, $B = 56 \pm 18$, $C = 84^{+18}_{-17}$, $D = 0.16^{+0.05}_{-0.04}$, $x_0 = 0.493^{+0.004}_{-0.005}$, $\Gamma = 0.061^{+0.010}_{-0.007}$ (68% CL).

To compare the two models, we can use the fraction of their respective posterior probabilities.

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

Using Bayes' theorem, this can be expressed through the likelihoods of the data given the models and the prior probability for the models. The Bayes factor K is defined as the middle fraction and describes how our beliefs should change in the light of the data.

$$K = \frac{P(\mathbb{D}|M_1)}{P(\mathbb{D}|M_2)}$$

Since our models are dependent on a number of parameters $\vec{\lambda}$, these need to be integrated out, yielding

$$K = \frac{\int d\vec{\lambda}_1 \; P(\mathbb{D}|M_1, \vec{\lambda}_1) \; P(\vec{\lambda}_1)}{\int d\vec{\lambda}_2 \; P(\mathbb{D}|M_2, \vec{\lambda}_2) \; P(\vec{\lambda}_2)}.$$

For the problem at hand, we choose flat priors for both models $P(\vec{\lambda}_1) = \text{const}$, $P(\vec{\lambda}_2) = \text{const}$. Then, the Bayes factor can be evaluated numerically by calculating the integrals of both models within the whole parameter space. This was approximated using the marginalizations obtained above and assuming that the likelihood factors in the parameters. While this is a crude approximation of the real distribution, the real value can only be obtained using advanced statistical methods like Markov Chain Monte Carlo or brute force calculation on an extremely large grid.

The resulting Bayes factor was calculated to be $K = 8.8 \cdot 10^{-10}$. A Bayes factor of less than 0.01 is generally considered to be very strong evidence for the second model. This much smaller K therefore rules out M_1 with almost certainty.

Exercise L1: Maximum likelihood for the Bernoulli distribution

The family of Bernoulli distributions have the probability density $P(x|p) = p^x(1-p)^{1-x}$.
(a) Calculate the Fisher information $I(p) = -\mathbb{E}[\frac{\partial^2 \ln P(x|p)}{\partial p^2}]$
(b) What is the maximum likelihood estimator for p?
(c) What is the expected distribution for $\hat{p} - p$?

Solution:

(a) We calculate the Fisher information using the provided formula.

$$\begin{aligned} I(p) &= -\mathbb{E}[\frac{\partial^2 \ln P(x|p)}{\partial p^2}] \\ &= -\mathbb{E}[\frac{\partial^2}{\partial p^2}(x \ln p + (1-x)\ln(1-p)] \\ &= -\mathbb{E}[\frac{\partial}{\partial p}(\frac{x}{p} - \frac{1-x}{1-p})] \\ &= -\mathbb{E}[-\frac{x}{p^2} - \frac{1-x}{(1-p)^2}] \\ &= \frac{p}{p^2} + \frac{1-p}{(1-p)^2} \\ &= \frac{(1-p)+p}{p(1-p)} = \frac{1}{p(1-p)} \end{aligned}$$

(b)

For N measurements, the likelihood function is

$$P(\mathbb{D}|p) = \prod_{i=1}^{N} p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{N-\sum_i x_i} = p^r(1-p)^{N-r}.$$

where $r \equiv \sum_i x_i$. To find the maximum, we solve

$$\max_p P(\mathbb{D}|p) \Leftrightarrow \frac{\partial P(\mathbb{D}|p)}{\partial p} = 0 \Leftrightarrow \frac{\partial \ln P(\mathbb{D}|p)}{\partial p} = 0$$

which yields

$$\frac{\partial \ln P(\mathbb{D}|p)}{\partial p} = \frac{\partial}{\partial p}[r \ln p + (N-r)\ln(1-p)] = \frac{r}{p} - \frac{N-r}{1-p} \stackrel{!}{=} 0$$

The maximum is therefore located at $p = \frac{r}{N} \equiv \hat{p}$ and will be used as our maximum likelihood estimator for p.

(c)

The deviation of our estimator $\hat{p}$ from the real value of p, $P(\hat{p} - p)$ gives an indication of how good the estimator is. We know the probability distribution of $r = \hat{p}N = \sum_i x_i$ is a binomial distribution in r

$$P(r) = \binom{N}{r} p^r (1-p)^{N-r}$$

Therefore using the transformation of random variables, the distribution of $\Delta p \equiv \hat{p} - p$ is

$$P(\Delta p) = N \binom{N}{N(\Delta p + p)} p^{N(\Delta p + p)} (1-p)^{N - N(\Delta p + p)}.$$

which has its mode at $\Delta p = 0$ as expected. For $N \to \infty$, this curve will approximate a normal distribution around zero with variance $\mathbb{V}[\Delta p] = \frac{p\ (1-p)}{N}$.

Exercise L2: Maximum likelihood for the exponential distribution

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

(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 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 CLTL (see lecture notes) and discuss.

Solution:

First, we need to find a maximum likelihood estimator for the exponential distribution $P(x|\lambda) = \lambda e^{-\lambda x}$. For N measurements x_i, the likelihood is

$$P(\mathbb{D}|\lambda) = \prod_{i=1}^{N} \lambda e^{-\lambda x_i} = \lambda^N e^{-\lambda \bar{x} N}$$

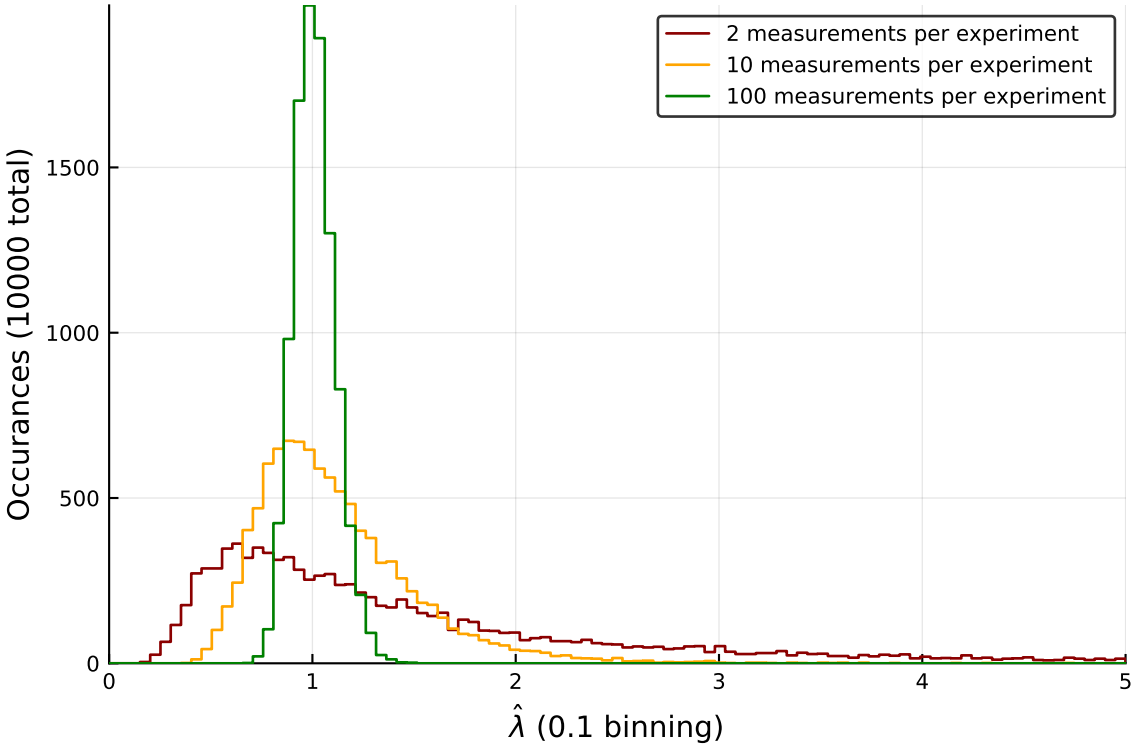
where $\bar{x} \equiv \frac{1}{N} \sum_{i=1}^{N} x_i$. For the maximum of the likelihood, we require

$$\frac{dP(\mathbb{D}|\lambda)}{d\lambda} = \frac{d}{d\lambda}[\lambda^N e^{-\lambda \bar{x} N}] = \lambda^N e^{-\lambda \bar{x} N} (\frac{N}{\lambda} - \bar{x} N) \stackrel{!}{=} 0.$$

which is fulfilled for $\hat{\lambda} = \frac{1}{\bar{x}}$, which will be our estimator for λ.

(a)

For this task, we investigate the distribution of $\hat{\lambda}$ by simulating experiments via sampling from the exponential distribution for $\lambda = 1$. 10,000 experiments have been simulated on the computer for n=2, 10 or 100. The resulting distributions of $\hat{\lambda}$ are shown below.



(b)

The law of large numbers (LLN) says that for a large number of experiments $N \to \infty$, the error $\Delta\lambda \equiv \hat{\lambda} - \lambda$ will go towards zero. This behavior can be seen in the plot, where the distribution $P(\hat{\lambda})$ becomes more and more sharply peaked around the true $\lambda = 1$ as n increases.