Acknowledgement
Preface
# Contents

1 Introduction to Probabilistic Reasoning 11
  1.1 What is Probability ? ..................................... 12
    1.1.1 Mathematical Definition of Probability ................. 13
  1.2 Scientific Knowledge ....................................... 15
    1.2.1 Updating Scheme ....................................... 16
    1.2.2 Schools of Data Analysis ................................ 18
    1.2.3 Which probability ? ..................................... 19
  1.3 Warm-up Exercise ........................................... 20
    1.3.1 Apparatus and Event Type ............................... 20
    1.3.2 Efficiency of Apparatus ................................. 21
    1.3.3 Exercise in Bayesian Logic .............................. 22

2 Binomial and Multinomial Distribution 25
  2.1 Binomial Distribution ....................................... 25
    2.1.1 Some Properties of the Binomial Distribution .......... 26
    2.1.2 Example ................................................. 27
    2.1.3 Central Interval ......................................... 27
    2.1.4 Smallest Interval ........................................ 28
  2.2 Neyman Confidence Level Construction ....................... 29
    2.2.1 Example ................................................. 31
  2.3 Likelihood .................................................. 31
  2.4 Limit of the Binomial Distribution .......................... 32
    2.4.1 Likelihood function for a Binomial Model ............... 32
    2.4.2 Gaussian Approximation for the Binomial Probability ... 33
  2.5 Extracting Knowledge using Bayes Equation .................. 36
    2.5.1 Flat Prior ............................................... 36
    2.5.2 Defining Probable Ranges ............................... 38
    2.5.3 Nonflat Prior ............................................ 42
2.5.4 A Detailed Example ........................................ 43
2.6 Multinomial Distribution .................................... 49
  2.6.1 Bayesian Data Analysis ................................. 49
  2.6.2 Likelihood .................................................. 50
  2.6.3 Adding the Prior .......................................... 51
3 Poisson Distribution ............................................. 57
  3.1 Introduction .................................................. 57
  3.2 Derivation ..................................................... 57
    3.2.1 Limit of the Binomial Distribution .................. 58
    3.2.2 Constant Rate Process & Fixed Observation Time .... 59
  3.3 The Poisson Distribution ................................... 60
    3.3.1 Properties .............................................. 60
    3.3.2 Examples and Graphical Display ..................... 61
  3.4 Bayesian Analysis .......................................... 65
    3.4.1 Discussion and Examples: Flat Prior Results ....... 66
    3.4.2 Non-flat Priors ......................................... 69
  3.5 Superposition of two Poisson Processes .................. 72
    3.5.1 Likelihood for Multiple Poisson Processes ........... 73
  3.6 Feldman-Cousins Intervals ................................ 75
  3.7 Bayesian Analysis of Poisson Process with Background ... 78
    3.7.1 Comparison of the Bayesian and Feldman-Cousins Intervals 78
  3.8 Uncertain Background ....................................... 82
    3.8.1 Frequentist Analysis .................................. 82
    3.8.2 Bayesian Analysis ...................................... 82
  3.9 Bayesian Analysis of the On/Off Problem ................ 83
    3.9.1 Source discovery with On/Off data ................... 87
    3.9.2 On/off with factorized Jeffreys priors ................ 88
4 Gaussian Probability Distribution Function .................. 97
  4.1 The Gauss Distribution as the Limit of the Binomial Distribution 98
  4.2 The Gauss Distribution as the Limit of the Poisson Distribution 98
  4.3 Gauss’ derivation of the Gauss distribution .............. 99
  4.4 Some properties of the Gauss distribution ................ 101
  4.5 Characteristic Function .................................... 103
    4.5.1 Sum of independent variables ....................... 104
    4.5.2 Characteristic Function of a Gaussian .............. 106
    4.5.3 Sum of Gauss distributed variables ................ 106
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5.6</td>
<td>$\chi^2$ distribution after fitting parameters</td>
<td>169</td>
</tr>
<tr>
<td>5.5.7</td>
<td>$\chi^2$ for Poisson distributed data</td>
<td>171</td>
</tr>
<tr>
<td>5.5.8</td>
<td>General $\chi^2$ Analysis for Gauss distributed data</td>
<td>172</td>
</tr>
<tr>
<td>5.5.9</td>
<td>$\chi^2$ for Bayesian and Frequentist Analysis</td>
<td>173</td>
</tr>
<tr>
<td>5.6</td>
<td>Maximum Likelihood analysis</td>
<td>174</td>
</tr>
<tr>
<td>5.6.1</td>
<td>Parameter uncertainties in likelihood analysis</td>
<td>175</td>
</tr>
<tr>
<td>5.6.2</td>
<td>Likelihood ratio as test statistic</td>
<td>178</td>
</tr>
<tr>
<td>5.6.3</td>
<td>Likelihood ratio for Poisson distributed data</td>
<td>181</td>
</tr>
<tr>
<td>5.7</td>
<td>Model Selection</td>
<td>183</td>
</tr>
<tr>
<td>5.7.1</td>
<td>Fisher’s significance testing</td>
<td>184</td>
</tr>
<tr>
<td>5.7.2</td>
<td>Neyman-Pearson hypothesis testing</td>
<td>184</td>
</tr>
<tr>
<td>5.7.3</td>
<td>Bayesian posterior probability test</td>
<td>185</td>
</tr>
<tr>
<td>5.8</td>
<td>Test of $p$-value definitions, from Beaujean et al.</td>
<td>185</td>
</tr>
<tr>
<td>5.8.1</td>
<td>Data with Gaussian uncertainties</td>
<td>186</td>
</tr>
<tr>
<td>5.8.2</td>
<td>Example with Poisson uncertainties</td>
<td>188</td>
</tr>
<tr>
<td>5.8.3</td>
<td>Exponential Decay</td>
<td>194</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction to Probabilistic Reasoning

Our high-level goal is to understand knowledge acquisition so that we may best optimize the process. In this chapter, we will introduce what we mean by knowledge in a scientific context, and show the role of data in increasing our store of knowledge. Knowledge will not be just a collection of facts but more importantly will be coded into a model that allows us to make calculations (predictions) concerning future outcomes. The modeling will typically have two components - one that we may call the scientific model, and the second the statistical model. The former will contain our description of nature (e.g., the Standard Model of particle physics, a model of the Earth’s layers, etc.) while the latter will be experiment specific and will describe how we expect possible data outcomes to be distributed given a specific experimental setup and the physical model. The physical model is specific to the branch of science and, while we will use a variety of examples from different sources, will not be our concern. Rather, we are interested in how to formulate the statistical model, and how to use the data to then improve our knowledge of the physical model. This aspect is quite general so that the concepts and techniques you will learn here can be used in a wide range of settings.

In short, you will learn:

- the basics of probabilistic reasoning;
- how to define an appropriate statistical model for your experimental situation;
- how to use pre-existing information in the probabilistic analysis;
• many numerical techniques for solving complicated problems.

1.1 What is Probability?

Although we all have an intuition what we mean with the word probability, it is difficult to define. Let us think of a few examples:

• the probability that when I flip a coin, heads will come up;

• the probability of the result TTTTTTTTH when I flip a coin ten times (T=tails, H=heads);

• the probability that it will rain tomorrow;

• the probability that there is extraterrestrial life;

In the first example, you probably think - oh, this is easy, the probability is 50%. Where did this come from? You made a model, reasoning as follows:

1. there are two possible outcomes (lets ignore the possibility of the coin landing on its side);

2. I have no reason to prefer one outcome over the other (Laplace’s rule of insufficient reason), so I assign the same probability to each;

3. the sum of the probabilities should be one (the coin must end up on one of its sides);

4. therefore, the probability is 50%.

But this result depended on several assumptions; i.e., the probability is a number that we created and that is the result of our model. If the coin was not fair, then the ‘real probability’ would be different. If we knew about this, we would also assign a different probability to heads or tails. I.e., probability is a number that we create based on a model that we think represents the situation, but that we are free to change as we get more information. Probability is not inherent in the coin, but in our model - and our model represents what we currently believe reflects the physical situation.
1.1. WHAT IS PROBABILITY?

In the second example, you may have said that the result is \((1/2)^{10} \approx 0.001\) because every sequence has the same probability (again taking our model of a fair coin) and there are \(2^{10}\) possible sequences. After seeing the result, you may suspect your model is not quite right and wonder if the probability should really be so small. And maybe you did not calculate the probability of the sequence, but rather the probability of getting 9T and 1H in ten tosses of your coin. If you used your fair-coin model, then you would have found for the probability 0.01 rather than 0.001. I.e., even in the same model and with the same data, you can calculate different probabilities. Which is correct? Is one probability somehow better than the other?

If we skip to the last example, then we have a situation where many models can be created with very different resulting probabilities. E.g., we can estimate the number of solar systems in the Universe - let’s say it is \(10^{12}\), estimate how many of these will have planets with conditions suitable for life - this will have very large uncertainty, then we need to have a distribution for the probability that life actually developed given a possible environment. Here we are left with pure guess work. We know that it is possible - we are here - but we also know that the number cannot be too close to 1 or we would have already found life (or it would have found us) in a nearby region of our galaxy. So our result for the probability of extraterrestrial life will depend completely on what we pick for this distribution, and without further information we will not have strong belief in any prediction being bandied about.

When we are discussing the frequency of outcomes within a mathematical model, then we can calculate a number that we call probability (e.g., heads in a Binomial model), but when we are discussing the probability of a data outcome, then we are more properly talking about a degree-of-belief, since we are never sure that the mathematical model we have created applies to the physical situation we have at hand. The model that we have chosen will itself have a degree-of-belief assigned to it. In the case of the coin flipping, it could have a high degree of belief, but in the case of extraterrestrial life, it should have a low degree of belief, and will very much depend on the individual at hand. The developers of the models will typically believe much more strongly in their models than others.

1.1.1 Mathematical Definition of Probability

The mathematical definition of probability, due to Kolmogorov, is rather straightforward and obvious. We use the following graphic to define the concepts:

Here \(S\) represents a set of possible ‘states’, and \(A\) and \(B\) subsets of that set.
Following Kolmogorov, we define probability as any real-valued function acting on the states of $S$, here $P(\cdot)$ satisfying the following conditions:

1. For every subset $A$ in $S$, $P(A) \geq 0$.
2. For disjoint subsets $A \cap B = \emptyset$, $P(A \cup B) = P(A) + P(B)$
3. $P(S) = 1$

Conditional probability (probability of $B$ conditioned on $A$; i.e., $A$ is given) is defined as follows:

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

Since $P(B \cap A) = P(A \cap B)$, we immediately get

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

which is known at Bayes’ Theorem. Furthermore, if we use the Law of Total Probability,:

$$ P(B) = \sum_{i=1}^{N} P(B|A_i)P(A_i) $$

where $A_i \cap A_j = \emptyset \quad i \neq j$ and $\sum_{i=1}^{N} A_i = S$, we have

$$ P(A|B) = \frac{P(A|B)P(B)}{\sum_{i=1}^{N} P(B|A_i)P(A_i)} $$

which we shall call the Bayes-Laplace Theorem since Laplace was the first to use Bayes’ Theorem in the analysis of scientific data. This formula is the basis for
1.2. SCIENTIFIC KNOWLEDGE

all scientific knowledge updates, where used explicitly (using Bayes Theorem) or implicitly (when individuals make decisions based on probabilities of data in the context of a model).

1.2 Scientific Knowledge

Our goal as scientists is to increase what we can call scientific knowledge. What is this scientific knowledge? We will define it as follows:

Scientific Knowledge = Justified Belief

Note that I leave out 'true' on the right hand side (a standard epistemologists definition is 'justified true knowledge') since we know we will never be able to prove our theories or models to be 'true'. For this, we would need to be able to rule out all possible other models that could result in the same set of observations that we have, and this seems inconceivable. 'Belief' itself sounds very weak to a scientist, since everyone can have beliefs about anything, so obviously the key is in 'justified'. Where does this justification come from? It comes from

1. having a logically consistent model that can explain known facts (measurements);
2. the ability to make predictions and test these against future experiments.

We will steadily increase our belief in the adequacy of models that make predictions that are testable and for which experimental results are in agreement with the predictions. Models that are untestable will have degrees of belief that vary widely across the scientific community - proponents of the models will generally proffer strong belief, whereas people outside the specific community will (rightly) be much more skeptical. Models that make predictions that are testable and that agree with the data will have strong and universally high degrees of belief.

Given several models that can explain all known data, it is often argued that we should use 'Occam's razor', and take the simplest of the models as the best. There are several arguments in favor of this approach:

- there is beauty in simplicity and we associate 'correctness' with beauty;
• it is easier to make predictions from simple models and they are more easily tested - complicated models often have little predictive power and are therefore not very useful;

• the history of physics has taught us that simpler ideas are more often correct (e.g., Ptolemaic vs Copernican systems).

However, these are all weak justifications and in general we can say that all models that explain are data are equally valid and the model selection should be based on their ability to correctly predict future data.

The models themselves often include parameters that have more or less well defined values, so that the scientific knowledge update occurs at several levels. The belief in the model as a whole can be updated, as well as the belief in possible values for particular parameters within the model. We will investigate how to use data for both of these tasks. In many cases, the models are not intended to have scientific content, but are simply ‘useful parameterizations’ of data that should however be useable in different contexts. I.e., they should serve as a way of transporting data from one experiment to be used as input knowledge in a different setting. In these cases, we do not test the belief in the model, but rather the adequacy as a representation of the data. We will discuss criteria for deciding whether such parameterizations of data are indeed adequate or not.

1.2.1 Updating Scheme

Our general scheme for updating knowledge is given in Fig. 1.1. On the left-hand side of the diagram, we display our model and work out its predictions. The model has the name $M$, and contains a set of fundamental parameters, $\vec{\lambda}$, that need to be specified to make quantitative predictions. The distributions expected from the model are often in the form of probability densities (e.g., the angular distribution expected in a scattering process, the energy distribution expected from radiation from a star, etc.) and we denote the predictions from the model with $g(\vec{y}|\vec{\lambda},M)$, where $\vec{y}$ are the ‘observables’ for which the model makes predictions (e.g., the angular distribution of the scattered particles, the energy distribution of photons, ).

We will want to compare the predictions of the model to data taken with some experimental setup. The experimental apparatus always has some limitations - it cannot have infinitely good resolution, the angular acceptance is limited, it has a limited speed at which it can accumulate data, etc., and these limitations imply that the observed distributions cannot be given by $g(\vec{y}|\vec{\lambda},M)$ but are distorted
1.2. SCIENTIFIC KNOWLEDGE

Figure 1.1: General scheme for knowledge updates.
versions. An extra step is necessary in making our predictions for what we expect to observe - the modeling of the experimental conditions. This modeling is often performed via computer simulations of the experimental setup, leading to predictions for frequency distributions of observable data. This modeling often relies on parameters that also need to be specified - these are denoted by $\vec{\nu}$ in the figure, and the predicted distribution for the quantities that are to be measured is given by $f(\vec{y}|\vec{\lambda}, \vec{\nu}, M)$. The actual values of $\vec{\nu}$ are often not of interest, and these parameters are called 'nuisance parameters'.

Given this information, it is now possible to compare the experimental data, denoted here by $\vec{D}$, with the predictions.

### 1.2.2 Schools of Data Analysis

The scientific community is broadly divided into two schools regarding how to analyze the experimental data. In the Bayesian approach, Bayes’ theorem is used to form a degree of belief in a model or learn about the probable values of parameters as outlined above. In the 'Classical' school, no prior information is used, only the predicted frequency distribution of observable results from the model and the observed data. In this case, statements of the kind 'given model $M$ and parameters $\vec{\lambda}$, the observed data is in the central 68 % range of expected outcomes'. Parameter ranges can then be found, e.g., where the observed data is within a defined interval of expected results. This type of statement can be viewed as a summary of the data in the context of the model, but does not allow conclusions about the probable values of the parameters themselves. The parameters are said to have 'true values' and the data should lead us to the truth.

Many aspects of the analysis are the same in a 'Classical' analysis of the data as for a Bayesian analysis. The formulation of the model and the modeling of the experimental setup are required for both approaches. The difference lies in the fact that in the Bayesian analysis, additional information - the 'prior information' - is used in the evaluation of our belief in the model $M$ and the probable values of the parameters. Bringing in the extra information 'biases' the results in the sense that our conclusions are not based entirely on the data from the single experiment. This is of course desirable if the goal is to make the best possible statements concerning the model or its parameter values, but if the result is to be interpreted solely as a summary of the data from a particular experiment, then this misunderstanding can lead to incorrect conclusions.

Since misunderstanding abounds concerning the conceptual foundations of the Bayesian and Classical approaches, we will analyze many examples from both
1.2. SCIENTIFIC KNOWLEDGE

perspectives and discuss the meaning of the results. We note that there is no ‘right’ and ‘wrong’ approach, but rather ‘right’ and ‘wrong’ interpretations of the numerical results that are produced, assuming of course that no errors were made in the process.

1.2.3 Which probability ?

What we call the probability of the data can take on many different values for the same data, depending on how we choose to look at it. For example, let us compare the probability we assign to an experimental result where we flip a coin ten times and have either

\[ S_1 : \text{THTHHTHTTH} \]

or the result

\[ S_2 : \text{TTTTTTTTTT} \]

where \( T \) represents 'tails' and \( H \) 'heads'. Which of these results do we consider more probable? Or do they have the same probability?

For this example, suppose we are justified in assuming that the coin has equal probability to come up heads and tails - we can still come up with many different probabilities for these data outcomes. One choice is the probability that we get the given sequence. Since for each flip we have probability \( \frac{1}{2} \) to get heads or tails, the probability of the two sequences are the same

\[
P(S_1|M) = (1/2)^{10} \quad P(S_2|M) = (1/2)^{10}.
\]

However, we can also calculate the probability to get the number of \( T \) in the sequence:

\[
P(S_1|M) = \binom{10}{5} (1/2)^{10} = 252 \cdot (1/2)^{10} \\
P(S_2|M) = \binom{10}{10} (1/2)^{10} = 1 \cdot (1/2)^{10}
\]

and we find that \( S_1 \) is 252 times more likely than \( S_2 \). We have the same data, but in one case we specified a specific order of the outcomes and in the second case only counted how often each outcome occurred, and find very different probabilities. This will in general be true - it will be possible to analyze the same data but
assign very different probabilities within the context of the models. The choice of which probability to use is a very important aspect of any data analysis and will require case-by-case analysis of the situation.

1.3 Warm-up Exercises

We will finish this introductory chapter with some exercises to familiarize ourselves with probabilistic reasoning.

1.3.1 Apparatus and Event Type

Let’s start with the following situation: imagine you have an apparatus that will give a positive signal 95 % of the time when an event of type \( A \) occurs, while it gives a positive signal 2 % of the time when an event of type \( B \) occurs, and otherwise gives no signal. I.e., we can write

\[
P(signal|A) = 0.95 \quad P(signal|B) = 0.02 \quad P(signal|A \text{ or } B) = 0.0
\]

and we perform a measurement on some source to determine if it produces events of type \( A \) or \( B \) (we know it should be one or the other). Let’s say we find that we get a positive signal. What can we now conclude?

The answer is, very little, unless you provide more information. In a Bayesian analysis, you would write:

\[
P(A|signal) = \frac{P(signal|A)P_0(A)}{P(signal|A)P_0(A) + P(signal|B)P_0(B)}
\]

It is necessary to have some information about the probability of \( A \) and \( B \) prior to the measurement to draw some conclusion. Suppose that we don’t know anything about these probabilities, and we think it is reasonable to give them equal initial probability. I.e., we set

\[
P_0(A) = P_0(B) = 1/2
\]

Then we find

\[
P(A|signal) = \frac{0.95}{0.95 + 0.02} = 0.98
\]

and we conclude with 98 % probability that our source is of type \( A \).
If we had instead some prior knowledge and knew that the initial probability for $B$ was 1000 greater than for $A$:

$$P_0(A)/P_0(B) = 10^{-3}.$$ 

then we would find

$$P(A|\text{signal}) = \frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.02} = 0.045.$$ 

I.e., the conclusion depends very much on what is known about the initial probabilities that we assign to the possible sources.

Suppose that we make nine more measurements and all of them give us a positive signal. What can we say about the probability that our source of events is of type $A$ (we assume all events are either of type A or B)? In this case, we can write the probability of our data given the models as

$$P(\text{signals}|A) = 0.95^{10} \approx 0.6 \quad P(\text{signals}|B) = 0.02^{10} \approx 10^{-17}$$

and the results for our two different prior assumptions are effectively identical

$$P(A|\text{signals}) \approx 1.$$ 

I.e., if the data are very powerful, the prior choice carries little weight in forming the final probability.

### 1.3.2 Efficiency of Apparatus

Now suppose the task is to determine the efficiency of our apparatus for generating a signal for different types of particles. In this case, we know exactly what type of particle is aimed at the detector, but do not know the efficiency for the detector to record a signal. Imagine we shot $N = 100$ particles at our detector, and found a positive response $s = 70$ times. What can we say about the efficiency? Obviously, it should be a number near 0.70. What is the uncertainty on this number? Often, a rule-of-thumb is to use $\sqrt{s}/N$ to estimate the uncertainty on the efficiency. We will see where these rules-of-thumb come from, and also when they apply. Consider a not uncommon case, where $N = 100$ and $s = 98$. What are the efficiency and uncertainty on the efficiency in this case. If we use

$$\epsilon = \frac{s}{N} \pm \frac{\sqrt{s}}{N} = 0.98 \pm 0.10$$
then the result looks strange. The possible values of $\epsilon$ only range up to 1.0, so the uncertainty as quoted above is clearly meaningless. This shows that the rule-of-thumb approach should be used with caution, and you should know when it applies.

### 1.3.3 Exercize in Bayesian Logic

You are presented with three boxes which contain two rings each. In one of them both are gold, in a second one is gold and one is silver, and in the third both are silver. You choose a box at random and pick out a ring. Assume you get a gold ring. You are now allowed to pick again. What is the best strategy to maximize the chance to get a second gold ring?

This is a case where our knowledge increases after the first experiment (opening the first box) and we therefore update our probabilities.

Let us define box $A$ as having the two golden rings, box $B$ as having one gold and one silver ring, and box $C$ as having two silver rings. Given no other information, it is reasonable to assign each a probability of $1/3$ that we will initially pick them. I.e., we have

$$P(A) = P(B) = P(C) = \frac{1}{3}.$$  

We picked a box, not knowing whether it is $A, B$ or $C$. We take a ring out, and find it is golden. We call this event $E$. We use Bayes’ Theorem to update our probabilities for the chosen box:

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|B)P(B) + P(E|C)P(C)}$$

$$= \frac{1 \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{2}{3}.$$  

Similarly, we find $P(B|E) = 1/3$ and $P(C|E) = 0$. Given this new information, we would therefore increase the probability that we assign to our box being $A$, and would decide to stick with our initial choice.

This example shows what is meant by ’Probability does not Exist’\(^1\). The probability we assign is a number we choose, and is not a property of any object. It has no existence in itself. Of course, thinking about Quantum Mechanics makes this statement even more interesting. For a fun read, please look at http://www.nature.com/news/physics-qbism-puts-the-scientist-back-into-science-1.14912.

---

\(^1\) de Finetti, Theory of Probability
1.3. WARM-UP EXCERCIZES

Literature

There is a huge literature for data analysis techniques, usually either explaining the frequentist approach or the Bayesian approach. In principle, these lecture notes are intended to be self-contained and provide an introduction to both schools of thought. The aim is that, by putting them side-by-side, the student can see for her/himself what these techniques are based on and what they say.

Nevertheless, it can often be useful to see how others describe similar situations, so here some texts with Bayesian and Classical perspectives. You can figure out which is which.


Exercises

1. You meet someone on the street. They tell you they have two children, and has pictures of them in his pocket. He 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: person takes out both pictures, looks at them, and is required to show you a picture of a girl if he has one. What is now the probability that the second child is also a girl ?

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

3. 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 ?
4. 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 Mongolian swamp fever if you go to the doctor with all or with any three out of four of these symptoms? (From R.Barlow)
Chapter 2

Binomial and Multinomial Distribution

2.1 Binomial Distribution

The Binomial distribution is appropriate when we have the following conditions:

- Two possible outcomes, with fixed probabilities \( p \) and \( q = 1 - p \);
- A fixed number of trials.

A typical situation is when you analyze the efficiency for some apparatus. E.g., you want to see how often your detector responds when a certain type of particle impinges on it, so you go to a 'test beam' where you count how many particles entered, \( N \), and count how often your detector responded, \( r \). Another example is that you observe a source of events that can be of two types, \( A \) and \( B \), and you count a fixed number of trials (e.g., flipping a coin). We define

\[
\begin{align*}
p & \quad \text{probability of a success} \quad 0 \leq p \leq 1 \\
N & \quad \text{independent trials} \quad 0 \leq N < \infty \\
r & \quad \text{number of successes} \quad 0 \leq r \leq N
\end{align*}
\]

The probability that we assign for \( r \) given \( N \) and \( p \) is

\[
P(r|N,p) = \frac{N!}{r!(N-r)!} p^r q^{N-r} \quad (2.1)
\]
where the Binomial factor
\[
\binom{N}{r} = \frac{N!}{r!(N-r)!}
\]
gives the number of possible ways to get \( r \) successes in \( N \) trials (i.e., here we do not consider the order of the outcomes). This probability assignment corresponds to the relative frequency for getting \( r \) successes in the limit that we could repeat the experiment (the \( N \) trials) an infinite number of times. Since we have a mathematical model, we can carry out this calculation of this relative frequency and it is our best choice for the degree-of-belief to assign to a particular outcome under our assumptions (\( N, p \) fixed and known). We call these types of probability assignment - relative frequencies from our mathematical model - **direct probabilities**.

### 2.1.1 Some Properties of the Binomial Distribution

The following can be (in some cases) easily proven (please see the Appendix for the notation). The Binomial distribution is normalized:

\[
\sum_{r=0}^{N} P(r|N,p) = 1
\]  
\( (2.2) \)

The expectation value for \( r \) is \( Np \):

\[
E[r] = \sum_{r=0}^{N} r P(r|N,p) = Np
\]  
\( (2.3) \)

The variance of \( r \) is \( Np(1-p) \):

\[
V[r] = \left[ \sum_{r=0}^{N} r^2 P(r|N,p) \right] - E[r]^2 = Np(1-p)
\]  
\( (2.4) \)

The mode of \( r \) (value with highest probability is):

\[
r^* = 0 \text{ for } p = 0 \text{ and } r^* = N \text{ for } p = 1
\]  
\( (2.5) \)

or, for the most common case

\[
r^* = \lfloor (N+1)p \rfloor \text{ for } p \neq 0,1 \text{ and } (N+1)p \notin \{1, \ldots, N\}
\]  
\( (2.6) \)

or, for the special case that \( (N+1)p \) is an integer, there is a double mode:

\[
r^* = \lfloor (N+1)p \rfloor \text{ and } \lfloor (N+1)p \rfloor - 1 \text{ for } (N+1)p \in \{1, \ldots, N\}
\]  
\( (2.7) \)
2.1. BINOMIAL DISTRIBUTION

2.1.2 Example

Consider the simple example: \( N = 5, p = 0.6 \). We calculate our direct probabilities and give them in Table 2.1.

Table 2.1: Probabilities assigned to possible outcomes for a Binomial model with \( N = 5 \) and \( p = 0.6 \). In addition to the value of \( r \) and its probability, we also give the cumulative probability starting from \( r = 0 \), the ordering of the outcomes according to a Rank defined by the probabilities of the outcomes, and a cumulative probability according to the Rank.

| \( r \) | \( P(r | N = 5, p = 0.6) \) | \( F(r | N = 5, p = 0.6) \) | Rank | \( F(\text{Rank}) \) |
|---|---|---|---|---|
| 0 | 0.0102 | 0.0102 | 6 | 1 |
| 1 | 0.0768 | 0.0870 | 5 | 0.9898 |
| 2 | 0.2304 | 0.3174 | 3 | 0.8352 |
| 3 | 0.3456 | 0.6630 | 1 | 0.3456 |
| 4 | 0.2592 | 0.9222 | 2 | 0.6048 |
| 5 | 0.0778 | 1 | 4 | 0.9130 |

From this table, we see that the mode is at \( r^* = 3 \) which is as expected \( (r^* = \lfloor (N + 1)p \rfloor) \). We can create various intervals based on probability requirements. This will be a warm-up for the discussion of Confidence Intervals. To this end, we define the Central Probability Interval and Smallest Probability Interval containing a probability at least \( 1 - \alpha \).

2.1.3 Central Interval

This interval is constructed by defining a smallest member of the set of possible outcomes, \( r_1 \), and a largest member, \( r_2 \), subject to the criterion that

\[
P(r < r_1) \leq \alpha/2
\]

and

\[
P(r > r_2) \leq \alpha/2
\]

while maximizing these probabilities. Stating this precisely, we write
\[ r_1 = 0 \text{ if } P(r = 0|N, p) > \alpha/2 \text{ else } \]  
\[ r_1 = \sup_{r \in 0, \ldots, N} \left[ \sum_{i=0}^{r} P(i|N, p) \leq \alpha/2 \right] + 1 \]  
\hspace{1cm} (2.8) 
\[ r_2 = N \text{ if } P(r = N|N, p) > \alpha/2 \text{ else } \]  
\[ r_2 = \inf_{r \in 0, \ldots, N} \left[ \sum_{i=r}^{N} P(i|N, p) \leq \alpha/2 \right] - 1 . \]  
\hspace{1cm} (2.9)

The set of results in our Central Interval is
\[ \mathcal{O}_{1-\alpha}^C = \{r_1, r_1 + 1, \ldots, r_2\} . \]

Choosing \( 1 - \alpha = 0.68 \) and considering the numbers in Tables 2.1, we find \( \mathcal{O}_{0.68}^C = \{2, 3, 4\} \). The total probability contained in this set of results is \( F(r = 4) - F(r = 1) = 0.8352 \), which means that the interval over covers; i.e., it contains more probability than the minimum that we have set.

### 2.1.4 Smallest Interval

The second option that we consider for the probability interval is to use the smallest interval containing at least a given probability. For our discrete case, the set making up the smallest interval containing probability at least \( 1 - \alpha \), \( \mathcal{O}_{1-\alpha}^S \), is defined by the following algorithm

1. Start with \( \mathcal{O}_{1-\alpha}^S = \{r^*\} \). If \( P(r^*|N, p) \geq 1 - \alpha \), then we are done. An example where this requirement is fulfilled is, for \( 1 - \alpha = 0.68 \), when \( (N = 1, p = 0.001, r^* = 0, \mathcal{O}_{0.68}^S = \{0\}) \).

2. If \( P(r^*|N, p) < 1 - \alpha \), then we need to add the next most probable number of observations. I.e., we keep adding members to our set according to their Rank, with the Rank defined by the value of the probability of the outcome. Our set would then contain \( \mathcal{O}_{1-\alpha}^S = \{r(\text{Rank} = 1), r(\text{Rank} = 2)\} \). We would continue to add members to the set \( \mathcal{O}_{1-\alpha}^S \) until the condition

\[ P(\mathcal{O}_{1-\alpha}^S) \geq 1 - \alpha \]
Choosing again $1 - \alpha = 0.68$ and considering the numbers in Tables 2.1, we find $\mathcal{O}^{S}_{0.68} = \{2, 3, 4\}$; i.e., the same result in this case as for the Central Interval.

### 2.2 Neyman Confidence Level Construction

We have seen how to define intervals that contain at least a minimum probability. Our intervals depended on the value of $N$ as well as $p$. Suppose we do not know the value of $p$, and want to see how our intervals depend on this chosen value. We can repeat the exercise above for all possible values of $p$ ($0 \leq p \leq 1$) and see how our interval evolves. As a concrete example, consider Fig. 2.1, constructed for $N = 5$ and $1 - \alpha = 0.68$. From the figure, we see that for very small $p$, say $p = 0.02$, the only member of our Central Interval is $r = 0$. For $p = 0.6$, we find $\mathcal{O}^{C}_{0.68} = \{2, 3, 4\}$ as calculated previously. Neyman has defined a procedure based on such a band plot to define Confidence Levels for the parameter $p$ for a given observation.

The Neyman procedure is as follows:

1. Construct the band plot for the specified $1 - \alpha$ and $N$ as in our example.
   You first need to decide how to define the interval (Central Interval, Smallest Interval, ).

2. Perform your experiment and count the number of successes. Let’s call this $r_D$.

3. For the measured $r_D$, find the range of $p$ such that $r_D \in \mathcal{O}^{C}_{0.68}$.

4. The resulting range of $p$ is said to the the $1 - \alpha$ Confidence Level interval for $p$.

The Neyman Confidence Level interval specifies a range of $p$ for which $r_D$ is within the specified probability interval for outcomes $r$. Imagine there is a ‘true value’ for $p$. Then, in a large number of repetitions of our experiment, the observed $r_D$ would be one of the members of $\mathcal{O}^{C}_{0.68}$ for the true $p$ in at least 68 % of our experiments. Since the Confidence Level band for $p$ would then include the true value in these cases, we can conclude that our CL intervals will contain the true value of $p$ in at least $1 - \alpha$ of a large number of repetitions of our experiment.
Figure 2.1: 68% Central Interval bands for a Binomial probability distribution with $N = 5$. The red line gives the value of $r_1$, the smallest element in the set, while the blue bands gives $r_2$, the largest element in the set.
2.3. LIKELIHOOD

2.2.1 Example

Imagine that we take $N = 5$ and $1 - \alpha = 0.68$, and we perform an experiment and observe $r_D = 0$. From Fig. 2.1, we find $0 \leq p \leq 0.31$. For these values of $p$, the result $r_D = 0$ is in the 68% Central Interval. This is the conclusion that we can give. I.e., we have transformed the observed number of events into the language of our probabilistic model. If the value of $p$ is related to a model parameter, then we can further translate our experimental result into a statement in the language of the model. E.g., if $p$ represents an efficiency of a detector, we can say that for efficiencies lower than 31%, the observed number of successes is within the 68% Central Interval. Note that we only report on the probability of the data within the context of the model. There is no statement possible concerning what we think are the preferred values of the parameter. For this, one needs to perform a Bayesian analysis (we will get there soon).

Suppose we then perform a second experiment and find $r_D = 2$. In this case, we would find, at 68% CL, that $0.15 < p < 0.72$. I.e., for these values of $p$, our measurement lies within the 68% Central Interval. How would we combine this information with that from the previous measurement? This is in general an unsolved problem. If the experimental data can be added (as in our simple example), then we could perform one CL analysis based on $N = 10$ trials and $r_D = 2$. However, in most cases when experiments are carried out by different teams this combination of data is not possible. In the Neyman scheme, what one should rather do is collect a large number of confidence intervals and somehow intuit the true value from this connection. Exactly how to do this is not specified.

2.3 Likelihood

Figure 2.2 shows the probability distribution for $r$ for selected values of $p$ and for $N = 5$. For fixed $p$, we have the probability of different data outcomes. If we instead consider $r$ fixed and let $p$ vary, then we have a function of $p$. For example, $P(r = 3|N = 5, p)$ can be viewed as a function of $p$, the probability that we assign to $N = 5, r = 3$ for different values of $p$. Viewed in this way, the probability of the data is called the likelihood and often given the symbol $\mathcal{L}$. I.e.,

$$\mathcal{L}(p) = P(r = 3|N = 5, p).$$

Note that the likelihood is not a probability density for $p$, however, and it is not properly normalized. I.e., when working with the likelihood, we are dealing
2.4 Limit of the Binomial Distribution

Figure 2.3 shows the probability distribution for $r$ for selected values of $p$ and for $N = 50$. We observe that the probability distribution of $r$ is now rather symmetric for all values of $p$. Indeed, the probability distribution approaches a Gaussian shape in the limit $N \to \infty$ and $p$ not too close to zero. This is demonstrated below.

2.4.1 Likelihood function for a Binomial Model

The likelihood function has the shape

$$ \mathcal{L}(p) \propto p^r (1 - p)^{N-r} $$
with the mode of the likelihood function given by

\[ p^* = \frac{r}{N}, \]

the expectation value given by

\[ E[p] = \frac{r + 1}{N + 2} \]

and the variance

\[ V[p] = E[(p - E[p])^2] = \frac{E[p](1 - E[p])}{(N + 3)}. \]

The likelihood function becomes symmetric when \( r \) and \( N \) are both large. We can show this by calculating the skew of the likelihood:

\[ \gamma = \frac{E[(p - E[p])^3]}{(V[p]^{3/2})} = \frac{2(N - 2r)}{N + 4} \sqrt{\frac{N + 3}{(r + 1)(N - r + 1)}} \]

which is zero for \( r = N/2 \) as we would expect. If we take the limit \( r \to \infty \) (implying \( N \to \infty \)), then we find

\[ \gamma \to \frac{2(N - 2r)}{\sqrt{Nr(N - r)}} \to 0 \]

and the likelihood function becomes symmetric. Given that this limit is valid, we can also be tempted to use a Gaussian approximation for the likelihood function, with a mean \( \mu = r/N \) and standard deviation \( \sigma = \sqrt{r(N - r)/N^3} \). Note that this approximation should be used only in appropriate circumstances - i.e., when \( N \) and \( r \) are both large.

### 2.4.2 Gaussian Approximation for the Binomial Probability

(This section can be skipped w/o further consequences)

We first recall Stirling’s approximation for the natural logarithm of the factorial:

\[ \ln n! \approx \ln \sqrt{2\pi n} + n \ln n - n \]

or

\[ n! \approx \sqrt{2\pi n} \left( \frac{n}{e} \right)^n \]
and substitute this into the Binomial probability distribution

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

\[ \approx \frac{\sqrt{2\pi N}(N/e)^N}{\sqrt{2\pi r}(r/e)^r \sqrt{2\pi(N-r)((N-r)/e)^{N-r}}} p^r (1-p)^{N-r} \]

\[ = \frac{1}{\sqrt{2\pi N}} \left( \frac{r}{N} \right)^{-r-1/2} \left( \frac{N-r}{N} \right)^{-N+r-1/2} p^r (1-p)^{N-r} \]

We now make a change of variables, and work with the quantity \( \xi = r - Np \). We know the variance of the values of \( r \), \( V[r] = Np(1-p) \) so that typical values of \( \xi \) will be of order \( \sqrt{N} \) and for large \( N \), \( \xi/N \approx 1/\sqrt{N} \to 0 \). So we can use series approximations such as

\[ \ln(1+x) \approx x - \frac{1}{2} x^2 \quad \ln(1-x) \approx -x - \frac{1}{2} x^2 \]
Let us rewrite our expression for $P(r|N,p)$ in terms of $\xi$. We need
$$
\left( \frac{r}{N} \right)^{-r-1/2} = (p + \xi/N)^{-r-1/2} = p^{-r-1/2} \left( 1 + \frac{\xi}{Np} \right)^{-r-1/2}
$$
and
$$
\left( \frac{N-r}{N} \right)^{-N+r-1/2} = (1-p)^{-N+r-1/2} \left( 1 - \frac{\xi}{N(1-p)} \right)^{-N+r-1/2}
$$
and we substitute in our expression for the probability of $r$ to get
$$
P(r|N,p) \approx \frac{1}{\sqrt{2\pi Np(1-p)}} \left( 1 + \frac{\xi}{Np} \right)^{-r-1/2} \left( 1 - \frac{\xi}{N(1-p)} \right)^{-N+r-1/2}.
$$
We now use an exponential form so that we can use our approximations for the natural log given above:
$$
P(r|N,p) \approx \frac{1}{\sqrt{2\pi Np(1-p)}} \exp \left[ (-r - 1/2) \ln \left( 1 + \frac{\xi}{Np} \right) + (-N + r - 1/2) \ln \left( 1 - \frac{\xi}{N(1-p)} \right) \right]
$$
We expand the piece in the brackets and get
$$
(-Np - \xi - 1/2) \left( \frac{\xi}{Np} - \frac{1}{2} \left( \frac{\xi}{Np} \right)^2 \right) + (-N(1-p) + \xi - 1/2) \left( -\frac{\xi}{N(1-p)} - \frac{1}{2} \left( \frac{\xi}{N(1-p)} \right)^2 \right).
$$
In the limit that $p$ is not too small and not too big, we can keep only terms up to $\xi^2$ and get
$$
-Np \left( \frac{\xi}{Np} - \frac{1}{2} \left( \frac{\xi}{Np} \right)^2 \right) - N(1-p) \left( -\frac{\xi}{N(1-p)} - \frac{1}{2} \left( \frac{\xi}{N(1-p)} \right)^2 \right) = -\frac{1}{2} \frac{\xi^2}{Np(1-p)}
$$
so
$$
P(r|N,p) \approx \frac{1}{\sqrt{2\pi Np(1-p)}} \exp \left( -\frac{1}{2} \frac{(r - Np)^2}{Np(1-p)} \right).
$$
So, we see that we can use the formula for the Gaussian distribution to approximate the probability of getting $r$ successes in $N$ trials in the limit where both $r$ and $N$ are large. Whether or not the Gaussian approximation is appropriate depends on how precise your calculations needs to be. In the exercises, you get to compare some values. Here just a couple examples:
N = 10, r = 3, p = 0.5 \quad P(r = 3|N = 10, p = 0.5) = 0.117 \quad G(3, \mu = 5, \sigma^2 = 2.5) = 0.183

N = 100, r = 35, p = 0.5 \quad P(r = 35|N = 100, p = 0.5) = 8.6 \cdot 10^{-4} \quad G(35, \mu = 50, \sigma^2 = 25.) = 8.9 \cdot 10^{-4}

2.5 Extracting Knowledge using Bayes Equation

We now turn from describing the Binomial distribution itself and the associated Likelihood function to the extraction of information about the parameter in the Binomial Distribution, p. I.e., we have now performed an experiment with N trials and found r successes. To now learn about the parameter p, we use Bayes’ Theorem:

\[
P(p|N, r) = \frac{P(r|N, p)P_0(p)}{P(D)} = \frac{\int P(r|N, p)P_0(p)dp}{\int P(r|N, p)P_0(p)dp}
\]

To go further, we need to specify a prior probability function \(P_0(p)\). This is of course situation dependent, and therefore we can only consider examples here. For any data analysis based on a ‘real-life’ situation, the prior probability should be evaluated on a case-by-case basis. While this may seem frustrating, it is very important to understand that there is no alternative to this. Any analysis that wants to claim information on the parameters of the model via data analysis must use Bayes’ Theorem and therefore must specify priors. As discussed above, ‘Classical Approaches’ such as the Neyman Confidence Intervals do not provide statements on probable values of the parameters but are rather restatements of the data in the language of the model under study. They are often confused with probability statements about the parameters, but they are not, and we will not make this logical mistake here.

2.5.1 Flat Prior

So, we need to specify a prior probability and we will choose a few examples to understand the features of data analysis in more detail. We start by taking a ‘flat prior’, meaning that we assign the same probability to all values of p. I.e., we have
no reason to give a preference to any values or range of values. This may or may not be a reasonable choice, and, again, depends on the specifics of the problem at hand. Let’s see what happens if we make the choice. Since the possible values for \( p \) are \([0, 1]\), we have for our normalized probability density

\[
P_0(p) = 1
\]

Bayes equation for the Binomial model now reduces to

\[
P(p|N, r) = \frac{p^r(1 - p)^{N-r}}{\int_0^1 p^r(1 - p)^{N-r} dp}
\]

where the integral in the denominator is a standard Beta Function

\[
\beta(r + 1, N - r + 1) = \int_0^1 p^r(1 - p)^{N-r} dp = \frac{r!(N - r)!}{(N + 1)!}
\]

yielding

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

We now have a probability distribution for \( p \) given the experimental result and can analyze this further. We notice that the function looks very much like the Binomial probability distribution we started with, except that now we have a function of \( p \) rather than a function of \( r \). Also, the numerator is different by a factor \((N + 1)\). We easily see that the mode of this function occurs at \( p^* = r/N \). Note that this is the same result as for the mode of the Likelihood, and this is no accident. There is a simple relationship between a Bayesian analysis with flat prior probabilities and a Likelihood analysis, as described below. We calculate the mean and variance for \( p \) given this probability density:

\[
E[p] = \int_0^1 p P(p|N, r) dp = \int_0^1 \frac{(N + 1)!}{r!(N - r)!} p^{r+1}(1 - p)^{N-r} dp
\]

We use the following property of the \( \beta \) function to solve the integral

\[
\beta(x + 1, y) = \beta(x, y) \cdot \frac{x}{x + y}
\]

which gives

\[
E[p] = \frac{r + 1}{N + 2}
\]
To get the variance, we apply the relation twice and get
\[ E[p^2] = \frac{r + 1}{N + 2} \frac{r + 2}{N + 3} \]
yielding
\[ V[p] = E[p^2] - E[p]^2 = \left( \frac{r + 1}{N + 2} \right) \left( \frac{N - r + 1}{N + 3} \right) = \frac{E[p](1 - E[p])}{N + 3} \]

We note that these formulae are also valid for \( N = 0, r = 0 \), and in this case we get the mean and variance of \( p \) from the prior distribution.

To get a feel for this function, we plot it in Fig. 2.4 for specific examples. The posterior probability density is plotted for four different data sets \( \{N = 3, r = 1\}, \{N = 9, r = 3\}, \{N = 30, r = 10\}, \{N = 90, r = 30\} \). A few points can be made about the plots:

- As the amount of data increases, the probability distribution for \( p \) narrows, indicating that we can make more precise statements about its value;
- The distribution becomes more symmetric as the amount of data increases, and at some point can be accurately reproduced with a Gaussian distribution.

### 2.5.2 Defining Probable Ranges

The posterior probability distribution contains all the information that we have about \( p \). However, it is sometimes convenient to summarize this knowledge with a couple numbers. An obvious choice would be to use the mean and variance of the distribution, which are just the first moment and second (central) moment of the function. However, we can make other choices which are also consistent based on probability content. We consider here the pairs (median and central interval) and (mode and smallest interval).

#### Median and Central Interval

The median is the value of \( p \) for which the cumulative probability reaches 50 %. I.e., it is defined via
\[
F(p_{\text{med}}) = \int_0^{p_{\text{med}}} P(p|N, r)dp = \int_{p_{\text{med}}}^1 P(p|N, r)dp = 0.5
\]
We can then define a central probability region as the region containing a fixed probability $1 - \alpha$ and defined such that there is equal probability in the low and high tails. I.e.,

$$F(p_1) = \int_0^{p_1} P(p|N, r) dp = 1 - F(p_2) = \int_{p_2}^1 P(p|N, r) dp = \frac{\alpha}{2}.$$ 

The interval $[p_1, p_2]$ is called the Central Probability Interval and contains probability $1 - \alpha$. We would then say that the parameter $p$ is in this interval with $1 - \alpha$ probability. Note that this is a direct statement about the probable values of the parameter $p$, as opposed to the Confidence Level interval.

**Mode and Shortest Interval**

Another logical pair or numbers that can be used to summarize the probability density for $p$ are the mode and the shortest interval containing a fixed probability. The mode is the value of $p$ that maximizes $P(p|N, r)$ and in general depends on the prior distribution chosen. For a flat prior, it is $p^* = r/N$. The shortest interval
for a unimodal distribution is defined as \([p_1, p_2]\) with the conditions

\[
1 - \alpha = \int_{p_1}^{p_2} P(p|r, N) dp \quad P(p_1|r, N) = P(p_2|r, N) \quad p_1 < p^* < p_2
\]

**Graphical Summary**

These quantities are graphically demonstrated in Fig. 2.5 for the function

\[
P(x) = \frac{1}{2} x^2 e^{-x} \quad 0 \leq x < \infty
\]

The mode of this function is at \(x^* = 2\), the mean is \(E[x] = 3\), and the variance is \(V[x] = 3\). To find the median, central interval and smallest interval, we need to use numerical techniques. Let’s first write down the cumulative probability:

\[
F(x) = \int_0^x \frac{1}{2} x^2 e^{-x'} dx' = \frac{1}{2} \left[ 2 - (x^2 - 2x - 2) e^{-x} \right]
\]

To find the median, we set \(F(x_{med}) = 0.5\) and find \(x_{med} \approx 2.68\). To find the 68 % central interval, we solve using the cumulative and find \(x_1 \approx 1.37\) and \(x_2 \approx 4.65\).

To find the smallest interval, we need to solve

\[
F(x_2) - F(x_1) = 1 - \alpha = \frac{1}{2} \left[ (x_1^2 + 2x_1 + 2) e^{-x_1} - (x_2^2 + 2x_2 + 2) e^{-x_2} \right]
\]

under the condition \(P(x_2) = P(x_1)\). This simplifies our expression to

\[
1 - \alpha = \left[ (x_1 + 1) e^{-x_1} - (x_2 + 1) e^{-x_2} \right].
\]

Taking again the 68 % interval, we find (numerically) for this case \(x_1 \approx 0.86, x_2 \approx 3.86\).

These values are shown in Fig. 2.5. We note that the ’point estimate’ and the intervals are quite different for the three definitions chosen. The moral is that it is often not sufficient to try to summarize a functional form with two numbers. For symmetric unimodal distributions, the Central and Smallest Intervals will be the same, and the Median, Mean and Mode will also be the same. However, in general the interval defined by \(\pm \sigma\) where the Standard Deviation \(\sigma = \sqrt{V[x]}\) does not have a well-defined probability content. For the Gaussian distribution, the intervals defined by \(\pm \sigma\) is the same as the 68 % Central and Smallest Intervals.
Figure 2.5:
2.5.3 Nonflat Prior

In the general case, we have a complicated prior and do not have a standard set of formulas that can be used, and numerical approaches are needed to solve for the posterior probability density. However, let us see what happens if we start with a flat prior and perform two successive experiments meant to give us knowledge about the same Binomial parameter. After the first experiment, we have, as derived above,

$$P(p|N_1, r_1) = \frac{(N_1 + 1)!}{r_1!(N_1 - r_1)!} p^{r_1}(1 - p)^{N_1-r_1}$$

where now the subscripts on $N, r$ indicate that these are from the first measurement. We can now use this as the Prior probability density in the analysis of the data from the second experiment. I.e., we have

$$P(p|N_2, r_2) = \frac{P(r_2|N_2, p)P_0(p)}{\int P(r_2|N_2, p)P_0(p)dp} = \frac{P(r_2|N_2, p)P(p|N_1, r_1)}{\int P(r_2|N_2, p)P(p|N_1, r_1)dp}.$$ 

This gives:

$$P(p|N, r) = \frac{p^{r_2}(1 - p)^{N_2-r_2}p^{r_1}(1 - p)^{N_1-r_1}}{\int p^{r_2}(1 - p)^{N_2-r_2}p^{r_1}(1 - p)^{N_1-r_1}dp} = \frac{p^{r_1+r_2}(1 - p)^{N_1+N_2-r_1-r_2}}{\int p^{r_1+r_2}(1 - p)^{N_1+N_2-r_1-r_2}dp}.$$

Defining $r = r_1 + r_2$ and $N = N_1 + N_2$, we see that we are back to the same equation as when we analyzed the flat prior case. We can then write down the answer as

$$P(p|N_1, N_2, r_1, r_2) = \frac{(N_1 + N_2 + 1)!}{(r_1 + r_2)!(N_1 + N_2 - r_1 - r_2)!} p^{r_1+r_2}(1 - p)^{N_1+N_2-r_1-r_2}$$

and we note that we have exactly the same answer as if we have first added the data sets together and analyzed them in one step starting from the flat Prior distribution. This result is in general true for cases where the data sets are not correlated - the Bayes formulation for knowledge update is consistent in the sense that 'equal states of knowledge' yield 'equal probability assignments'.

We note also that if we use a Beta distribution for a Prior Probability, we get out another Beta distribution as posterior probability. We say that the Beta distributions provide a family of conjugate priors for the Binomial distribution.
2.5. EXTRACTING KNOWLEDGE USING BAYES EQUATION

2.5.4 A Detailed Example

Imagine that you perform a test of some equipment to determine how well it works. You want to extract a success rate. In a first test, you perform $N = 100$ trials and find $r = 100$ successes. In a second test of the same equipment, and under very similar conditions, you again perform $N = 100$ trials and this time find $r = 95$ successes. The description of the setup indicates that we should use a Binomial distribution to calculate the probability of the data. Since the equipment is the same and the conditions, as far as you can tell, identical, you assume that the success probability parameter, $p$, should be the same for both experiments. Here are some possible questions:

- what can we say about the efficiency after the first experiment?
- what can we say after the second experiment?
- what happens if we combine the data and treat them as one data set?
- are the results compatible with each other?
- can we tell if our modeling is correct or adequate?

First Data Set

Let’s start with a flat prior and analyze the first data set. The result, as derived above, is

$$P(p|N_1, r_1) = \frac{(N_1 + 1)!}{r_1!(N_1 - r_1)!} p^{r_1}(1 - p)^{N_1 - r_1} = 101p^{100}$$

The mode is at $p^* = r_1/N_1 = 1$, the mean is $E[p] = (r + 1)/(N + 2) = 101/102 \approx 0.99$, the variance is $\sigma^2 = (E[p](1 - E[p]))/(N + 3) \approx 10^{-4}$. I.e., the efficiency is very high. A plot of the posterior probability density is shown in Fig. 2.6.

In situations where the distribution peaks at one of the extreme values of the parameter, we often want to quote a limit on the parameter value. In this case, we can specify that $p$ is greater than some value at some probability. For this, we need the cumulative probability distribution:

$$F(p) = \int_0^p P(p'|N = 100, r = 100)dp' = p^{101}.$$
In setting limits, it is often the case that 90 or even 95 % probabilities are chosen. If we ask 'what is the value to have 95 % probability that $p$ is greater than this cutoff', we solve

$$0.05 = F(p) = p^{101} \quad p_{95} \approx 0.97.$$ 

So, we conclude that, at 95 % probability, $p > 0.97$. The most probable value is $p^* = 1$.

**Second Data Set**

Let us suppose we analyze only the second data set, ignoring the first result. In this case, starting with a flat prior, we have

$$P(p|N_2, r_2) = \frac{(N_2 + 1)!}{r_2!(N_2 - r_2)!}p^{r_2}(1-p)^{N_2-r_2} = \frac{101 \cdot 100 \cdot 99 \cdot 98 \cdot 97 \cdot 96}{5 \cdot 4 \cdot 3 \cdot 2}p^{95}(1-p)^5.$$ 

The resulting distribution is shown in Fig. 2.7 as the blue curve, together with the result from the first experiment. The distribution peaks at $p^* = 95/100 =$
0.95. The resulting distribution from analyzing both data sets (either by adding the data sets and analyzing at once, or from using the posterior pdf from the first experiment as prior for the second) is also shown.

![Figure 2.7: Posterior probability distribution $P(p)$ for $N_2 = 100$ and $r_2 = 95$ (blue), for $N_1 = 100$ and $r_1 = 100$ (red), and from analyzing both data sets (black-dashed).](image)

Are the results compatible? How would we decide - or - does the question even make sense? Let’s compare two possibilities:

1. Model M1: The two results come from the same underlying process, with the same Binomial success parameter $p$.

2. Model M2: The situation was different in the two cases, and we should therefore have two different Binomial success parameters, $p_1$ and $p_2$.

We have $P(M1) + P(M2) = 1$. The prior odds (before we get the experimental results) are:

$$O_0 = \frac{P_0(M1)}{P_0(M2)}$$

and, given the way we set up the experiments, take this to be a rather large value. Let’s take $O_0 = 10$ for specificity.
We evaluate the posterior odds ratio:

\[ O = \frac{P(M_1)}{P(M_2)} = \frac{P(N_1, r_1, N_2, r_2|M_1)P_0(M_1)}{P(N_1, r_1, N_2, r_2|M_2)P_0(M_2)} = B(M_1, M_2)O_0 \]

where we have defined the Bayes’ Factor, \( B(M_1, M_2) \) as

\[ B(M_1, M_2) = \frac{P(N_1, r_1, N_2, r_2|M_1)}{P(N_1, r_1, N_2, r_2|M_2)} \]

The Bayes’ factor then tells us how, from the measured data, we should adjust our odds ratio for our two competing models. Let’s evaluate it for our specific example. Starting with model \( M_1 \), we expand the probability for the model using the Law of Total Probability and then integrate over the possible values of \( p \):

\[ P(N_1, r_1, N_2, r_2|M_1) = \int P(N_1, r_1, N_2, r_2|M_1, p)P(p|M_1)dp \]

where \( P(p) \) is the probability distribution we assign to \( p \) (we can choose here either to use the prior probability for \( p \) or the posterior probability after seeing the data - see later discussion). For the probability of the data, let us take the product of the individual probabilities of the data sets:

\[ P(N_1, r_1, N_2, r_2|M_1, p) = \frac{N_1!N_2!}{r_1!r_2!(N_1 - r_1)!(N_2 - r_2)!}p^{r_1+r_2}(1-p)^{N_1+N_2-r_1-r_2} \]

where \( r = r_1 + r_2 \) and \( N = N_1 + N_2 \). Using our data values, we find \( P(N_1, r_1, N_2, r_2|M_1) \approx 1.5 \cdot 10^{-4} \).
2. We use the posterior probability distribution for $p$, if we are interested in seeing if our model, adjusted after seeing the data, gives a good description of the data set. We would presumably use this choice if all we were interested in was whether the modeling gives a good description of the observed data. In this case, we find:

$$P(N_1, r_1, N_2, r_2|M_1) = K \frac{(N+1)!}{r!(N-r)!} \frac{(2r)!(2N-2r)!}{(2N+1)!}$$

Using our data values, we find $P(N_1, r_1, N_2, r_2|M_1) \approx 3.6 \cdot 10^{-3}$, which is larger than the previous case because now our probability distribution for $p$ is weighted to the preferred values.

3. We use our best fit value of $p$, say $p^*$ resulting from the analysis of both data sets to see if with this single best choice we get a reasonable probability for the observed data. If the probability is reasonably high, then we could say the modeling is adequate since there is at least one value of $p$ for which both data sets can be explained. In this case, we find

$$P(N_1, r_1, N_2, r_2|M_1) = K(p^*)^r (1 - p^*)^{N-r}$$

with $p^* = r/N = 0.975$. We find $P(N_1, r_1, N_2, r_2|M_1) \approx 5.3 \cdot 10^{-3}$ which is somewhat larger again because we now use the mode of the posterior distribution for $p$.

Let us now turn to the second Model. In this case, we say that the two data sets have distinct Binomial success probabilities, and we therefore factorize as follows:

$$P(N_1, r_1, N_2, r_2|M_2) = P(N_1, r_1|M_2)P(N_2, r_2|M_2)$$

$$= \int P(N_1, r_1|M_2, p_1)P(p_1|M_2)dp_1 \int P(N_2, r_2|M_2, p_2)P(p_2|M_2)dp_2$$

Using a similar notation as above

$$P(N_1, r_1|M_2, p_1) = \frac{N_1!}{r_1!(N_1-r_1)!} p_1^{r_1}(1-p_1)^{N_1-r_1}$$

$$= K_1 \cdot p_1^{r_1}(1-p_1)^{N_1-r_1}$$
and similarly
\[ P(N_2, r_2|M2, p_2) = K_2 \cdot p_2^{r_2}(1 - p_2)^{N_2-r_2} \]
with \( K_1 = 1 \) and \( K_2 \approx 7.5 \cdot 10^7 \).

We again need to specify probability distributions for \( p \), and consider the same three cases as above.

1. Flat priors for \( p_1 \) and \( p_2 \):
\[
P(N_1, r_1, N_2, r_2|M2) = K_1^{r_1!(N_1 - r_1)!} \cdot K_2^{r_2!(N_2 - r_2)!} \approx 9.8 \cdot 10^{-5}
\]

2. using the individual posterior probability densities for the two data sets:
\[
P(N_1, r_1, N_2, r_2|M2) \approx 0.13
\]

3. using the value of \( p \) at the modes of each data set:
\[
P(N_1, r_1, N_2, r_2|M2) \approx 0.18
\]

If we now calculate the Bayes Factor, we have for the three choices of \( P(p) \):

1. \( B(M1,M2) =1.5; \)
2. \( B(M1,M2) =0.03; \)
3. \( B(M1,M2) =0.03. \)

So, if we wanted to compare the results for models that were considered fully specified before the data was taken, then we would have case (1) and we would conclude that we favor the situation with a single Binomial probability to describe the two data sets. Here the improved description of the data by allowing for two distinct Binomial distributions is offset by the penalty assigned from using two Prior probabilities. This type of result - where the more complicated model is penalized via the prior density - can be in line with scientific expectations where ‘Occam’s Razor’ can be a useful tool.
However, if we instead take the point of view that we are willing to adjust the model parameters to best describe the data, and then compare whether we prefer a single Binomial to describe both data sets or allow two distinct Binomial distributions, then we prefer having the latter. Our data is better described with the two Binomials case.

How should we decide? There is no general rule, and which view is taken depends very much on the individual situations being analyzed.

### 2.6 Multinomial Distribution

We keep the fixed number of trials, \(N\), but now we allow more than two possible outcomes. We’ll use \(m\) for the number of possible outcomes, \(p_j, \quad 1 \leq j \leq m\), is the probability in our Model for a particular outcome to end up in class \(j\) and \(r_j\) is the number of outcomes in category \(j\). We assume the probabilities are fixed (in an Urn example, where we can pull different colored balls from an Urn, this would be sampling with replacement). There are \(N!\) different sequences of possible outcomes. Since we do not distinguish between different orderings of the \(r_j\) outcomes, we have

\[
P(\vec{r}|\vec{p}, N) = \frac{N!}{r_1! \cdot r_2! \cdots r_m!} p_1^{r_1} \cdots p_m^{r_m}.
\]

For \(m = 2\), this reduces to the Binomial probability distribution.

Note that we now have \(m\)-dimensional collections of outcomes, denoted by \(\vec{r}\) and \(m\)-dimensional sets of success probabilities, \(\vec{p}\). The normalization condition

\[
\sum_{j=1}^{m} p_j = 1
\]

means that only \(m - 1\) of these are independent.

#### 2.6.1 Bayesian Data Analysis

We perform an experiment with \(N\) trials, and measured \(\vec{r}\) values, and now want to extract the probability distributions for the parameters \(\vec{p}\). Bayes’ equation tells us

\[
P(\vec{p}|\vec{r}, N) = \frac{P(\vec{r}|N, \vec{p}) P_0(\vec{p})}{\int P(\vec{r}|N, \vec{p}) P_0(\vec{p}) d\vec{p}}
\]
For $m = 2$, the Binomial case, we saw that we have a lot of freedom in defining the Prior. For $m > 2$, the Prior must satisfy the normalization condition, so we immediately have a more complicated situation. E.g., it is not possible to assign flat independent prior probabilities for each direction. Let’s see what would happen if we tried:

$$P_0(\vec{p}) = \prod_{j=1}^{m} P_0(p_j)$$

$$\int P_0(\vec{p})d\vec{p} = \prod_{j=1}^{m} \int P_0(p_j)dp_j = \prod_{j=1}^{m} 1 = 1$$

$$p = \sum_{j=1}^{m} p_j \quad 0 \leq p \leq m$$

I.e., our total probability does not sum to 1 but can be less than or greater than 1.

### 2.6.2 Likelihood

Let us first consider the probability of the data outcome, $P(\vec{r}|N, \vec{p})$ viewed as a function of the parameters, $\vec{p}$. I.e., we consider the likelihood function:

$$\mathcal{L}(\vec{p}) = P(\vec{r}|N, \vec{p}) = \frac{N!}{r_1! \cdot r_2! \cdots r_m!} p_1^{r_1} \cdots p_m^{r_m}$$

We can introduce a Lagrange multiplier, and set all derivatives of

$$L(\vec{p}, \lambda) = \mathcal{L}(\vec{p}) - \lambda \left( \sum_{j=1}^{m} p_j - 1 \right)$$

to zero. The derivative wrt $\lambda$ returns our constraint equation, while derivatives wrt one of the $p_j$ gives

$$\lambda \propto (r_j/p_j) \mathcal{L}(\vec{p}).$$

I.e., we see that $p_j \propto r_j$, since this expression is true for all $j$. Using the normalization condition and $\sum_{j=1}^{m} r_j = N$, we have $p_j^* = r_j/N$. 

2.6. MULTINOMIAL DISTRIBUTION

2.6.3 Adding the Prior

The denominator in Bayes equation is a number, so if we want to find the maximum of the Posterior probability, we need to find the maximum of the numerator. We now have to take the derivatives of

\[ L(\vec{p}, \lambda) = \mathcal{L}(\vec{p})P_0(\vec{p}) - \lambda \left( \sum_{j=1}^{m} p_j - 1 \right) \]

and set them to zero. Suppose that the prior is symmetric such that it has the same form for all \( p_j \). Using the chain rule, we have

\[ \lambda = \frac{\partial \mathcal{L}(\vec{p})}{\partial p_j} P_0(\vec{p}) + \mathcal{L}(\vec{p}) \frac{\partial P_0(\vec{p})}{\partial p_j} \]

and the second term is the same for all \( j \) given our symmetry assumption. Therefore we again get \( p_j^* = r_j/N \).

I.e., we can get the mode of the \( m \)-dimensional posterior probability density under these conditions, but we would like to go further and also define some probability intervals. We have reached the limits of what we can do analytically, and actually finding the posterior pdf will require numerical techniques. We come back to this problem after we have learned some Monte Carlo techniques.

Literature
Exercises

1. Prove the relations given in Section 2.1.1.

2. Construct a Table similar to 2.1 for $N = 6, p = 0.3$ and find the following:
   
   (a) The Central Interval for $1 - \alpha = 0.68$ and $1 - \alpha = 0.90$
   
   (b) The Smallest Interval for $1 - \alpha = 0.68$ and $1 - \alpha = 0.90$

3. Show that the likelihood based on the Binomial distribution is not properly normalized, and is therefore not a probability density for $p$.

4. Prove that the mode of the likelihood function for a Binomial distribution is given by $r/N$.

5. Derive the relations given in Section 2.4.

6. Compare the (normalized) likelihood function to a Gaussian approximation for the cases:
   
   (a) $N = 5; r = 2$
   
   (b) $N = 20; r = 17$
   
   (c) $N = 100; r = 63$

7. Show that the variance of a flat prior is given

   \[
   \frac{E[p](1 - E[p])}{N + 3} = \frac{1}{12}
   \]

8. For the following function

   \[ P(x) = xe^{-x} \quad 0 \leq x < \infty \]

   (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 mode and 68 % smallest interval
9. 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.

10. Analyze the data in the table from a frequentist perspective by finding the 90 % confidence level interval for $p$ as a function of energy.

11. We now analyze the results in the table above using regularization. I.e., we put in prior knowledge in the form of the expectation that the efficiencies are expected to vary smoothly with energy. One way of doing this is to use the posterior probability density at a neighboring energy as the prior to analyze the data. Using the results of section 2.5.3 and starting with a flat prior for the lowest energy point, find the posterior probability density as a function of energy using this scheme for regularization.

12. Let us see what happens if we reuse the same data multiple times. We have $N$ trials and measure $r$ successes. Show that is 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 $n \rightarrow \infty$?

13. A Bernoulli experiment is performed with 12 trials, resulting in 9 successes. What is the probability for this outcome? Now imagine that instead a
Bernoulli experiment is performed until 3 failures occur. What is the probability to have found 9 successes? The result contains a Negative Binomial and is

$$\binom{9+2}{2} p^9 (1-p)^3.$$  

Why are the results different?

14. Two measurements are to be performed of a quantity \(x\) which is assumed to have a flat probability distribution between \([0, 1]\); i.e., \(P(x) = 1\) if \(0 \leq x \leq 1\); \(P(x) = 0\) otherwise.

(a) Show that the probability distribution for \(z = x_1 + x_2\), where \(x_1, x_2\) are the two measurements of \(x\), is given by

\[
P(z) = \begin{cases} 
  z & 0 \leq z \leq 1 \\
  2 - z & 1 \leq z < 2 
\end{cases}
\]

(b) Show that the probability distribution for \(z = x_1 \cdot x_2\), where \(x_1, x_2\) are the two measurements of \(x\), is given by

\[P(z) = \ln z\]

(c) Show that the probability distribution for \(z = x_1/x_2\), where \(x_1, x_2\) are the two measurements of \(x\), is given by

\[
P(z) = \begin{cases} 
  1/2 & 0 \leq z \leq 1/2 \\
  \frac{1}{2z^2} & z > 1/2 
\end{cases}
\]
2.6. MULTINOMIAL DISTRIBUTION

n!
Chapter 3

Poisson Distribution

3.1 Introduction

A Poisson distribution is the distribution of choice to model the probabilities for numbers of events (counts) when we do not know the number of trials, only that it is a very large number, and we assume there is a fixed probability of ‘success’ in each trial. We also assume the trials are independent of each other. A continuous time process with a constant rate will produce a Poisson distributed number of events in a fixed time interval. An example from high energy physics is the distribution of the numbers of events expected for different processes that can occur in particle collisions. The particles in the collider pass through each other at a huge rate - at the LHC of order $10^{11}$ particles have the chance to collide every 25 nanoseconds. The chance to get an event of interest in any one crossing of the beam particles is extremely small, and the number of ‘good’ events is therefore expected to follow a Poisson distribution.

3.2 Derivation

We will look at two different derivations of the Poisson distribution. In the first derivation, we start with the Binomial distribution, and take the limit of $N \to \infty$ while keeping $\nu = Np$ finite. In the second derivation, we start from the assumption of a constant rate process, and calculate the number of events expected in a fixed time interval $T$. While the Poisson probability distribution will depend on the parameter $\nu$, usually the quantity of interest is related to the probability $p$ and a relation will be needed between $\nu$ and the parameters in the scientific model.
CHAPTER 3. POISSON DISTRIBUTION

This will be discussed via detailed examples.

3.2.1 Limit of the Binomial Distribution

Recall the Binomial probability distribution for the number of successes \( n \) in \( N \) trials given a success probability \( p \) per trial:

\[
P(n|N,p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}.
\]

We introduce \( \nu = Np \) and use this to replace \( p \):

\[
P(n|N,\nu) = \frac{N!}{n!(N-n)!} \left( \frac{\nu}{N} \right)^n \left( 1 - \frac{\nu}{N} \right)^{N-n}.
\]

In the limit \( N \to \infty \) and \( n \) not too large, we have

\[
\frac{N!}{(N-n)!} \to N^n.
\]

Also, we can write

\[
\lim_{N \to \infty} \left( 1 - \frac{\nu}{N} \right)^{N-n} \to \lim_{N \to \infty} \left( 1 - \frac{\nu}{N} \right)^N \to e^{-\nu}
\]

and putting the pieces together we are left with

\[
P(n|\nu) = \frac{e^{-\nu} \nu^n}{n!} \tag{3.1}
\]

which is the famous Poisson distribution. It only depends on the one parameter, \( \nu \), rather than the two parameters \( N, p \) of the Binomial distribution. The Poisson distribution, together with the Gaussian distribution, is the most important probability distribution for use in the physical sciences. Note that we now have an infinite number of trials, so that the number of successes, \( n \), is also unbounded. The quantity \( \nu \) is the expectation of \( n \):

\[
E[n] = \nu.
\]

We will discuss further properties of the Poisson distribution after going through the second derivation.
3.2. DERIVATION

3.2.2 Constant Rate Process & Fixed Observation Time

Imagine we have a process producing ‘events’ at a constant rate, \( R \). We will be interested in the distribution of numbers of events in a time \( T \). We start by calculating the probability that no event has occurred up to time \( T = n\Delta t \), where our time \( T \) has been divided up into \( n \) intervals of length \( \Delta t \). The probability of no event up to time \( T \) is:

\[
P(\text{no event in } T) = P(\text{no event in } \Delta t)^n = (1 - R\Delta t)^n .
\]

We now take the limit \( n \to \infty \) and get the well-known result

\[
\lim_{n \to \infty} (1 - R\Delta t)^n = \lim_{n \to \infty} (1 - RT/n)^n = e^{-RT} .
\]

The probability to have no events in a fixed time interval in a constant rate process follows an exponential law.

We now look at the probability to have one event in the time span \( T \). Imagine that this event occurs at time \( t \) in the time interval \( dt \), as indicated in the sketch:

```
  |   |   |
  |   |   |
----------
t = 0  t  T
```

We can decompose the time interval \( T \) into three parts: the time from \( 0 \to t \) where no event took place, the time from \( t \to t + dt \) where we have an event, and the time from \( t + dt \to T \) where again no event took place. The probability for this sequence is:

\[
P(\text{one event at time } t) = e^{-Rt}R dt e^{-R(T-(t+dt))} = e^{-RT} e^{Rdt} R dt .
\]

We assume the event occurs in infinitesimal time, so that \( e^{Rdt} = 1 \), and then integrate over all possible times \( t \) within \( T \) when the event could have occurred:

\[
P(\text{one event in } T) = \int_0^T e^{-RT} R dt = RT e^{-RT} .
\]
The product $RT$ is just the expected number of events in time $T$ and we will call this $\nu = E[n] = RT$, so that

\[ P(\text{one event in } T) = \nu e^{-\nu}. \]

We now consider an arbitrary number of events, $n$, in the time interval $T$. If events occur in infinitesimal time, then there will always be a total time $T$ when no event occurred, so that the factor $\exp(-RT)$ will always be present in the probability. Imagine the events occurred at times $t_1, t_2, \ldots, t_n$. We have probabilities $Rdt_1, Rdt_2, \ldots Rdt_n$ for the events that occurred. We now have to integrate over all possible times, taking care that event $j$ occurred after event $j - 1$:

\[ P(n|R, T) = e^{-RT} \int_0^T Rdt_1 \int_{t_1}^T Rdt_2 \cdots \int_{t_{n-1}}^T Rdt_n. \]

Performing the last integral, we get $\int_{t_{n-1}}^T Rdt_n = R(T - t_{n-1})$. Define $\xi = (T - t_{n-1})$. The penultimate integral is then

\[ \int_{t_{n-2}}^T R^2 (T - t_{n-1}) dt_{n-1} = R^2 \int_0^{T-t_{n-1}} \xi d\xi = R^2 \frac{(T - t_{n-2})^2}{2}. \]

We can continue with these integrations and get the final result

\[ P(n|R, T) = e^{-RT} \frac{R^n T^n}{n!} \]

or, using the notation with $\nu = RT$, we have

\[ P(n|\nu) = \frac{e^{-\nu} \nu^n}{n!} \quad (3.2) \]

### 3.3 The Poisson Distribution

We go through some of the properties of the Poisson distribution and then discuss with graphical examples.

#### 3.3.1 Properties

Normalizaton:

\[ \sum_{n=0}^{\infty} \frac{\nu^n e^{-\nu}}{n!} = e^{-\nu} \sum_{n=0}^{\infty} \frac{\nu^n}{n!} = e^{-\nu} e^\nu = 1 \]
3.3. THE POISSON DISTRIBUTION

Expectation value:

\[ E[n] = \sum_{n=0}^{\infty} n \frac{\nu^n e^{-\nu}}{n!} = \nu e^{-\nu} \sum_{n=1}^{\infty} \frac{\nu^{n-1}}{(n-1)!} = \nu e^{-\nu} e^\nu = \nu \]

Variance:

\[ V[n] = E[n^2] - E[n]^2 \]

\[ E[n^2] = \sum_{n=0}^{\infty} n^2 \frac{\nu^n e^{-\nu}}{n!} = \nu e^{-\nu} \sum_{n=1}^{\infty} n \frac{\nu^{n-1}}{(n-1)!} \]

Write \( n = (n - 1) + 1 \) and split the sum into two parts:

\[ \sum_{n=1}^{\infty} n \frac{\nu^{n-1}}{(n-1)!} = \nu \sum_{n=2}^{\infty} \frac{\nu^{n-2}}{(n-2)!} + \sum_{n=1}^{\infty} \frac{\nu^{n-1}}{(n-1)!} = (\nu + 1)e^\nu \]

so

\[ V[n] = E[n^2] - E[n]^2 = \nu e^{-\nu} (\nu + 1)e^\nu - \nu^2 = \nu . \]

The variance of \( n \) for a Poisson distribution is the same as the expectation value.

Mode:

\[ n^*_1 = \nu \quad \nu \notin \mathbb{Z} \]

\[ n^*_2 = \nu \quad n^*_2 = \nu - 1 \quad \nu \in \mathbb{Z} \]

I.e., the mode is the integer part of \( \nu \), unless \( \nu \) is an integer, in which case there is a double mode at \( \nu \) and \( \nu - 1 \).

3.3.2 Examples and Graphical Display

Figure 3.1 shows the Poisson distribution for several values of \( \nu \). We see that the Poisson distribution is very asymmetric for small values of \( \nu \), and becomes rather symmetric for large \( \nu \). We consider a specific example: \( \nu = 10/3 = 3.3 \). The probabilities, cumulative probabilities, rank and cumulative probability according to rank are given in Table 14.

We see from the table that the most probable value is \( n^* = 3 = \lfloor \nu \rfloor \). We can use such a table to form intervals containing a certain probability, as discussed
Figure 3.1: Probability distribution for $n$ for $\nu = 0.1$ (upper left); $\nu = 0.5$ (upper right); $\nu = 1$ (second row left); $\nu = 2$ (second row right); $\nu = 5$ (third row left); $\nu = 10$ (third row right); $\nu = 20$ (lower left); $\nu = 50$ (lower right).
Table 3.1: Values of $n$, the probability to observe such a value given $\nu = 3.3$ and the cumulative probability, rounded to four decimal places. The fourth column gives the rank in terms of probability - i.e., the order in which this value of $n$ is used in calculating the smallest set $O_{1-\alpha}^S$, and the last column gives the cumulative probability summed according to the rank.

| $n$ | $P(n|\nu)$ | $F(n|\nu)$ | $R$ | $F_R(n|\nu)$ |
|-----|-------------|-------------|-----|---------------|
| 0   | 0.0357      | 0.0357      | 7   | 0.9468        |
| 1   | 0.1189      | 0.1546      | 5   | 0.8431        |
| 2   | 0.1982      | 0.3528      | 2   | 0.4184        |
| 3   | 0.2202      | 0.5730      | 1   | 0.2202        |
| 4   | 0.1835      | 0.7565      | 3   | 0.6019        |
| 5   | 0.1223      | 0.8788      | 4   | 0.7242        |
| 6   | 0.0680      | 0.9468      | 6   | 0.9111        |
| 7   | 0.0324      | 0.9792      | 8   | 0.9792        |
| 8   | 0.0135      | 0.9927      | 9   | 0.9927        |
| 9   | 0.0050      | 0.9976      | 10  | 0.9976        |
| 10  | 0.0017      | 0.9993      | 11  | 0.9993        |
| 11  | 0.0005      | 0.9998      | 12  | 0.9998        |
| 12  | 0.0001      | 1.0000      | 13  | 1.0000        |
for the Binomial case in section 2.1.2. For example, the Central Interval with probability $1 - \alpha \geq 0.9$ contains the elements $O_{0.9}^C = \{1, \ldots, 7\}$ and the Smallest Interval containing probability $1 - \alpha \geq 0.9$ is $O_{0.9}^S = \{1, \ldots, 6\}$.

**Neyman Confidence Interval**

In performing a Neyman Confidence Level calculation, we need to produce a table such as the one in Tab. 14 for all values of $\nu$ that are possibly interesting. Then, upon defining a Confidence Level $1 - \alpha$ and choosing a construction for the interval (e.g., Central, Smallest, ), the smallest and largest members of the intervals, $n_1, n_2$ are found for each $\nu$. Taking $1 - \alpha = 0.9$ and the Central Interval yields the plot shown in Fig. 3.2.

Figure 3.2: 90 % Central Interval bands for the Poisson probability distribution. For a given $\nu$, the red line gives the value of $n_1$, the smallest element in the set, while the blue line gives $n_2$, the largest element in the set.

Let’s suppose we perform an experiment and measure $n_D = 3$. We go to our figure, and find that for $0.8 < \nu < 7.8$, the outcome $n = 3$ is included in $O_{0.9}^C$. We therefore conclude that the 90 % CL range for $\nu$ is $0.8 < \nu < 7.8$. Again, this is not a probabilistic statement on $\nu$, but is a summary of the data, here
n_D = 3, framed in the language of the statistical model. The conclusion is that, for 0.8 < \nu < 7.8, the outcome n_D = 3 is in the 90 \% Central Interval - nothing more is to be concluded from this measurement.

3.4 Bayesian Analysis

In order to make probabilistic statements about possible values of \nu, we need to perform a Bayesian analysis. Our master formula is

\[ P(\nu|n) = \frac{P(n|\nu)P_0(\nu)}{\int_0^\infty P(n|\nu)P_0(\nu)d\nu} = \frac{\nu^n e^{-\nu}P_0(\nu)}{\int_0^\infty \nu^n e^{-\nu}P_0(\nu)d\nu}. \] (3.3)

To go further, we need to specify a Prior probability density for \nu. Let’s start by taking a flat prior for \nu, \(P_0(\nu) = C\). We have to limit the possible values of \nu, however, or we will not be able to normalize our Prior pdf. Let’s say we have a maximum conceivable value of \nu, \nu_{\text{max}}. Then, \(C = 1/\nu_{\text{max}}\) and

\[ P(\nu|n) = \frac{\nu^n e^{-\nu}}{\int_0^{\nu_{\text{max}}} \nu^n e^{-\nu}d\nu}. \]

For \nu_{\text{max}} \gg n, we approximate the integral

\[ \int_0^{\nu_{\text{max}}} \nu^n e^{-\nu}d\nu \approx \int_0^\infty \nu^n e^{-\nu}d\nu = n! \]

and we find

\[ P(\nu|n) = \frac{e^{-\nu}\nu^n}{n!}. \] (3.4)

The RHS is our Poisson function, but now for \(n\) fixed and \nu as the variable (with \nu \geq 0). We note that this is also the likelihood. I.e., as we have seen previously, the likelihood is proportional to the Posterior pdf for the case where a flat prior is chosen. In this case, the likelihood is normalized to one. The mode of the Posterior is clearly at \(\nu^* = n\). Let us calculate a few more quantities that can be useful in summarizing the Posterior:

**Normalization**

\[ \int_0^\infty P(\nu|n)d\nu = \frac{1}{n!} \int_0^\infty \nu^n e^{-\nu}d\nu = \frac{n!}{n!} = 1 \]
CHAPTER 3. POISSON DISTRIBUTION

Expectation value of \( \nu \)

\[
E[\nu] = \int_0^{\infty} \nu \frac{e^{-\nu}}{n!} d\nu = \frac{1}{n!} \int_0^{\infty} \nu^{n+1} e^{-\nu} d\nu = \frac{(n + 1)!}{n!} = n + 1
\]

Variance of \( \nu \)

\[
V[\nu] = E[\nu^2] - (E[\nu])^2 = \int_0^{\infty} \nu^2 \frac{e^{-\nu}}{n!} d\nu - (n + 1)^2 = \frac{(n + 2)!}{n!} - (n + 1)^2 = n + 1
\]

Cumulative Probability of \( \nu \)

\[
F(\nu|n) = \int_0^{\nu} \frac{e^{-\nu'}}{n!} d\nu' = \frac{1}{n!} \left[ -\nu^ne^{-\nu} \bigg|_0^n + n \int_0^{\nu} \nu^{n-1} e^{-\nu'} d\nu' \right] = 1 - e^{-\nu} \sum_{i=0}^{n} \frac{\nu^i}{i!}
\]

3.4.1 Discussion and Examples: Flat Prior Results

The mode of the posterior is at \( \nu^* \) and the mean is always one larger. For \( n = 0 \), the most likely value for \( \nu \) is 0 and the expectation value is 1. Recall that we started out saying all values of \( \nu \) in the range \( 0 \leq \nu \leq \nu_{\text{max}} \) were equally likely. We now performed an experiment and found that \( P(\nu|0) = e^{-\nu} \) and \( E[\nu] = 1 \). This is a dramatic change in knowledge! Finding \( n = 0 \) does not mean that you have no information - on the contrary, we have completely changed our probability distribution that we assign to \( \nu \). Moral - take the \( n = 0 \) result seriously!

At this point, you might be wondering - every experiment that measures \( n = 0 \) gets the same information out for \( \nu \). What if one experiment observed for twice as long as the other? How can they give the same result? The answer is that the result for \( \nu \) is the same, but the result for the parameters in your model are different. Go back to our case of a constant rate process, where \( \nu = RT \). If \( T \) is twice as large, then the value of \( R \) that we extract from \( \nu \) will be \( 1/2 \) as large. \( R \) is the quantity of interest in the physical model! To make this concrete, let’s take the case \( n = 0 \) and calculate the 90 % probability upper limit on \( \nu \). We find the
value of \( \nu \) for which the cumulative reaches 0.9:

\[
0.9 = F(\nu|0) \implies \int_0^{\nu_{90}} e^{-\nu} d\nu = 1 - e^{-\nu_{90}} \implies \nu_{90} = 2.3.
\]

So, we say that with 90\% probability \( \nu < 2.3 \). If the had observed for \( T = 100 \) s, then our conclusion would be that the rate \( R < 0.023 \) Hz with 90\% credibility.

What if we look at a different extreme case, where \( n \) is large? Take \( n = 100 \) as an example. Then we find that the mean and mode are, relatively speaking, very close (101 vs 100) and our posterior probability distribution looks quite symmetric. In fact, the skew of the distribution is

\[
\gamma = \frac{E[(\nu - E[\nu])^3]}{(V[\nu])^{3/2}} = \frac{2}{\sqrt{n+1}}
\]

and goes to zero for large \( n \), and at some point it may be adequate to approximate the posterior pdf with a Gaussian function.

The posterior pdf is shown for four values of \( n \) in Fig. 3.4 and the corresponding cumulative probabilities are given in Fig. ??.
Figure 3.4: Cumulative probability distribution for different values of observed events, here labeled $x_t$. These results are for a flat prior pdf. The dashed horizontal line indicates $F(\nu) = 0.95$. 
3.4. BAYESIAN ANALYSIS

3.4.2 Non-flat Priors

Conjugate Prior

In the general case, solving for the Posterior pdf cannot be done analytically, and numerical calculations are required. However, if the prior can be specified in the form of a gamma distribution, then the posterior will also be a gamma distribution (i.e., the gamma distribution is the Conjugate Prior associated with the Poisson distribution). For a prior based on the gamma distribution, we have

\[ P_0(\nu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \nu^{\alpha-1} e^{-\beta\nu} \quad (3.5) \]

where here \( \Gamma(\alpha) \) is the gamma function defined by

\[ \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx. \]

The gamma function is defined for all complex arguments except the negative integers and zero. In our applications, \( \alpha \) will be a positive definite real number. For a measurement resulting in \( n \) events, the posterior distribution is then again a gamma distribution with \( \alpha \rightarrow \alpha + n \) and \( \beta \rightarrow \beta + 1 \). As an example, we can use this result to see what happens if we start with a flat prior for \( \nu \), and make two measurements that are assumed to have the same underlying Poisson mean and resulting in \( n_1 \) and \( n_2 \) events. As we have seen above, the result from the first measurement yields the posterior pdf

\[ P(\nu|n_1) = \frac{e^{-\nu\nu^{n_1}}}{n_1!} \]

We now want to use this as the prior for the second measurement. Comparing to the gamma distribution definition, we need to set \( \alpha_1 = n_1 + 1 \) and \( \beta_1 = 1 \). Our posterior after analyzing data set \( n_2 \) will then be a gamma distribution with \( \alpha_2 = n_1 + n_2 + 1 \) and \( \beta_2 = 2 \). I.e., we have

\[ P(\nu|n_1, n_2) = \frac{2^{n_1+n_2}\nu^{n_1+n_2}e^{-2\nu}}{\Gamma(n_1 + n_2 + 1)} \]

The denominator is just \( (n_1 + n_2)! \), so we can write this as

\[ P(\nu|n_1, n_2) = \frac{(2\nu)^{n_1+n_2}e^{-2\nu}}{(n_1 + n_2)!}. \]
We note that this is exactly the same result which we would get by analyzing the two data sets in one step starting from a flat prior. Defining \( n = n_1 + n_2 \), we have

\[
P(\nu'|n) = \frac{e^{-\nu'}\nu'^n}{n!}
\]

where now we have used the symbol \( \nu' \) for the expectation. Comparing the results, we see that \( \nu' = 2\nu \) which is as expected, since the experiment has been run twice as long.

**Detailed Example**

As a specific example, imagine that we are interested in finding the rate, \( R \), for a given process. We perform an experiment for time \( T_1 \) and count the number of observed events. The expected number of events is \( \nu_1 = R \cdot T_1 \) where \( T_1 \) is the observation time. Suppose we started with a flat prior on \( R \) up to a maximum values \( R_{\text{max}} \) and we observed 0 events. How do we proceed? We write Bayes’ equation in terms of the parameter of interest, \( R \):

\[
P(R|T_1, n = 0) = \frac{P(n = 0|R, T_1)P_0(R)}{\int P(n = 0|R, T_1)P_0(R)dR} = \frac{e^{-RT_1}(1/R_{\text{max}})}{\int e^{-RT_1}(1/R_{\text{max}})dR} = T_1e^{-RT_1}
\]

We now run a second experiment to test the same process, and observe \( n_2 \) events in time \( T_2 \). What can we say about the rate? Let us translate the prior for \( R \) (the posterior of the first experiment) into a prior for \( \nu_2 \), the expected number of events in the second experiment. We need to solve

\[
P_0(\nu_2) = P_0(R(\nu_2)) \left| \frac{dR}{d\nu_2} \right|
\]

where \( \nu_2 = R \cdot T_2 \). I.e.,

\[
P_0(\nu_2) = \frac{T_1}{T_2}e^{-\frac{T_1}{T_2}\nu_2}
\]

If we run the second experiment ten times as long, such that \( T_2 = 10 \cdot T_1 \), then we would have

\[
P_0(\nu_2) = 0.1 \cdot e^{-0.1 \cdot \nu_2}.
\]
Note that this prior, written in terms of $\nu_2$, is less steep than the posterior for $\nu_1$. The reason is of course because of the different observation times, but not handling this correctly is a common mistake, so be careful!

Referring to the previous discussion on conjugate priors, we can write our prior in terms of the gamma distribution with $\alpha_1 = 1$ and $\beta_1 = 0.1$ with $\Gamma(1) = 1$. The posterior for $\nu_2$ is then a gamma function with $\alpha_2 = 1 + n_2$ and $\beta_2 = 1.1$, so that we have

$$P(\nu_2) = \frac{(1.1)^{n_2+1}\nu_2^{n_2}}{\Gamma(1+n_2)}e^{-1.1\nu_2}$$

If $n_2 = 0$, then we end up with

$$P(\nu_2|n_2 = 0) = 1.1e^{-1.1\nu_2}$$

or, in terms of $R$

$$P(R|n_2 = 0) = P(\nu_2(R))\left|\frac{d\nu_2}{dR}\right| = 1.1T_2e^{-1.1RT_2}.$$ 

This distribution has expectation value for $R$ which is 11 times smaller than if we only had the result from the first experiment, and 10% smaller than if we only had the result from the second experiment.

**Jeffreys Prior**

The Jeffreys prior is intended to be a non-informative prior distribution, and is defined with relation to the Fisher information as follows:

$$P_0(\nu) \propto \sqrt{\det I(\nu)}$$

(3.6)

where the Fisher information, $I$, is defined as

$$I(\nu) = E\left[\left(\frac{\partial \log P(n|\nu)}{\partial \nu}\right)^2\right].$$

The Jeffreys prior is invariant under reparametrization of the probability of the data (see exercises) and is therefore favored by some. However, the whole point of Bayesian analysis is to be able to use informative priors in order to extract the best possible posterior probability statements. We therefore do not recommend its use, but rather insist that the analyzer does the best possible job in defining a prior.

For completeness, we note that the Jeffreys prior for the Poisson distribution is $P_0(\nu) = \sqrt{\frac{1}{\nu}}$ and is therefore not normalized.
3.5 Superposition of two Poisson Processes

We often have to deal with a superposition of two Poisson processes: the signal and the background, which are indistinguishable in the experiment. Usually we know the background expectations and want to know the probability that there is a signal in addition. If we don’t have a good estimate of the background expectation, we may run several experiments - some to learn about the background processes and others to learn about the possible presence of a signal. For example, in a high energy physics experiment, the signal for large extra dimensions may be the observation of events where momentum balance is (apparently) strongly violated. However this can be mimicked by neutrinos, energy leakage from the detector, etc. So, there are background processes that are indistinguishable from the signal, but we nevertheless want to make some statements about the possible presence of a signal. What do we do?

Let’s assume that both the background and signal events follow a Poisson distribution. E.g., we assume they would occur with constant rates, \( R_B \) and \( R_S \) and we observe for some time \( T \). The expected number of background events in then \( E[n_B] = \nu_B = R_B T \) and the expected number of signal events is \( E[n_S] = \nu_S = R_S T \). What is the probability of observing \( n \) events given that both of these processes are present? We decompose this into the possible combinations \((n_S = 0, n_B = n), (n_S = 1, n_B = n - 1), \cdots, (n_S = n, n_B = 0)\) and sum up the probabilities for each of these combinations:

\[
P(n|\nu_B, \nu_S) = \sum_{n_S=0}^{n} P(n_S|\nu_S)P(n-n_S|\nu_B) \\
= e^{-(\nu_S+\nu_B)} \sum_{n_S=0}^{n} \frac{\nu_S^{n_S} \nu_B^{n-n_S}}{n_S!(n-n_S)!} \\
= e^{-(\nu_S+\nu_B)} \frac{(\nu_S+\nu_B)^n}{n!} \sum_{n_S=0}^{n} \frac{n!}{n_S!(n-n_S)!} \left( \frac{\nu_S}{\nu_S+\nu_B} \right)^{n_S} \left( \frac{\nu_B}{\nu_S+\nu_B} \right)^{n-n_S}
\]

and we see that, defining \( p = \frac{\nu_S}{\nu_S+\nu_B} \) the argument in the sum is just the Binomial probability for observing \( n_S \) events given a probability \( p \) and \( n \) trials, and since the Binomial distribution is normalized, the sum is one. We therefore find:

\[
P(n|\nu_S, \nu_B) = \frac{e^{-(\nu_S+\nu_B)}(\nu_S+\nu_B)^n}{n!}
\]  (3.7)
which tells us that the combination of two Poisson processes yields another Poisson process with expectation value equal to the sum of the individual expectation values. This argument is clearly extendable to any number of Poisson processes, so that we can say the sum of any number of Poisson processes yields a Poisson process with Poisson parameter \( \nu = \sum_i \nu_i \), the sum of the individual Poisson parameters.

### 3.5.1 Likelihood for Multiple Poisson Processes

The formulae that we derived for a single Poisson process carry over to the case where we have several indistinguishable Poisson processes feeding into our event counts.

Let’s start with the situation where we have a background process, where we now denote the Poisson parameter with \( \lambda \) and a (possible) signal that we want to learn about, with Poisson expectation \( \nu \). The combined expectation is denoted \( \mu = \lambda + \nu \). We furthermore assume for now that we have figured out what \( \lambda \) should be from other information, and we consider it perfectly well known. We now perform an experiment to test if a signal is also present, and observe \( n \) events. What can we say about the possible signal expectation \( \nu \) ? The probability of the data, viewed as a function of the parameters (i.e., the Likelihood) is

\[
L(\mu) = P(n|\mu = \nu + \lambda) = \frac{e^{-\mu} \mu^n}{n!}.
\]

If \( \lambda \) is perfectly well known, then we can construct probability intervals for \( n \) for different values of \( \nu \). For example, if \( \lambda = 3.0 \), the probability table for \( \nu = 1/3 \) will be exactly the same as the one given in Table 14, where now \( \nu \) should be replaced with \( \mu \) in the table. Also, the band plot given in Fig. 3.2 carries over to \( \mu \). It is reproduced here with the axes relabeled as Fig. 3.5. If we now observe \( n_D \) events, we find all values of \( \mu \) for which \( n_D \) is in our interval (in this case we are looking at the Central Interval). This yields a Confidence Level interval for \( \mu \). The prescription to define an interval for \( \nu \) is then to subtract the known value of \( \lambda \).

Let us consider a couple of specific examples. In the first example, let’s use Fig. 3.5 and suppose \( \lambda = 4.0 \) and \( n_D = 4 \). From Fig. 3.5, we find (roughly) that

\[
\mu \in [1.4, 9.2].
\]

I.e., this is the range of \( \mu \) for which the observed number of events is within the 90\% central probability interval. If we now use the relation \( \mu = \lambda + \nu \) and
Figure 3.5: 90% Central Interval bands for the Poisson probability distribution for a combination of signal $\nu$ and background $\lambda$ expectations. For a given $\mu = \lambda + \nu$, the red line gives the value of $n_1$, the smallest element in the set, while the blue line gives $n_2$, the largest element in the set.
assume we know \( \lambda = 4.0 \) exactly, then we can say that at 90\% CL using the Central Interval,
\[
\nu \in [0, 5.2].
\]
We are therefore able to directly transfer the Confidence Interval construction defined for the single parameter Poisson distribution to the case where we have signal and background, and the background is perfectly known.

As a second case, let us take again \( \lambda = 4.0 \) but now assume that \( n_D = 0 \).

From Fig. 3.5, we find that
\[
\mu \in [0, 3].
\]
If we now extract the range of \( \nu \), we get a null interval:
\[
\nu \in [0, 0].
\]
I.e., there are no values of \( \nu \) for which the result falls into the central 90\% interval, because the probability of observing 0 events with a background expectation of \( \lambda = 4.0 \) is less than 5\%. While there is no logical difficulty with this result - the intervals are only designed to cover a true value in a fraction of all cases - the misuse of frequentist intervals (interpreting them as ranges where the parameters have high probability of having their true value) has led to the construction of different types of intervals. We consider a popular interval definition - the Feldman-Cousins Interval.

### 3.6 Feldman-Cousins Intervals

The Feldman-Cousins Interval is based on a different ranking of possible data outcomes. A new quantity is defined - the ratio of the probability assigned to a possible outcome to the probability of that outcome calculated with the value of the Poisson parameter that maximizes the outcomes probability, subject to any restrictions that might apply. Mathematically
\[
r = \frac{P(n|\mu = \lambda + \nu)}{P(n|\hat{\mu})}.
\] (3.8)
The outcomes \( n \) are then ranked according to \( r \) and the cumulative probability according to this ranking is used to form an interval. I.e., the set of outcomes included at \( 1 - \alpha \) probability follows the same logic as for the smallest interval, but instead of ranking results according to the probability of \( n \) given \( \mu \), the results are ranked according to the probability ratio \( r \).
To take a concrete example, we imagine that we have a fixed background expectation \( \lambda = 3.0 \). We then need to make a table of probabilities for outcomes \( n \) for possible \( \nu \). We choose for this example \( \nu = 1/3 \) so that we have the same table of outcomes as before. However, we now introduce additional columns to give the value of \( r \), the rank according to this value, and the cumulative probability according to the rank. The optimal value of \( \mu, \hat{\mu} \) is also listed. Note that we have the restriction \( \mu \geq \lambda \), so that \( \hat{\mu} \geq 3 \). The results are given in Table 3.2. For comparison, the Central Interval with probability \( 1 - \alpha \geq 0.9 \) contains the elements \( \mathcal{O}_{0.9}^C = \{1, \ldots, 7\} \) the Smallest Interval contains \( \mathcal{O}_{0.9}^S = \{1, \ldots, 6\} \) and the Feldman-Cousins interval \( \mathcal{O}_{0.9}^{FC} = \{0, \ldots, 6\} \). The three interval definitions each give a different set of values included in the \( 1 - \alpha \) set, indicating the importance of indicating the set construction used when quoting results.

Table 3.2: Values of \( n \), the probability to observe such a value given \( \mu = 3.3 \), the value of \( \mu \) that maximizes the probability of \( n \) given the constraint \( \mu \geq \lambda \), the probability of \( n \) given \( \hat{\mu} \), the ratio of the probabilities, the rank according to the probability ratio and the cumulative probability according to rank.

| \( n \) | \( P(n|\mu) \) | \( \hat{\mu} \) | \( P(n|\hat{\mu}) \) | \( r \) | Rank | \( F_R(n|\mu) \) |
|------|--------|------|------------|-----|------|-------------|
| 0    | 0.0357 | 3.0  | 0.050      | 0.717 | 5    | 0.7565      |
| 1    | 0.1189 | 3.0  | 0.149      | 0.796 | 4    | 0.7208      |
| 2    | 0.1982 | 3.0  | 0.224      | 0.885 | 3    | 0.6091      |
| 3    | 0.2202 | 3.0  | 0.224      | 0.983 | 1    | 0.2202      |
| 4    | 0.1835 | 4.0  | 0.195      | 0.941 | 2    | 0.4037      |
| 5    | 0.1223 | 5.0  | 0.175      | 0.699 | 6    | 0.8788      |
| 6    | 0.0680 | 6.0  | 0.161      | 0.422 | 7    | 0.9468      |
| 7    | 0.0324 | 7.0  | 0.149      | 0.217 | 8    | 0.9792      |
| 8    | 0.0135 | 8.0  | 0.140      | 0.096 | 9    | 0.9927      |
| 9    | 0.0050 | 9.0  | 0.132      | 0.038 | 10   | 0.9976      |
| 10   | 0.0017 | 10.0 | 0.125      | 0.014 | 11   | 0.9993      |
| 11   | 0.0005 | 11.0 | 0.119      | 0.004 | 12   | 0.9998      |
| 12   | 0.0001 | 12.0 | 0.114      | 0.001 | 13   | 1.0000      |

The Feldman-Cousins band plot for \( \lambda = 3.0 \) and \( 1 - \alpha = 0.9 \) is shown Fig. 3.6. With this result, we can now quote \( 1 - \alpha \) confidence level Feldman-Cousins intervals for the signal parameter \( \nu \). We see that a null interval is no longer possible.
for $\nu$, and for $n = 0$ we have $\nu \leq 1$ at 95% CL in the Feldman-Cousins construction for $\lambda = 3.0$. The upper value of $\nu$ at $1 - \alpha$ CL will still depend on the background expectation, however, and stronger signal limits will be achieved for experiments with larger backgrounds. This is again logically fine, since the intervals are not constructed to give the ranges where the parameters have high probabilities, but nevertheless leads to great confusion since it gives the impression that poorer experiments give stronger results.

Figure 3.6: lower and upper limits on the values of $n$ included in the Feldman-Cousins $1 - \alpha = 0.9$ set for different values of the signal expectation $\nu$ and for fixed background expectation $\lambda = 3.0$.

The example concerned the simplest case where the background expectation was known exactly. Treating the more realistic case of background known only within some uncertainty is treated below, after considering the Bayesian Analysis of Poisson processes in the presence of background.
3.7 Bayesian Analysis of Poisson Process with Background

Starting with the case of known background expectation $\lambda$, Bayes formula reads

$$P(\nu|n, \lambda) = \frac{P(n|\nu, \lambda)P_0(\nu)}{\int P(n|\nu, \lambda)P_0(\nu)d\nu} \tag{3.9}$$

Assuming our usual starting point of a flat distribution for $\nu$ up to some large $\nu_{\text{max}}$, we have

$$P(\nu|n, \lambda) = \frac{e^{-(\nu+\lambda)}(\lambda + \nu)^n}{\int_0^{\nu_{\text{max}}} e^{-(\nu+\lambda)}(\lambda + \nu)^n d\nu} \tag{3.10}$$

The integral in the denominator can be solved via integration by parts, taking $\nu_{\text{max}} \to \infty$, yielding

$$P(\nu|n, \lambda) = \frac{e^{-\nu}(\lambda + \nu)^n}{\sum_{i=0}^{n} \frac{\lambda^i}{i!}} \tag{3.11}$$

In your application, you should check whether the limit $\nu_{\text{max}} \to \infty$ is applicable. If it is not, you will need to perform the integration in the denominator numerically. The mode of the distribution in Eq. 3.11 occurs at $\nu^* = \max\{0, n - \lambda\}$ (the posterior probability is 0 for $\nu < 0$ because the prior is zero for negative $\nu$). If you choose to report the mode of the posterior as your best estimate for $\nu$, then you have the intuitive result that it is the number of observed events minus the expected background, or zero.

The cumulative probability is given by

$$F(\nu|n, \lambda) = 1 - \frac{e^{-\nu} \sum_{i=0}^{n} \frac{(\lambda+\nu)^i}{i!}}{n! \sum_{i=0}^{n} \frac{\lambda^i}{i!}} \tag{3.12}$$

and this result can be used to define probable ranges for $\nu$. Examples of these functions are shown in Fig. 3.7

3.7.1 Comparison of the Bayesian and Feldman-Cousins Intervals

Bayesian and frequentist intervals have different meaning, so in principle they should not be compared. However, since this has been the source of much discussion, we will also discuss the results here. Again, to clearly specify the definitions:
3.7. BAYESIAN ANALYSIS OF POISSON PROCESS WITH BACKGROUND

Figure 3.7: Posterior probability density for $\nu$ with $n = 5$ events observed and two different known background expectations (top). The cumulative of the posterior probability density (bottom).
• Frequentist intervals report the values of the parameters for which the observed data are members of predefined set of outcomes. The relative frequencies of possible data outcomes are fixed within the model for fixed values of the parameters, and we can define different constructions for specifying which data outcomes fall within the allowed set. We have seen three different constructions so far - the central probability interval, the smallest probability interval and the Feldman-Cousins construction. These are different ways of specifying which data outcomes will be considered in our set of outcomes, with the requirement that the total probability of the members of the set be at least $1 - \alpha$. I.e., in addition to the construction rules for the set, we also need to specify the probability content of the set. The set content is evaluated for all possible values of the parameters. Upon measuring a data outcome, you then keep all values of the parameters for which your observed data falls within the pre-defined sets.\footnote{As we have seen, it is possible that there are no allowed values of the parameters satisfying the criteria. This is fine, since the parameter values reported are not intended to be interpreted as probable values of the parameters. This is the point which generates great confusion, so please make sure you have understood this. The parameter ranges constructed in the frequentist way are intended to cover true values of the parameters in a statistical way - i.e., in $1 - \alpha$ of the experiments, the true value will be in your range. I have never seen anyone actually try to infer the true value of a parameter this way, so this point is not very important in practice, even if frequentists often lay great emphasis on this.}

• Bayesian posterior probabilities on the other hand are by construction designed to report the probability (degree-of-belief) that we assign to the possible parameter values. As we have seen, we can then report intervals specifying where we believe the correct value of the parameter is most likely to be. In contrast to the frequentist intervals, the resulting Bayesian intervals depend on the specification of prior information. This is a necessary condition for calculating posterior probabilities. For a given data outcome, there are as many possible resulting parameter ranges as there are prior possibilities - generally infinite. The prior choice should therefore always be explicitly given when quoting any result.

So there are deep conceptual differences in the meaning of the intervals, and it is therefore not surprising that the numerical values that result are also different. The Feldman-Cousins construction for $1 - \alpha = 0.9$ was used for a Poisson process and fixed background of $\lambda = 3.0$ and the results shown in Fig. 3.6. We look at the same case from a Bayesian analysis with flat prior and calculate the smallest
3.7. BAYESIAN ANALYSIS OF POISSON PROCESS WITH BACKGROUND

Probability intervals containing 90% probability. For this, we need to solve the set of equations

\[
\begin{align*}
F(\nu_2|n, \lambda = 3.0) - F(\nu_1|n, \lambda = 3.0) &= 0.90 \\
P(\nu_2|n, \lambda = 3.0) &= P(\nu_1|n, \lambda = 3.0)
\end{align*}
\]

This is done numerically, and the results are shown in Fig. 3.8.

Figure 3.8: Comparison of the Feldman-Cousins interval (red lines) and Bayesian (outer limits of interval given by the blue dots) assuming a flat prior and taking the smallest interval for fixed background expectation \(\lambda = 3.0\) and \(1 - \alpha = 0.9\).

As we see from the figure, for this set of conditions the ranges reported are, for \(n \geq 2\), quite similar. The main difference is for \(n = 0, 1\), where the upper included value of \(\nu\) is much lower for the FC interval than it is for the Bayesian interval. Since in this case \(n < \lambda\), we would report an upper limit, and the upper limit on the signal strength parameter would be numerically much more stringent in the frequentist analysis. This has led to some confusion, including statements that this is somehow a better result. This is of course nonsense, since we are comparing apples and oranges. As a recommendation, it would be wise to report the following:
• The data, including all relevant parameters needed to get from a signal strength parameter $\nu$ to a physical quantity;

• The background and how it was determined (and its uncertainty - see below)

• The expectations based on previous knowledge concerning $\nu$

• A frequentist interval specifying the range of $\nu$ for which the observed data is in the pre-defined set of outcomes. Here, the interval construction should be completely specified;

• A Bayesian interval on credible values of the parameters, specifying clearly the prior (or priors) chosen.

3.8 Uncertain Background

We usually do not know our background perfectly well, and either have additional data that helps us constrain it or other sources of information that allow us to specify a probability distribution for possible values of the background strength. We will now investigate how we handle this in our extraction of possible values for the signal strength. We start by pointing out that even with no knowledge of the background, we can still make statements about the signal strength. E.g., if we observe $n = 5$ events, then we know that $\nu \geq 10$ is very unlikely irrespective of the value of $\lambda$.

3.8.1 Frequentist Analysis

We postpone the discussion of a frequentist analysis of a Poisson process with uncertainty in both the signal and background until we introduce test statistics (which will allow us to return to 1D probability distributions).

3.8.2 Bayesian Analysis

There is no difficulty in generalizing the analysis of data for the case where also the background is unknown. In this case, we allow ourselves to learn about the background from the data by rewriting Bayes’ Formula as

$$ P(\nu, \lambda|n) = \frac{P(n|\nu, \lambda)P_0(\nu, \lambda)}{\int P(n|\nu, \lambda)P_0(\nu, \lambda)d\nu d\lambda} \quad (3.15) $$
3.9 Bayesian Analysis of the On/Off Problem

I.e., we get a 2-dimensional probability density from which we can extract information about both \( \nu \) and \( \lambda \). To proceed, we need to specify some priors. In the general case, the mathematics become intractable and we need to solve for the posterior pdf numerically. Let us start by considering a simple case where we specify flat priors for \( \nu \) and \( \lambda \) up to some large values. Then, performing the double integration in the denominator, we get

\[
P(\nu, \lambda | n) = \frac{e^{-(\nu+\lambda)}(\lambda + \nu)^n}{(n + 1)!}.
\]

(3.16)

We show this distribution for the case \( n = 5 \) in Fig. 3.9. We can find the 1-dimensional probability distribution for \( \nu \) by integrating out \( \lambda \) (this is called marginalization), and find:

\[
P(\nu | n) = \frac{e^{-\nu} \sum_{i=0}^{n} \nu^i}{n + 1}.
\]

(3.17)

The result for our example with \( n = 5 \) is also shown in Fig. 3.9. So we see that we can gain information on the possible values of the signal strength even when the background is not known. It is clear that the signal strength cannot be much larger than \( \nu = 5 \), independently of the background level.

3.9 Bayesian Analysis of the On/Off Problem

We now investigate the so-called On/Off problem, where we now have two data sets: one to help us learn about the background, and another where the background is possibly supplemented with a signal. As we have seen, when analyzing two data sets, we can either analyze them sequentially, or both at once. The general formula is again

\[
P(\vec{\nu}, \vec{\lambda} | \vec{n}) = \frac{P(\vec{n} | \vec{\nu}, \vec{\lambda}) P_{0}(\vec{\nu}, \vec{\lambda})}{\int P(\vec{n} | \vec{\nu}, \vec{\lambda}) P_{0}(\vec{\nu}, \vec{\lambda}) d\vec{\nu} d\vec{\lambda}}.
\]

(3.18)

where now we have indicated that there could be many data values and a corresponding signal and background expectation for each, which could be different if, e.g., the observation time is different. We assume that the background and signal rates are constants, so that
Figure 3.9: Two-dimensional posterior probability distribution for a Poisson process with flat priors for the signal and background and $n = 5$ events observed (left). The marginalized (over $\lambda$) distribution, showing the pdf for $\nu$ (right).
\[ \begin{align*}
\bar{\lambda} &= R_B \bar{T} \\
\bar{\nu} &= R_S \bar{T}
\end{align*} \]

with the different observation times given. So the uncertain parameters are \( R_B \), \( R_S \) and we can rewrite our Bayes’ equation as

\[ P(R_S, R_B|\bar{n}) = \frac{P(\bar{n}|R_S, R_B)P_0(R_S, R_B)}{\int P(\bar{n}|R_S, R_B)P_0(R_S, R_B)dR_SdR_B} \]. \tag{3.19} \]

Again, the general solution cannot usually be found analytically, and we need to resort to numerical methods. Taking a simple case, let’s assume we have two data sets. The first is used to determine the background; i.e., we look away from the source but with conditions that we assume will give the same background rate as when looking at the source. We observe this region for a time \( T_1 \) and observe \( n_1 \) events. Now we move to the source region and observe for a time \( T_2 \) and record \( n_2 \) events. Our probability of the data is therefore

\[ P(n_1, n_2|R_B, R_S) = \frac{e^{-R_B T_1}(R_B T_1)^{n_1}}{n_1!} \cdot \frac{e^{-(R_B+R_S) T_2}((R_B + R_S) T_2)^{n_2}}{n_2!} \]

If we can assume that the prior can be written as a constant for the \( R \)’s in the range of interest, then we have

\[ P(R_B, R_S|n_1, n_2) \propto e^{-R_B T_1}(R_B T_1)^{n_1}e^{-(R_B+R_S) T_2}((R_B + R_S) T_2)^{n_2} \]

where the expression is now a probability density for the parameters.

Here is a concrete example:

Data set 1 \( T_1 = 100 \quad n_1 = 100 \)

Data set 2 \( T_2 = 250 \quad n_2 = 300 \)

The shape of the posterior pdf is shown in Fig. ?? as a function of the rates for the background and signal.
Figure 3.10: Two-dimensional posterior probability distribution for a Poisson process with flat priors for the signal and background for the data given in the text. There are two pieces of data - one on source.
3.9. BAYESIAN ANALYSIS OF THE ON/OFF PROBLEM

3.9.1 Source discovery with On/Off data

Two hypotheses:

I. Only background, prior is $P_0(\lambda|I) = 1/\lambda_{\text{max}}$;

II. Signal+background, prior is

$$P_0(\lambda, \nu|\text{II}) = P_0(\nu|\text{II})P_0(\lambda|\text{II}) = 1/\lambda_{\text{max}}1/\nu_{\text{max}};$$

$$P(n_1, n_2|I) = \int P(n_1, n_2|\lambda)P_0(\lambda|I)d\lambda$$

$$= \frac{1}{n_1!n_2!\lambda_{\text{max}}} \int_0^{\lambda_{\text{max}}} e^{-2\lambda\lambda_{n_1+n_2}}d\lambda$$

for $\lambda_{\text{max}} \gg n_1 + n_2$, we have

$$P(n_1, n_2|I) \approx \left(\frac{1}{2}\right)^{n_1+n_2+1} \frac{(n_1 + n_2)!}{n_1!n_2!\lambda_{\text{max}}}$$

and for hypothesis II:

$$P(n_1, n_2|\text{II}) = \int P(n_1, n_2|\lambda, \nu)P_0(\lambda, \nu|\text{II})d\lambda d\nu$$

$$= \frac{1}{n_1!n_2!\lambda_{\text{max}}\nu_{\text{max}}} \int_0^{\nu_{\text{max}}} \int_0^{\lambda_{\text{max}}} e^{-2\lambda\lambda_{n_1+n_2}}(\lambda + \nu)^{n_1+n_2}d\lambda d\nu$$

The Bayes Factor for adding a second source of events (the signal) is suppressed by $\nu_{\text{max}}$. Note that there is a suppression also for $\nu_{\text{max}} < 1$ because of the limited range of the exponential.

Take as an example $n_1 = n_2 = 0^2$. Then we have

$$P(n_1, n_2|I) = \frac{1 - e^{-2\lambda_{\text{max}}}}{2n_1!n_2!\lambda_{\text{max}}}$$

$$P(n_1, n_2|\text{II}) = \frac{(1 - e^{-2\lambda_{\text{max}}})(1 - e^{-\nu_{\text{max}}})}{2n_1!n_2!\lambda_{\text{max}}\nu_{\text{max}}}$$

and the Bayes Factor is

$$BF = \frac{1 - e^{-\nu_{\text{max}}}}{\nu_{\text{max}}}$$

which is always less than one as intuitively expected.

\(^2\)this set of observations produces the strange result that hypothesis II is favoured in the non-factorized Jeffreys prior approach.
CHAPTER 3. POISSON DISTRIBUTION

3.9.2 On/off with factorized Jeffreys priors

We have the same hypotheses, but now we take $P_0(\lambda, I) = P_0(\nu|II)P_0(\lambda|II) \propto \lambda^{-1/2}\nu^{-1/2}$. Writing down the same equations as before:

$$P(n_1, n_2|I) = \int P(n_1, n_2|\lambda)P_0(\lambda|I)d\lambda$$

$$= \frac{1}{n_1!n_2!} \int_0^{\lambda_{\text{max}}} e^{-2\lambda n_1 + n_2 - 1/2}d\lambda$$

(3.28)

and for hypothesis II:

$$P(n_1, n_2|II) = \int P(n_1, n_2|\lambda, \nu)P_0(\lambda, \nu|II)d\lambda d\nu$$

$$= \frac{1}{n_1!n_2!} \int_0^{\lambda_{\text{max}}} \int_0^{\nu_{\text{max}}} e^{-2\lambda e^{-\nu} n_1 - 1/2(\lambda + \nu)}n_1 + n_2 - 1/2 d\lambda d\nu$$

(3.30)

If we now take $\nu_{\text{max}} \to \infty$ and $\lambda_{\text{max}} \to \infty$ (i.e., improper priors) and the special case $n_1 = n_2 = 0$, we find

$$P(n_1, n_2|I) = \frac{\sqrt\pi/2}{n_1!n_2!}$$

(3.32)

$$P(n_1, n_2|II) = \frac{\pi/\sqrt2}{n_1!n_2!}$$

(3.33)

and the Bayes factor is $BF = \sqrt\pi > 1$. I.e., we do not get a reasonable result just by using factorized priors. It is also important to have proper priors. For the same special case $n_1 = n_2 = 0$ but now taking $\lambda_{\text{max}}, \nu_{\text{max}}$ finite and normalising the priors, we find

$$P(n_1, n_2|I) = \frac{\sqrt\pi/2 \text{erf}(\sqrt2\lambda_{\text{max}})}{n_1!n_2! \sqrt{2\lambda_{\text{max}}}}$$

(3.34)

$$P(n_1, n_2|II) = \frac{\pi/\sqrt2 \text{erf}(\sqrt2\lambda_{\text{max}}) \text{erf}(\sqrt\nu_{\text{max}})}{n_1!n_2! \sqrt{2\lambda_{\text{max}}} \sqrt{2\nu_{\text{max}}}}$$

(3.35)

and the Bayes factor is

$$BF = \frac{\sqrt\pi \text{erf}(\sqrt\nu_{\text{max}})}{2\sqrt{\nu_{\text{max}}}}$$

(3.36)
3.9. BAYESIAN ANALYSIS OF THE ON/OFF PROBLEM

Figure 3.11: Comparison of Bayes Factors for different prior choices for the special case of no observed events in on/off runs.
which is always less than one as we expect. I.e., it is important to work with proper priors. A comparison of the two Bayes factors as a function of $\nu_{\text{max}}$ is shown in the figure.

So, the choice of priors is important, including the fact that the prior be a proper prior. It is possible to get sensible results for either the Jeffreys or flat priors, but an upper limit has to be set on the possible signal and background rates. This upper limit (and the choice of prior) should be in principle set case-by-case by the analyzer, and probably the best that can be done is to give some guidelines on appropriate choices.
3.9. BAYESIAN ANALYSIS OF THE ON/OFF PROBLEM

Literature


Allen Caldwell and Kevin Kröninger, Signal discovery in sparse spectra: A Bayesian analysis, Phys. Rev. D 74, 092003

Exercises

1. Derive the variance of an exponential distribution

2. Calculate the skew of the Poisson distribution

3. Find the distribution for the waiting time for the $k^{th}$ event in a process with constant rate $R$.

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?

5. For a Poisson process with mean 1.5, what is the probability to see 6 or more events? What is the probability of exactly 0 events?

6. Assume the probability distributions for the observed number of events in a measurement is described by a Poisson distribution. Assume a prior of the form $P_0(\lambda) \propto e^{-a\lambda}(a\lambda)^m$

   (a) Normalize the prior

   (b) Find the posterior pdf

   (c) For $n = m = 0$ and $a = 1$, find the 90% credible upper limit for the parameter

   (d) What physical situation could this refer to?

7. 9 events are observed in an experiment modeled with a Poisson probability distribution.

   (a) What is the 95% probability lower limit on the expected mean starting with a flat prior and assuming an underlying Poisson distribution?

   (b) What is the 68% confidence level interval for the mean of the Poisson distribution?
8. 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 $\lambda$

9. You have observed a part of the sky for $10^3$ s and observed 10 events.
   (a) Starting from Bayes Theorem and using a flat prior, calculate the probability distribution would you get for the rate of events?
   (b) What is the most probable value for the rate?
   (c) Find the 68% smallest interval.
   (d) Your colleague performed the experiment in a slightly different way. He waited until he saw the 10th event and then stopped his measurements. Call the time at which the 10th event arrived $T_2$. What is the probability distribution for the rate given $T_2$? What value of $T_2$ would have to be measured for you to get the same result for the most probable value of the rate?

10. Derive the Jeffreys prior (see Eq. (3.6) for the Poisson distribution (answer: $P_\theta(\nu) \propto \sqrt{\frac{1}{\nu}}$).

11. For the following data set:

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Source in/out</th>
<th>Run Time</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Out</td>
<td>1000 s</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>In</td>
<td>2000 s</td>
<td>250</td>
</tr>
</tbody>
</table>

   (a) Plot the probability distribution for the background rate from Data set 1 only
   (b) Analyze the two data sets simultaneously; plot the 2D probability density for the background and signal rates.
(c) Find the 68% central credibility interval for the decay rate.

(d) If your sample had a mass of one gram, and the isotope in the sample
has an atomic mass of $m_A = 110$ gm/mole, what is the lifetime of the
isotope (value with uncertainty)?

12. Perform the calculations necessary to derive formulas Eq. 3.16,3.17

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 $\lambda$ and defined and normalized in a range
$[a, b]$. $n$ events are measured with $x$ values $x_i$, $i = 1, \ldots, n$. The unbinned
likelihood is defined as the product of the densities

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

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 $\Delta_j$.
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 $\Delta$ 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.

14. In the Extended Likelihood approach, the overall normalization is taken into
account so as to have a penalty if the observed number of events is not in
3.9. BAYESIAN ANALYSIS OF THE ON/OFF PROBLEM

line with expectations. Referring to the notation in the previous problem, we define

\[ \Lambda = \int_a^b f(x|\lambda)dx \]

The normalized unbinned likelihood is now

\[ \mathcal{L}(\lambda)_{\text{normalized}} = \prod_{i=1}^{n} \frac{f(x_i|\lambda)}{\Lambda} \]

and the extended likelihood is defined as

\[ \mathcal{L}(\lambda)_{\text{extended}} = \frac{\Lambda^n e^{-\Lambda}}{n!} \prod_{i=1}^{n} \frac{f(x_i|\lambda)}{\Lambda} \]

Compare the value of the unbinned likelihood, the binned Poisson probability, and the binned Poisson probability for the data set in the table for the model \( f(x) = 10x \quad 0 \leq x \leq 1 \). Discuss the results.

<table>
<thead>
<tr>
<th>Set</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.5</td>
<td>0.75</td>
<td>0.9</td>
</tr>
</tbody>
</table>

15. Let’s investigate the relationship of a binned Poisson probability to a multinominal probability distribution. Using the notation above, we assume that the number of events \( n \) is fixed.

(a) Show that the expectation in a bin is \( \nu_j = p_j n \) where \( p_j \) is the multinominal probability for a measurement to be in bin \( j \).

(b) Show that the binned Poisson probability is related to the multinomial probability as

\[ P(\{n_j\}|\{\nu_j\}, n) = \prod_{j=1}^{K} \frac{e^{-\nu_j} \nu_j^{n_j}}{n_j!} = \frac{n^n e^{-n}}{n!} P(\{n_j\}|\{p_j\}, n) \]

where \( P(\{n_j\}|\{\nu_j\}, n) \) is the multinomial probability.
16. We consider a thinned Poisson process. Here we have a random number of occurrences, $N$, distributed according to a Poisson distribution with mean $\nu$. 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

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

17. For the Binomial case, the Conjugate Prior is the Beta Distribution:

$$P_0(p|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1 - p)^{\beta-1}$$

(a) What are $\alpha$ and $\beta$ if the prior is from the posterior pdf of an experiment where you started with a flat prior and had $r_0$ successes in $N_0$ trials?

(b) Prove that with a Binomial experiment with $N$ trials and $r$ successes, and using the prior above, the Posterior probability density is given by

$$P_0(p|\alpha, \beta, N, r) = \frac{\Gamma(\alpha + \beta + N)}{\Gamma(\alpha + r)\Gamma(\beta + N - r)} p^{r + \alpha - 1}(1 - p)^{N - r + \beta - 1}$$
Chapter 4

Gaussian Probability Distribution Function

The Gaussian probability distribution function (also called the Normal Distribution) is the most commonly used probability distribution. This is due to the fact that it is the limiting distribution for the Binomial and Poisson distributions when the expected number of events is large, and because of the Central Limit Theorem, which tells us that it is the limiting distribution for the sum (or average) of repeated samplings of the same distribution for any distribution satisfying rather mild conditions. It is also rather easy to use and its self-similar form means that having understood the Unit Normal distribution (the Gauss distribution with unit variance and mean zero) you know all the properties of the Gauss distribution. It appears in many contexts in a natural way, and has been the subject of nearly 300 years of study. In this chapter, we look at different derivations of the Gauss distribution to gain insight on when it is applicable to our analysis settings. The Central Limit Theorem is introduced and derived, and we show how it can be used in many cases. We then provide extensive samples of analyses based on Gaussian probability distributions for the data, and derive the chi squared distribution. This gives us a handy opportunity to discuss test statistics, model testing and \( p \)-values, which will be discussed in more detail in the next chapter.
4.1 The Gauss Distribution as the Limit of the Binomial Distribution

We have already seen in Section 2.4 that in the limit of $N \to \infty$ and $p, q$ not too small, the Binomial Distribution can be approximated by the Normal distribution. This is known as the de Moivre-Laplace Theorem. Specifically, de Moivre showed that the probability for the discrete variable counting the number of successes $r = Np + x\sqrt{Npq}$, where $N$ is the number of trials and $p = 1 - q$ is the success probability, for fixed $r$ and therefore fixed $x$, in the limit $N \to \infty$ is approximated by

$$P(r|N, p) = \binom{N}{r} p^r q^{N-r} \approx \frac{1}{\sqrt{2\pi Npq}} e^{-\frac{1}{2} \frac{(r-Np)^2}{Npq}} = \frac{1}{\sqrt{2\pi Npq}} e^{-\frac{x^2}{2}} \quad (4.1)$$

I.e., the Gauss probability distribution function can be used to estimate the Binomial probability. We note that this approximation should be used with care, as the tail probabilities can be far off if the $N$ is not large enough.

4.2 The Gauss Distribution as the Limit of the Poisson Distribution

Also for the Poisson distribution, we have seen that as the expectation value becomes large, the distribution for the number of events becomes more symmetric and resembles a Gaussian probability distribution function. We can make this connection more definite as follows. We let $x$ represent the fractional deviation between the number of events observed and the expectation:

$$x = \frac{\nu - n}{\nu} \quad \text{so that} \quad n = \nu(1 - x)$$

and substitute in the Poisson probability distribution

$$P(x|\nu) = \frac{e^{-\nu} \nu^{x(1-x)}}{(\nu(1-x))!}. \quad (4.2)$$

We now use Stirlings’ approximation for the factorial in the limit of large $n$:

$$n! \approx n^n e^{-n} \sqrt{2\pi n}$$
4.3. GAUSS’ DERIVATION OF THE GAUSS DISTRIBUTION

and get

\[ P(x|\nu) \approx \frac{e^{-\nu x}}{\sqrt{2\pi \nu (1-x)}} \left( \frac{\nu}{\nu(1-x)} \right)^{\nu(1-x)} = \frac{e^{-\nu x}}{\sqrt{2\pi \nu}} \left( \frac{1}{1-x} \right)^{\nu(1-x)+1/2}. \]  

(4.3)

Now note that the last piece can be rewritten as follows:

\[ \left( \frac{1}{1-x} \right)^{\nu(1-x)+1/2} = \exp \left[ -(\nu(1-x)+1/2) \ln(1-x) \right] \]

and now use the Taylor series expansion for \( \ln(1+x) \approx x - x^2/2 + O(x^3) \) which is valid for \( |x| < 1 \) to get

\[ \lim_{x \to 0} \{ \exp \left[ -(\nu(1-x)+1/2) \ln(1-x) \right] \} \approx \exp \left[ -\frac{\nu x^2}{2} + \nu x + x/2 \right] \]

so that we have (dropping the last term since \( \lim_{x \to 0} \exp(x/2) \approx 1 \))

\[ \lim_{x \to 0} \{ P(x|\nu) \} \approx \frac{1}{\sqrt{2\pi \nu}} e^{-\nu \bar{x}^2 \over 2} = \frac{1}{\sqrt{2\pi \nu}} e^{-\left(\nu-\nu^2\right)^2 \over 2\nu}. \]  

(4.4)

We will see later that we can also derive the Gaussian limit of the Poisson distribution using the Central Limit Theorem.

4.3 Gauss’ derivation of the Gauss distribution

Gauss considered the following problem: what is the form of the probability density function \( \phi(x_i - \mu) \) which gives a maximum for

\[ f(\vec{x}|\mu) = \prod_{i=1}^{n} \phi(x_i - \mu) \]  

(4.5)

when

\[ \mu = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \]

We recognize the product of probability densities as what we would call the probability of the data in a Bayesian analysis, and, when viewed as a function of \( \mu \), as the likelihood. The \( x_i \) are taken to be independent samplings from the same probability density \( \phi(x_i - \mu) \), and the observed average of the measurements
is $\bar{x}$. The form of the function $\phi$ is initially unspecified, but we will see that with Gauss’ optimization criterion we can show that the function is the Gauss probability distribution.

We start by taking the derivative of $f$ wrt to $\mu$ and requiring that it vanishes when $\mu = \bar{x}$:

$$
\left. \frac{df}{d\mu} \right|_{\mu=\bar{x}} = 0 \implies \sum_{i=1}^{n} \frac{d\phi(x_i - \mu)}{d\mu} \prod_{j=1, j\neq i}^{n} \phi(x_j - \mu) \bigg|_{\mu=\bar{x}} = 0.
$$

(4.6)

Now, assuming that $f(\vec{x}|\mu) \neq 0$, we divide by $f$ to get

$$
\sum_{i=1}^{n} \frac{\phi'(x_i - \mu)}{\phi(x_i - \mu)} \bigg|_{\mu=\bar{x}} = 0
$$

(4.7)

where

$$
\phi'(x_i - \mu) = \frac{d\phi(x_i - \mu)}{d\mu}.
$$

Define $\psi = \frac{\phi'(x_i - \bar{x})}{\phi(x_i - \bar{x})}$ and $z_i = x_i - \bar{x}$. Then we have

$$
\sum_{i=1}^{n} z_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} \psi(z_i) = 0
$$

for all possible values of $z_i$. This is only possible (assuming $\psi(z) \neq 0 \ \forall z$) if

$$
\psi(z_i) \propto z_i.
$$

We define the proportionality constant to be $-k$ and therefore have the differential equation

$$
\frac{d\phi(z)/dz}{\phi(z)} = -kz
$$

(4.8)

with solution

$$
\phi(z) \propto e^{-\frac{kz^2}{2}}.
$$

Requiring

$$
\int_{-\infty}^{\infty} \phi(z)dz = 1
$$

yields the constant of integration and we have finally

$$
\phi(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}}
$$
4.4. SOME PROPERTIES OF THE GAUSS DISTRIBUTION

and

\[ \phi(x - \mu) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]  

(4.9)

where we have set \( k = 1/\sigma^2 \). This is the famous Gauss probability distribution function.

4.4 Some properties of the Gauss distribution

The Gauss distribution is shown in Fig. 4.1. We will usually write the function in this way:

\[ G(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} \]  

(4.10)

Figure 4.1: Offset and scaled Gauss distribution on a linear scale (left) and logarithmic scale (right).

The cumulative of the Gauss distribution does not have a closed form expression, but is expressed in terms of the ‘error function’, defined as

\[ erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, dt \]  

A plot of the error function is shown in Fig. 4.2.
In terms of the error function, we write the cumulative of the Gauss function as

$$F(x|\mu, \sigma) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{x'-\mu}{\sigma})^2} dx' \quad (4.11)$$

$$= \frac{1}{2} \left( 1 + erf\left(\frac{x - \mu}{\sqrt{2}\sigma}\right) \right) \quad (4.12)$$

We give here some of the properties of the Gauss distribution:

**mode**

$$x^* = \mu$$

**Expectation value**

$$E[x] = \mu$$

**median**

$$F(x) = 0.5 \implies x = \mu$$

**variance**

$$Var[x] = \sigma^2$$

**FWHM**

The full width at half maximum is

$$FWHM = 2.355\sigma$$
### 4.5 Characteristic Function

We will make extensive use of the characteristic function in the next sections, so please look at this carefully. The characteristic function for a density \( p(x) \) is
defined as the expectation value of $\exp(ikx)$ for that density:

$$\phi(k) = \int_{-\infty}^{\infty} e^{ikx} p(x) dx$$  \hspace{1cm} (4.14)$$

which is, up to a factor, the Fourier transform of the probability density.

For discrete variables, the characteristic function can be written as

$$\phi(k) = \sum_i P(x_i) e^{ix_i k}$$

and for example for the Poisson case, we have

$$\phi(k) = \sum_{n=0}^{\infty} P(n) e^{ink}$$

which is just a Fourier series, so that we can get the $P(n)$ via the inverse transformation

$$P(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi(k) e^{-ink} dk .$$

We now go through a few properties of the Characteristic Function.

### 4.5.1 Sum of independent variables

The characteristic function of a sum of independent variables is the product of the individual Characteristic Functions.

**Probability density**

If two variables, here $x, y$, are independent, then the joint probability density factorizes:

$$P(x, y) = P_x(x)P_y(y) .$$  \hspace{1cm} (4.15)$$

The subscripts on the function names are there just to distinguish the functions. Now consider the Characteristic Function for

$$z = x + y$$
4.5. CHARACTERISTIC FUNCTION

\[ \phi_z(k) = \int e^{ikz} P_z(z) dz \]
\[ = \int e^{ikz} \left[ \int \delta(z - x - y) P(x, y) dx dy \right] dz \]
\[ = \int e^{ikz} \left[ \int \delta(z - x - y) P_x(x) P_y(y) dx dy \right] dz \]
\[ = \int \left[ \int e^{ikz} \delta(z - x - y) dz \right] P_x(x) P_y(y) dx dy \]
\[ = \int e^{ikx} P_x(x) dx \int e^{iky} P_y(y) dy \]
\[ = E[e^{ikx}] P_x(x) E[e^{iky}] P_y(y) \]
\[ = \phi_x(k) \phi_y(k) \]

so that we see that the characteristic function for the sum of independent variables is the product of the characteristic functions. This will allow us to show some neat things about the Gauss density function.

Discrete probability

We assume again that our variables \( x, y \), are independent, so that the joint probability density factorizes:

\[ P(x, y) = P_x(x) P_y(y) . \quad (4.16) \]

For
\[ z = x + y \]
we have
\[ P_z(z) = \sum_x P_x(x) P_y(y = z - x) \]
and
\[ \sum_z P_z(z) = \sum_z \sum_x P_x(x) P_y(y = z - x) \]
\[ = \sum_x P_x(x) \sum_z P_y(y = z - x) \]
\[ = \sum_x P_x(x) \sum_y P_y(y) \]
Now the characteristic function for $z$ is:

$$
\phi_z(k) = \sum_z e^{ikz} P_z(z)
$$

$$
= \sum_x \sum_y P_z(x) P_y(y) e^{ik(x+y)}
$$

$$
= \sum_x P_z(x) e^{ikx} \sum_y P_y(y) e^{iky}
$$

$$
= \phi_x(k) \phi_y(k)
$$

### 4.5.2 Characteristic Function of a Gaussian

We now calculate the Characteristic Function of a Gauss function:

$$
\phi(k) = \int e^{ikx} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx
$$

$$
= \frac{1}{\sqrt{2\pi\sigma}} \int e^{ik\mu - \frac{k^2x^2}{2\sigma^2}} e^{-\frac{1}{2} \left(z - \frac{\mu}{\sigma} + ik\sigma\right)^2} dx
$$

$$
= \frac{1}{\sqrt{2\pi\sigma}} e^{ik\mu - \frac{k^2x^2}{2}} \int e^{-\frac{1}{2} \left(z - \frac{\mu}{\sigma} + ik\sigma\right)^2} dx
$$

and now make use of the following integral:

$$
\int_{-\infty}^{\infty} e^{-\frac{z^2}{2\sigma^2}} dz = a\sqrt{\pi}
$$

where $z$ is complex to get

$$
\phi(k) = e^{ik\mu - \frac{k^2x^2}{2}}. \quad (4.17)
$$

This shows that the unit Normal distribution (where $\mu = 0$ and $\sigma = 1$) is an eigenfunction of the Fourier Transformation (see exercises).

### 4.5.3 Sum of Gauss distributed variables

We now combine the results from the two previous sections to find the probability distribution for the sum of two independent Gauss distributed variables: $z = x + y$.

$$
\phi_z(k) = \phi_x(k) \phi_y(k)
$$

$$
= e^{ik\mu_x - \frac{k^2x^2}{2}} e^{ik\mu_y - \frac{k^2y^2}{2}}
$$

$$
= e^{ik(\mu_x + \mu_y)} e^{-\frac{k^2(\sigma_x^2 + \sigma_y^2)}{2}}
$$
4.5. CHARACTERISTIC FUNCTION

We now see that this is the Characteristic Function for a Gauss function with

\[
\begin{align*}
\mu_z &= \mu_x + \mu_y \\
\sigma_z^2 &= \sigma_x^2 + \sigma_y^2
\end{align*}
\]

so that the probability distribution for the sum of two Gauss distributed random variables is another Gauss distributed variable. We can go from this to the general case of adding any number of Gauss distributed independent variables. The mean of the resulting distribution will be the sum of the individual means, and the variance will be the sum of the variances. If

\[
x = \sum_{i=1}^{n} x_i
\]

\[
P(x_1, x_2, \ldots, x_n) = \prod_{i} P_i(x_i)
\]

\[
P(x_i) = \mathcal{G}(x_i|\mu_i, \sigma_i^2)
\]

then

\[
P(x) = \mathcal{G}(x|\mu, \sigma^2)
\]

\[
\mu = \sum_{i} \mu_i
\]

\[
\sigma^2 = \sum_{i} \sigma_i^2.
\]

This is the source of ‘adding uncertainties in quadrature’ discussed below.

4.5.4 Graphical Example

We give a graphical example in Fig. 4.3 to make this all more concrete. As is seen in the figure, the distribution of the sum of two Gauss distributions behaves according to the formulas above. Since the sum of two Gauss distributed variables yields a Gauss distributed variable, we can conclude that the sum of any number of independent Gauss distributed variables yields a Gauss distributed variable with

\[
\mu = \sum_{i=1}^{n} \mu_i \quad \text{(4.18)}
\]

\[
\sigma^2 = \sum_{i=1}^{n} \sigma_i^2 \quad \text{(4.19)}
\]
4.5.5 Adding uncertainties in Quadrature

Imagine you are making a measurement of a quantity $x$, and that there are several sources of uncertainty in your measurement.

$$x_{\text{meas}} = x_{\text{true}} + \delta_1 + \delta_2 + \cdots + \delta_n$$

Examples of sources for the shifts of the measured quantity from the 'true' or 'source' value could be resolution effects in your detector, an unknown efficiency, an unknown calibration, an unknown background, etc. In each case, you do the best you can in accounting for each of the unknown effects so that you would say the shift from the correct value should be centered on zero and could be described by a Gauss function. I.e., your model is

$$P(\delta_i) = G(\delta_i|0, \sigma_i)$$

by which we mean the $\delta_i$ are random numbers centered on 0 and following a Gauss distribution with variance $\sigma_i^2$. In this case, the probability distribution for $x_{\text{meas}}$ is then a Gauss function with

$$P(x_{\text{meas}}) = G(x_{\text{meas}}|x_{\text{true}}, \sigma)$$
This is known as addition in quadrature, since the uncertainties are added in quadrature. The distribution of the measured variable follows a Gaussian probability distribution function around the true value.

Usually the assumption of a Gauss distribution to represent the effect of the resolution of a detector is justified from knowing that the limit of a Poisson distribution is a Gauss distribution. Where does the Poisson distribution come from? Imagine that your detector measures an energy, say by counting the number of visible photons produced. The number of visible photons produced is usually very large, of the order $E_{\text{incident}}/(\text{few eV})$, so for an incident energy of MeV say, on the order of a million photons are produced. The chance to see any one of these is very small (many are reabsorbed, your photon counter only sees a small solid angle, the photon counter has a small efficiency, etc.) so that the number of photons actually detected by your detector is Poisson distributed. If the number of photons detected is reasonably large, say at least 20 or 25, then as we have discussed the distribution can be approximated with a Gauss probability distribution. Your measured energy is then quoted as

$$E_{\text{meas}} = A n_{\text{detected}}$$

where $A$ is a calibration factor converting from the number of observed photons, $n_{\text{detected}}$, to an energy, and the distribution of

$$P(E_{\text{meas}} - E_{\text{incident}}) \approx G(\delta E | 0, A \sqrt{n}) . \quad (4.20)$$

The justification for using Gauss probability distributions for the so-called systematic uncertainties that result from mis calibration, incorrect determination of the efficiencies, etc. is usually not so clear, and Gauss distributions are used with vague reference to the Central Limit Theorem (see below). In actuality, we typically do not have any indication of the correct probability distribution to use for these systematic uncertainties, and use Gauss functions because they are convenient. The determination of the variance of the chosen Gauss function is part of the experimental art and relies on experience and judgment.
4.6 Adding Correlated Variables

We can extend this to the case where the variables are correlated, so that we cannot factorize the probability density as in Eq. (4.16). We define a correlation coefficient as follows:

\[ \rho_{xy} = \frac{\text{cov}[x, y]}{\sigma_x \sigma_y} \]

with

\[
\text{cov}[x, y] = E[(x - E[x])(y - E[y])] P(x, y) \\
= E[xy] P(x, y) - E[x] E[y] P_x P_y
\]

The meaning of \( E[\cdot]_{P(\cdot)} \) is that the expectation value is to be evaluated using the probability density \( P(\cdot) \). The probability densities \( P_x(x) \), \( P_y(y) \) are the so-called marginal distributions. Marginalization in this case is imply the integration over the unwanted variables:

\[
P_i(x_i) \equiv \int P(x_1, x_2, \ldots, x_n) \prod_{j \neq i} dx_j \quad (4.21)
\]

With this definition, we have for the sum of correlated random variables:

\[
x \equiv \sum_i x_i \quad (4.22)
\]

\[
E[x] = \sum_{i=1}^n E[x_i] \quad (4.23)
\]

\[
\text{Var}(x) = \sum_{i=1}^n \sum_{j=1}^n \text{cov}(x_i, x_j) \quad (4.24)
\]

\[
= \sum_{i=1}^n \text{Var}(x_i) + 2 \sum_{i<j}^n \text{cov}(x_i, x_j) \quad (4.25)
\]

The proof is saved for the exercises.

4.6.1 Correlation Coefficient - examples

Several examples of 2-dimensional correlation are shown in Fig. 4.4. The correlation coefficient ranges from \(-1 \leq \rho \leq 1\), where \( \rho = -1 \) implies complete anti
4.6. ADDING CORRELATED VARIABLES

correlation, $\rho = 0$ to no correlation, and $\rho = 1$ to complete positive correlation. The contours indicated in the plot satisfy the condition

$$P(x, y) = \text{constant}$$

$$\int_{\text{boundary}} P(x, y) \, dx \, dy = \text{fixed probability}$$

so that a fixed probability is contained within the contour (say 68%, e.g.) and all points on the boundary of the contour have the same probability density. Note that the slope of the plots is not really relevant since they can be changed with a change of scale. Nevertheless, the plots are intended to give some intuition about the meaning of the correlation coefficient.

Figure 4.4: Graphical depictions of correlated probability densities.

4.6.2 Sum of Correlated Gauss Distributed Variables

We can show that the resulting probability distribution is Gauss distributed as follows. The joint probability distribution for correlated Gauss variables is written as
\[ P(x_1, \ldots, x_n) = \frac{1}{(2\pi|\Sigma|)^{n/2}} e^{-\frac{1}{2}(\bar{x} - \bar{\mu})^T \Sigma^{-1} (\bar{x} - \bar{\mu})} \]  

(4.26)

where \( \Sigma \) is the covariance matrix

\[ \Sigma_{ij} = \text{cov}(x_i, x_j) \]

and \(|\Sigma|\) is the determinant of the matrix. We can diagonalize \( \Sigma \) with a similarity transformation

\[
\begin{align*}
\Sigma_y &= A \Sigma A^{-1} \\
\bar{\mu}_y &= A \bar{\mu} \\
\bar{y} &= A \bar{x}
\end{align*}
\]

and since the covariance matrix for \( \bar{y} \) is diagonal we can use our previous results to find the parameters of the resulting Gauss distribution. We have for

\[ x \equiv \sum_i x_i \]

that

\[
\begin{align*}
E[x] &= \sum_i E[x_i] \\
\text{Var}(x) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \text{cov}(x_i, x_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij} \sigma_i \sigma_j \\
&= \sum_{i=1}^{n} \sigma_i^2 + 2 \sum_{i<j}^{n} \rho_{ij} \sigma_i \sigma_j .
\end{align*}
\]

### 4.7 Central Limit Theorem

We are now ready to attack the Central Limit Theorem, one of the most important results you will learn in this course. Please pay close attention to the limiting conditions under which it is derived, since it is not always valid and often misused.
4.7. CENTRAL LIMIT THEOREM

We consider making \( n \) independent measurements of the same real-valued quantity, \( x \), which has a probability density \( P(x) \) (we will look at discrete variables a little further on). The average of our measurements is

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .
\]

The expectation value for \( \bar{x} \) is

\[
E[\bar{x}] = E\left[\frac{1}{n} \sum_{i=1}^{n} x_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[x_i] = E[x] = \mu .
\]

We now consider the probability distribution for \( z = \bar{x} - \mu \), the distribution of the observed average from the expected average. Let’s use \( P(z) \) to denote this probability density. Since we can write \( z \) as the sum

\[
z = \sum_{i=1}^{n} \frac{x_i - \mu}{n}
\]

the characteristic function for \( P(z) \) is the product of the characteristic functions for each term

\[
\phi_z(k) = \prod_{i=1}^{n} \phi_{\frac{x_i - \mu}{n}}(k)
\]

(4.27)

Now

\[
\phi_{\frac{x - \mu}{n}}(k) = \int e^{ik\left(x - \mu\right)} P(x)dx .
\]

(4.28)

We note that since \( P(x) \) is normalized and the norm of the integrand is bounded by 1, then the characteristic function is finite for all values of \( k \). We expand the exponential and evaluate the individual terms that arise

\[
\int e^{ik(x - \mu)} P(x)dx = \int \left(1 + ik\frac{E[x - \mu]}{n} - \frac{k^2}{2n^2}(x - \mu)^2 + \frac{ik^3}{6n^3}(x - \mu)^3 \ldots\right) P(x)dx
\]

\[
= 1 + ik\frac{E[x - \mu]}{n} - \frac{k^2}{2n^2}E[(x - \mu)^2] + O\left(\frac{E[(x - \mu)^3]}{n^3}\right) + \ldots
\]

The second term is obviously 0. If we take the limit \( n \to \infty \) and assuming that the moments of \( P(x) \) are finite, then we can drop the higher order terms and we are left with
\[ \phi_{x-\mu}(k) \approx 1 - \frac{k^2}{2n^2\sigma_x^2}. \]  

(4.29)

Then

\[ \phi_z(k) = \prod_{i=1}^{n} \phi_{x_i-\mu}(k) \]  

(4.30)

\[ \lim_{n \to \infty} \phi_z(k) \approx \lim_{n \to \infty} \left(1 - \frac{k^2}{2n^2\sigma_x^2}\right)^n \]  

(4.31)

\[ \approx e^{-\frac{k^2}{2n\sigma_x^2}}. \]  

(4.32)

Comparing with Eq. (4.17), we see that this is the Characteristic Function of a Gauss distribution with mean 0 and variance \( \sigma_x^2 \). We therefore conclude that

\[
P(z) = \frac{1}{\sqrt{2\pi\sigma_z}} e^{-\frac{1}{2} \frac{z^2}{\sigma_z^2}}
\]  

(4.33)

with \( \sigma_z = \sigma_x/\sqrt{n} \). I.e., the probability distribution for the deviation of the mean of our measurements from the expected mean is a Gauss distribution, and the width of the distribution scales as \( 1/\sqrt{n} \).

If we considered \( z = \bar{x} \) rather than \( z = \bar{x} - \mu \), then we would have

\[
P(z) = \frac{1}{\sqrt{2\pi\sigma_z}} e^{-\frac{1}{2} \frac{(z-\mu)^2}{\sigma_z^2}}
\]  

(4.34)

and if we considered \( z = \sum_{i=1}^{n} x_i \) then we would again have a Gauss distribution with

\[
P(z) = \frac{1}{\sqrt{2\pi\sigma_z}} e^{-\frac{1}{2} \frac{(z-n\mu)^2}{\sigma_z^2}}
\]  

(4.35)

where now \( \sigma_z = \sqrt{n\sigma_x} \).

### 4.7.1 Example of Central Limit Theorem

We illustrate the CLT with a simple example, shown in Figs. 4.5, 4.6. We consider the uniform probability distribution

\[
P(x) = 1 \quad 0 \leq x \leq 1
\]
4.8. CENTRAL LIMIT THEOREM FOR POISSON DISTRIBUTION

which has

\[ \mu_x = \frac{1}{2} \quad \sigma_x = \frac{1}{\sqrt{12}} \]

and perform 1000 experiments where for each experiment we have 10 samples of \( x \). The results for the first 10 experiments are shown in Fig. 4.5. The individual values of \( x \) are shown as well as \( \bar{x} \), which is denoted with the red bar.

We now plot the distribution of \( \bar{x} \) for the 1000 experiments in Fig. 4.6 and compare to the expectation from the CLT, which is a Gauss distribution with mean \( \mu_z = \mu_x = 1/2 \) and standard distribution \( \sigma_z = \sigma_x / \sqrt{n} = \frac{1}{\sqrt{12\sqrt{10}}} \) and find excellent agreement. I.e., the CLT works very well already for the rather small value of \( n = 10 \). The convergence to the Gaussian distribution is quick in this case because of the small higher moments of \( P(x) \). You can try out the CLT for yourself in the exercises.

4.8 Central Limit Theorem for Poisson Distribution

The CLT can also be applied to discrete distributions, but we have to be particularly careful here of the conditions under which it applies. We describe its use for the Poisson distribution. The application to a Binomial distribution is left for the exercises.

As noted above, for the Poisson case, we have

\[ \phi(k) = \sum_{n=0}^{\infty} P(n) e^{ink} \]

Let’s evaluate the characteristic function of the Poisson distribution:

\[
\begin{align*}
\phi(k) &= \sum_{n=0}^{\infty} P(n) e^{ink} \\
&= \sum_{n=0}^{\infty} e^{-\nu} \frac{\nu^n}{n!} e^{ink} \\
&= e^{-\nu} \sum_{n=0}^{\infty} \frac{(e^{ik} \nu)^n}{n!} \\
&= e^{-\nu} e^{\nu e^{ik}} \\
&= e^{\nu(e^{ik}-1)}
\end{align*}
\]
Figure 4.5: Results for ten experiments where we sample from the uniform distribution. The black bars show the measured values of $x$, while the red bar shows the average of the 10 measurements in the experiment.
4.8. CENTRAL LIMIT THEOREM FOR POISSON DISTRIBUTION

![Figure 4.6: Distribution of \( \bar{x} \) for the 1000 experiments compared to the expectation from the Central Limit Theorem.](image)

We can see the parallel to our development of the CLT for a continuous variable as follows. As we have seen in the previous chapter, the probability distribution for events coming from different Poisson processes is again a Poisson probability distribution with expectation value given by the sum of the individual processes. We imagine that we have a total expectation \( \nu \), and break this down into \( n \) subprocesses, each with rate \( \nu/n \) and now consider that our observed number of events

\[
X = \sum_{j=1}^{n} X_j
\]

where \( X_j \) is the number of events for subprocess \( j \). The probability distribution of \( X \) will be Poisson distributed with expectation value \( \nu \), which is what we want. Now using our result for the characteristic function of the Poisson distribution, we have

\[
\phi_X(k) = \prod_{j=1}^{n} \phi_{X_j}(k)
\]

\[
= \prod_{j=1}^{n} e^{\frac{\nu}{n}(e^{ik} - 1)}
\]

\[
= e^{\nu(e^{ik} - 1)}.
\]
Let’s expand the exponential in the exponent:

\[ e^{ik - 1} \approx ik - k^2/2 + \ldots \]

which is 0 for \( k = 0 \) else the real part is negative. For \textbf{large} \( \nu \), the characteristic function of \( X \) will therefore be strongly suppressed for all values of \( k \) not near 0, so we keep only the first terms in the expansion

\[ \phi_X(k) \approx e^{ik\nu - k^2\nu/2} \]

which is the characteristic function of a Gauss function with mean \( \nu \) and variance \( \nu \) and we therefore have

\[ P(X|\nu) \approx \frac{1}{\sqrt{2\pi \nu}} e^{-\frac{(X-\nu)^2}{2\nu}}. \]

Note that we explicitly assumed that \( \nu \) is large to get the Gaussian approximation for the probability density. This is required even though \( n \) is arbitrarily large and the moments of \( P(X_i) \) are arbitrarily small.

The Poisson distribution for \( \nu = 50 \) is compared to the Gauss probability density function in Fig. 4.7. As can be seen in the figure, the agreement is quite good, particular in the heart of the distributions. There is a clear discrepancy in the tails, where the Gauss function overshoots at smaller values of \( n \) and under predicts the probability at large \( n \). Since the means of the distributions are the same, there is also a slight shift in the central part of the distribution. For \( n = 50 \), the Poisson probability is 0.0563 and the Gauss density at that point is 0.0564; i.e., the agreement is much better than 1%. If we look at \( n = \nu \pm 20 \), we see a much bigger disagreement. For \( n = 30 \), the Poisson probability is 6.8 \cdot 10^{-4} while at \( n = 70 \), it is 1.36 \cdot 10^{-3}. In both cases, the Gauss density is 1.0 \cdot 10^{-3}, so we see that at these values, roughly three standard deviations from the peak, we have discrepancies approaching 50%. This is true in general - using the Gauss probability density distribution as an approximation for the probability for the Poisson or Binomial distribution will be much worse in the tails of the distribution.

### 4.9 Practical use of the Gauss function for a Poisson or Binomial probability

We have seen that the Poisson and Binomial probabilities can be approximated with a Gauss function in the right limit. For the Binomial case, we set the mean
Figure 4.7: The Poisson probability for a given number of events, \( n \), assuming \( \nu = 50 \), is shown as the solid symbols. The Gauss probability density function for \( \mu = \nu \) and \( \sigma^2 = \nu \) is overlaid.
of the Gauss to be $Np$ and the variance to $Npq$, whereas for the Poisson case we set the mean to $\nu$ and the variance also to $\nu$. In Fig. 4.8 we show some example distributions, where in each case the distributions have the same mean but the variance of the Gauss function is set to the variance of the Binomial distribution. It is clear that the approximation with a Gauss function is not very good for small event counts, but improves for larger expectation and more symmetric distributions. Table 4.2 gives some example results.

Table 4.2: Comparison of the Binomial, Poisson and Gauss distributions for the examples given in Fig. 4.8.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$E[r]$</th>
<th>Var$r$</th>
<th>$E[(r - E[r])^2]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial $N = 2, p = 0.9$</td>
<td>1.8</td>
<td>0.18</td>
<td>-0.14</td>
</tr>
<tr>
<td>Poisson $\nu = 1.8$</td>
<td>1.8</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>Gauss $\mu = 1.8, \sigma^2 = 1.8$</td>
<td>1.8</td>
<td>1.8</td>
<td>0</td>
</tr>
<tr>
<td>Binomial $N = 10, p = 0.4$</td>
<td>4</td>
<td>2.4</td>
<td>0.48</td>
</tr>
<tr>
<td>Poisson $\nu = 4$</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Gauss $\mu = 4, \sigma^2 = 2.4$</td>
<td>4</td>
<td>2.4</td>
<td>0</td>
</tr>
<tr>
<td>Binomial $N = 100, p = 0.1$</td>
<td>10</td>
<td>9</td>
<td>7.2</td>
</tr>
<tr>
<td>Poisson $\nu = 10$</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Gauss $\mu = 10, \sigma^2 = 9$</td>
<td>10</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

4.10 Frequentist Analysis with Gauss distributed data

4.10.1 Neyman confidence level intervals

The Gauss distribution is specified by two parameters, the mean and the standard deviation, and in general if both are unknown then we should make our band plot allowing both to vary. This leads to some difficulties and we put off this discussion till later. We initially consider that the width of the Gaussian distribution is perfectly known.

We now look at the probability distribution for results $x$ for different choices of $\mu$. The central and smallest intervals containing $1 - \alpha$ probability are the same for the Gauss distribution since we have a symmetric unimodal distribution. Furthermore, we can solve exactly for the values of $x$ which define the boundaries of
4.10. FREQUENTIST ANALYSIS WITH GAUSS DISTRIBUTED DATA

Figure 4.8: Top Left: Comparison of the Binomial, Poisson and Gauss distributions each with mean 1.8. Top Right: Comparison of the Binomial, Poisson and Gauss distributions each with mean 4.0. Bottom: Comparison of the Binomial, Poisson and Gauss distributions each with mean 10. In each case, the variance of the Gauss is set to the variance of the Binomial distribution.
our intervals:

\[
\frac{\alpha}{2} = F(x_1|\mu, \sigma) = 1 - F(x_2|\mu, \sigma) .
\]

Note that the integral

\[
\frac{\alpha}{2} = \int_{-\infty}^{x_1} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = \int_{-\infty}^{(\mu-x_1)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \, dy
\]
depends only on the quantity \((x_1 - \mu)/\sigma\) and similarly

\[
\frac{\alpha}{2} = \int_{x_2}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = \int_{(x_2-\mu)/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \, dy
\]

so that the intervals \([x_1, x_2] = [\mu - N\sigma, \mu + N\sigma]\) where \(N\) is a real number will give us our interval, with \(N\) defined by the choice of \(1 - \alpha\). Choosing \(N = 1\), we have \(1 - \alpha = 0.68\). Other common choices are given in Table 4.1.

The band plot for the case \(\sigma = 1\) is shown in Fig. 4.9. If we now measure a value for \(x\), and assuming we take \(N = 1\), it will be in the interval \([\mu - \sigma, \mu + \sigma]\) if and only if \(\mu\) is in the range \([x - \sigma, x + \sigma]\). The example \(x = 1\) is shown in the figure.

Now imagine we made many measurements of the same quantity, each time with the same resolution \(\sigma = 1\). Each measurement \(i\) would yield a \(1 - \alpha = 0.68\) confidence interval for \(\mu\) of \(\mu \in [x_i-1, x_i+1]\). As an example, 100 such results are shown in Fig. 4.10, where the data have been generated on the computer assuming \(\mu_{\text{true}} = 10\). The point at the center of the interval represents the value of \(x_i\), but the intervals are to be understood as belonging to \(\mu\). The true value of \(\mu\) is to be inferred from this plot. In the limit of an infinite number of measurements, it would be the value of \(\mu\) which lies in the specified intervals exactly \(1 - \alpha\) of the time. Since we never have an infinite number of measurements, a different combination technique is needed. There is no procedure for combining confidence intervals to result in a new confidence interval, so the best approach is to analyze all data in one pass and derive a new confidence interval from the total data set.

### 4.10.2 Feldman-Cousins confidence level intervals

As with the case of Poisson distributed signal and background, we can run into non-intuitive results using the Neyman construction for Gaussian distributed data. Consider Fig. 4.9 and imagine that our parameter \(\mu\) is required to be non-negative for physical reasons. E.g., it might correspond to the mass of a particle or a cross...
Figure 4.9: Neyman construction for the $1 - \alpha = 0.68$ (central, smallest) intervals
Figure 4.10: Neyman 68 % intervals for $\mu$ for 100 experiments. The data were generated on the computer under the assumption $\mu_{\text{true}} = 10$ and $\sigma = 1$. The red line indicates the value of $\mu_{\text{true}}$. 
4.11. BAYESIAN ANALYSIS OF GAUSS DISTRIBUTED DATA

A reaction section for a reaction to take place. If the value of $x$ which is measured is e.g. $x = -2$, which can routinely happen due to resolution effects, then the $1 - \alpha = 0.68$ interval for $\mu$ from the Neyman construction will be empty. Again, there is no logical difficulty for this - we just realize that there is no value of $\mu$ for which the data is within the central $1 - \alpha$ probability interval. However, since these results are often misinterpreted as a statement on probable values of $\mu$, we again have a non-intuitive result. This does not happen with the Feldman-Cousins construction as we demonstrate.

Let us put the physical boundary at $\mu \geq 0$. Then, for $x \geq 0$, we have $\hat{\mu} = x$ since that is the value of $\mu$ which will maximize $G(x|\mu, \sigma)$. The Feldman-Cousins variable $r$ is then

$$r = \frac{G(x|\mu, \sigma)}{G(x|\hat{\mu}, \sigma)} = e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} \quad x \geq 0$$

For $x < 0$, we have $\hat{\mu} = 0$. The Feldman-Cousins variable $r$ is then

$$r = e^{x\mu/\sigma^2 - \mu^2/2\sigma^2} \quad x < 0.$$ 

The resulting value of $r(x)$ for $\mu = 1.0, \sigma = 1$ is shown in Fig. 4.11. As can be seen in the figure, there is a long tail to negative $x$, which then implies that negative $x$ values will get a higher rank then they would if the standard Gauss probability was used. This in turn means that the intervals for the FC interval definition will be shifted relative to the Neyman intervals. For $\mu = 1, \sigma = 1$, we find ...

### 4.11 Bayesian analysis of Gauss distributed data

We now turn to a Bayesian analysis, which allows us to make statements about credible values of the parameters of the Gaussian probability distribution function. We start with the case treated above - that we have one measurement of a quantity, $x$ that is expected to be distributed according to a Gaussian pdf, and that the width of the Gaussian is perfectly known. There is then only one free parameter, the mean of the Gauss pdf. Bayes’ Theorem then tells us

$$P(\mu|x, \sigma) = \frac{P(x|\mu, \sigma)P_0(\mu)}{\int P(x|\mu, \sigma)P_0(\mu)d\mu}$$

(4.41)
Figure 4.11: The Feldman-Cousins ranking variable $r$ as a function of $x$ for $\mu = 1$ and $\sigma = 1$ (solid curve). The Gaussian pdf $G(x|\mu = 1, \sigma = 1)$ is shown as the dashed line for comparison.
Let us see what happens if we take a flat prior for $\mu$ which extends well beyond measurable values of $x$. Then

$$\int_{\mu_{\text{min}}}^{\mu_{\text{max}}} P(x|\mu,\sigma)P_0(\mu)d\mu = \int_{\mu_{\text{min}}}^{\mu_{\text{max}}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \frac{1}{\mu_{\text{max}} - \mu_{\text{min}}} d\mu$$

$$\approx \frac{1}{\mu_{\text{max}} - \mu_{\text{min}}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] d\mu$$

$$= \frac{1}{\mu_{\text{max}} - \mu_{\text{min}}}$$

$$= P_0(\mu).$$

We therefore find

$$P(\mu|x,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(4.42)

which is the same function as $P(x|\mu,\sigma)$ but now viewed as a function of $\mu$ with $x$ fixed. So, the probability distribution for $\mu$ is a Gaussian probability distribution centered on $x$ with variance $\sigma^2$. We therefore have as central (and smallest) probability intervals for $\mu$ the values in Table 4.3.

Table 4.3: The smallest (and central) range of $\mu$ values to get the desired probability content.

<table>
<thead>
<tr>
<th>Probability Content (in %)</th>
<th>$\mu$ range</th>
</tr>
</thead>
<tbody>
<tr>
<td>68.3</td>
<td>$x \pm \sigma$</td>
</tr>
<tr>
<td>90.0</td>
<td>$x \pm 1.65\sigma$</td>
</tr>
<tr>
<td>95.0</td>
<td>$x \pm 1.96\sigma$</td>
</tr>
<tr>
<td>99.0</td>
<td>$x \pm 2.58\sigma$</td>
</tr>
<tr>
<td>99.7</td>
<td>$x \pm 3\sigma$</td>
</tr>
</tbody>
</table>

What if we make a second measurement, $x_2$, which is of the same quantity and we assume that we again have a Gaussian probability distribution for $x_2$ centered on the true value? The resolution of our measuring apparatus could however be different, so that we have for the first measurement a sampling from a Gauss distribution with mean $\mu$ and width $\sigma_1$ and for the second measurement a sampling from a Gauss distribution with mean $\mu$ and width $\sigma_2$. We use the posterior
probability density for \( \mu \) from measurement \( x_1 \) as the prior for the analysis of the second measurement:

\[
P_2(\mu|x_2) = \frac{P(x_2|\mu, \sigma_2)P_1(\mu)}{\int P(x_2|\mu, \sigma_2)P_1(\mu) d\mu} \tag{4.43}
\]

where we take

\[
P_1(\mu) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu)^2}{2\sigma_1^2}}
\]

and

\[
P(x_2|\mu, \sigma_2) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_2-\mu)^2}{2\sigma_2^2}}.
\]

The integral in the denominator of Eq. (4.43) is

\[
\int \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{(x_2-\mu)^2}{2\sigma_2^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma_1^2}} d\mu
\]

which can be solved by ‘completing the square’. We skip the somewhat tedious calculation and quote the result for the posterior probability distribution for \( \mu \) using both measured data points:

\[
P(\mu|x_1, x_2, \sigma_1, \sigma_2) = \frac{1}{\sqrt{2\pi}\sigma_A} e^{-\frac{(x_A-\mu)^2}{2\sigma_A^2}} \tag{4.44}
\]

with

\[
x_A = \frac{x_1/\sigma_1^2 + x_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} \tag{4.45}
\]

\[
\frac{1}{\sigma_A^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \tag{4.46}
\]

We find that we again have a Gaussian probability distribution function for \( \mu \) which is centered on a weighted sum of the two measurements, where the weight is the inverse of the variance assumed for the measurements. The variance of the resulting Gauss distribution for \( \mu \) is found by adding the inverses of the two variances. It is then immediately obvious how to add more measurements and get the probability distribution for \( \mu \). The result will be of the form given in Eq. (4.44),
4.11. BAYESIAN ANALYSIS OF GAUSS DISTRIBUTED DATA

with

\[
x_A = \frac{\sum_i x_i / \sigma_i^2}{\sum_i 1 / \sigma_i^2} \tag{4.47}
\]

\[
\frac{1}{\sigma_A^2} = \sum_i \frac{1}{\sigma_i^2} \tag{4.48}
\]

\[
\tag{4.49}
\]

As important points, we can note that the weight of experiments is proportional to the inverse of the square of their resolution. I.e., you win quickly by improving resolution. Also, as seen by the variance calculation rule, every experiment will bring a reduction in the total variance under our assumption of independent measurements.

4.11.1 Example

We look at a straightforward example to see the formulas derived in the previous section in action. We measure a quantity \( x \), and we model our measurement uncertainty with a Gaussian probability density distribution. In a first experiment, we measure \( x_1 = 0 \) and have a resolution \( \sigma_1 = 2 \). I.e., we assume that our data outcomes have probability density

\[
P(x_1 | \mu, \sigma_1) = \mathcal{G}(x | \mu, \sigma_1).
\]

As a result of this measurement, a Bayesian analysis with flat prior for \( \mu \) would yield

\[
P(\mu | x_1, \sigma_1) = \mathcal{G}(\mu | 0, 2).
\]

This result is shown as the red curve in Fig. 4.12. In a separate experiment of the same quantity, with twice as good resolution, \( \sigma_2 = 1 \), a value \( x_2 = 1 \) was measured. The experimenters started with a flat prior and their result is shown as the blue curve in Fig. 4.12. Now, combining the two results (either analyzing them simultaneously or using the posterior from experiment 1 as prior for the second experiment), the result is

\[
P(\mu | x_1, \sigma_1, x_2, \sigma_2) = \mathcal{G}(\mu | x_A, \sigma_A)
\]

\[
= \mathcal{G}(\mu | 0.8, \sqrt{0.8})
\]

which is shown as the black curve in the figure. As is seen, the final result is much closer to the result with the better resolution.
Figure 4.12: Posterior probability distributions for $\mu$ from an experiment with resolution $\sigma_1 = 2$ finding result $x_1 = 0$, red curve, from an experiment with resolution $\sigma_2 = 1$ finding result $x_2 = 1$, blue curve, and from a combination of both measurements, black curve.
4.12. **ON THE IMPORTANCE OF THE PRIOR**

As we have seen, frequentist analyses result in ranges for parameters where the data fulfill some probability criteria, and the results are dependent on the rule for summing the probabilities of the possible outcomes. Bayesian results are dependent on the choice of prior, and different prior choices will yield different credible intervals for our parameters. Let us see how important the prior choice is in a simple case of Gaussian distributed measurements. We imagine there is a true value for our parameter of interest, called here \(x_0\) with \(x_0 = 0.75\), and that we measure it with a resolution of \(\sigma = 0.1\). Imagine that the possible values of \(x_0\) are form \([0, 1]\), and that two different scientists have very different prior beliefs on the value of \(x_0\). Experimenter 1 is quite confident that \(x_0\) is small and assigns as prior

\[
P_1^0(x_0) = 10e^{-10x_0}
\]

Meanwhile, Experiment 2 is quite confident that \(x_0\) is large and assigns as prior

\[
P_1^1(x_0) = 10e^{-10(1-x_0)}
\]

The first measurement yields \(x = 0.8\). How different are the posterior probability densities? This is seen in Fig. 4.13, where indeed a shift in the posterior distribution is seen. Now imagine that 100 measurements are taken, where we have generated values of \(x_i\) around the true value \(x_0\) with \(\sigma = 1\). The result after these 100 measurements are also shown in the figure. There is now no visible (and very small numerical) difference between the two results. The moral of the story is: if you have strong data (large data sets, precise data), then the prior does not make much difference. On the other hand, if the data is not very constraining, then the posterior pdf can depend in a significant way on the choice of prior. This is as it should be!

---

\(^1\)Note that this is not normalized exactly to 1, but close enough for our purposes
Figure 4.13: Top panels: prior and posterior probability distributions for a quantity $x_0$ for two different priors (shown in red). The results are for a single measurement $x = 0.8$. Lower panels: the prior and posterior pdfs compared after 100 measurements with the $x$ values drawn from a Gaussian pdf with $\mu = 0.75$ and $\sigma = 0.1$. 
4.12. ON THE IMPORTANCE OF THE PRIOR

**Literature** For references - Peele’s pertinent puzzle
Exercises

1. Solve the differential equation (4.8) and find the pre factor by requiring that

\[ \int_{-\infty}^{\infty} \phi(z) dz = 1. \]

A useful integral is

\[ \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \]

2. Prove the first 5 properties of the Gauss distribution in section 4.4 (i.e., mode through FWHM).

3. Prove Equation (4.13)

4. Show that the correlation matrix for a set of independent variables is diagonal.

5. Using the result in (4.17), show that the Normal Distribution (Gauss function with \( \mu = 0 \) and \( \sigma = 1 \)) is an eigenfunction of the Fourier Transformation. Please be careful with the \( \sqrt{2\pi} \) factors.

6. Derive the relations in Eq. (4.22). Some useful facts:

\[
\text{cov}(x, y) = \text{cov}(y, x) \\
\text{cov}(x, x) = \text{var}(x) = \sigma_x^2
\]

If \( X, Y, Z \) are real-valued random variables and \( c \) is a constant, then

\[
\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z) \\
\text{cov}(cX, Y) = c \cdot \text{cov}(X, Y)
\]

For \( (X_1, X_2, \ldots, X_n) \) and \( (Y_1, Y_2, \ldots, Y_m) \) real-valued random variables and \( (a_1, a_2, \ldots, a_n) \) and \( (b_1, b_2, \ldots, b_m) \) constants. Then

\[
\text{cov} \left( \sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{cov}(X_i, Y_j)
\]

7. Try out the CLT in the following. You will need a way to generate random numbers according to different probability distributions.
4.12. ON THE IMPORTANCE OF THE PRIOR

(a) \( P(x) = 1 \quad -0.5 < x < 0.5. \)
Perform 100 experiments where for each experiment you have 10 samples from \( P(x). \)

(b) \( P(x) = 0.001 \quad -500 < x < 500. \)
Perform 100 experiments where for each experiment you have 10 samples from \( P(x). \)

(c) \( P(x) = \frac{a}{x^2} \quad 0.1 < x < \infty. \)
Perform 100 experiments where for each experiment you have 100 samples from \( P(x). \)

8. The resolution of a calorimeter (used to measure energy) is often parametrized with a form
\[
\sigma^2 = \sigma^2_{\text{noise}} + A E_{\text{incident}} + B E_{\text{incident}}^2
\]
where the first term is the noise term, the second is the resolution term and the third is a calibration term. Argue why these different contributions have the energy dependence as given in the formula. You can look at Eq. (4.20) for inspiration.

9. In liquid Argon, 40000 photons in the visible range are produced per MeV of deposited energy. Imagine that you have a detector that has an overall efficiency of 0.001 for detecting these photons. Your detector produces a signal of 1 Volt/detected photon.

(a) What is the calibration constant of the detector (MeV/V) ?
(b) What is the resulting resolution of the detector as a function of incident energy (resolution is defined as the standard deviation measured in energy units, MeV).

10. Cumulative of the Gauss function: make a plot of the cumulative of the Gauss function for \( \mu = -5 \) and \( \sigma = 2. \)

11. With a plotting program, draw contours of the bivariate Gauss function (see Section ??) 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) \( \mu_x = 1, \mu_y = -2, \sigma_x = 1, \sigma_y = 2, \rho_{xy} = -0.7 \)
12. Bivariate Gauss probability distribution

(a) Show that the pdf 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

\[
\begin{align*}
\mu_z &= \mu_x - \mu_y \\
\sigma_z^2 &= \sigma_x^2 + \sigma_y^2 - 2\rho \sigma_x \sigma_y
\end{align*}
\]

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 \( x \).

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

What is the predicted distribution for the observed quantity \( y \)?

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

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>30(^0)</th>
<th>45(^0)</th>
<th>90(^0)</th>
<th>120(^0)</th>
<th>150(^0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross section</td>
<td>11</td>
<td>13</td>
<td>17</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>Error</td>
<td>1.5</td>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
<td>1.5</td>
</tr>
</tbody>
</table>
4.12. ON THE IMPORTANCE OF THE PRIOR

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?

15. Imagine that 2 measurements are made of the same quantity, and that the data are assumed to be distributed according to a Gauss distribution with resolutions $\sigma_1, \sigma_2$. Show that the posterior pdf is the same where both results are analyzed simultaneously or sequentially, with the posterior from the first measurement taken as the prior for the second. You can assume starting with a flat prior.

16. What is the conjugate prior for Gaussian distributed data? Deriving the updating rule for a single measurement $x$ with resolution $\sigma$. 
Chapter 5

Model Fitting and Model selection

In the previous chapters, we have seen how to analyse data to extract parameter values for the underlying statistical distributions (Binomial, Poisson, Gauss). The data were assumed to follow a probability distribution from one of these distributions, and the task was to gain information about the parameters that control those distributions: the success probability in the Binomial case, the expectation value in the Poisson case, and the mean and variance in the Gauss case. We now go one step further and imagine that we are interested in parameters of a function, where the function is used to predict an expectation which is then compared to the data. As an example, we can imagine that we have a function that predicts an even rate as a function of an angle and that the function depends on some parameters; i.e., we have an $f(\theta|\lambda)$ where the parameters are labeled $\lambda$. We make measurements at some known values of $\theta$ and want to extract information on the $\lambda$. We will discuss techniques for gaining information on the parameters. But you may also want to know if your function is a good representation of the obtained data. How would you go about deciding? Or maybe there are alternative descriptions of the your data, say a function $g(\theta|\lambda')$ and you want to decide which of the two functions does a better job in representing your data. What can you do in this case? These will be the topics of this chapter. We will discuss how to approach these issues in both the Frequentist and Bayesian way, and will come across some presumably known quantities such as $\chi^2$ fits, Bayes factors, $p$-values, etc.
CHAPTER 5. MODEL FITTING AND MODEL SELECTION

5.1 Binomial Data

We start by considering a model $f(x|\lambda)$ where $x$ is some measurable quantity such as an angle, an energy, a frequency, etc. and $\lambda$ represents one or more parameters. The function predicts an expectation at a value of $x$ which depends on the parameters $\lambda$. We perform an experiment where for given values of $x$ we have an observed event count. We denote the values of $x$ where we made the measurements $x_i$ and the observed result $r_i$ for a number of trials $N_i$. For each $x_i$, we therefore have for the probability of observing $r_i$ the probability

$$P(r_i|N_i, f(x_i|\lambda)) = \binom{N_i}{r_i} f(x_i|\lambda)^{r_i} (1 - f(x_i|\lambda))^{N_i-r_i}$$ (5.1)

The probability for the full set of observations ($k$ measurements) is then

$$P\left(\{r\}|\{N\}, f(x_i|\lambda)\right) = \prod_{i=1}^{k} \binom{N_i}{r_i} f(x_i|\lambda)^{r_i} (1 - f(x_i|\lambda))^{N_i-r_i}$$ (5.2)

5.1.1 Bayesian Analysis

We start this time with the Bayesian analysis since we do not need to introduce any new concepts. We can write directly the posterior probability density function for the parameter $\lambda$:

$$P(\lambda|\{r\}, \{N\}) = \frac{P(\{r\}|\{N\}, \lambda)P_0(\lambda)}{\int P(\{r\}|\{N\}, \lambda)P_0(\lambda)d\lambda}$$ (5.3)

where we have

$$P(\{r\}|\{N\}, \lambda) = P(\{r\}|\{N\}, \{f(x|\lambda)\}) .$$

I.e., if we specify $\lambda$ and the locations where we measure, $\{x\}$, then we have the prediction for the success probability $f(x|\lambda)$. At this point, we need to look at concrete examples to continue.

5.1.2 Example

Let’s imagine we are studying the efficiency for a device to register a ‘hit’, and we expect that for very small values of $x$, which could be e.g. an energy, we expect the efficiency to be small, whereas for large values of $x$, we expect the efficiency
to approach 1. We need a functional form which can represent this. A commonly
used functional form in this case is a Sigmoid Function:

\[ S(x) = \frac{1}{1 + e^{-x}}. \quad (5.4) \]

The function \( S(x) \) is shown in Fig. 5.1.

If \( x \) should represent an energy, then we should shift the function by replacing
\( x \) with \( E - E_0 \), where \( E_0 \) is the energy where we think our efficiency reaches 50%.
We can also introduce a scaling of the energy since we do not know the units of \( E \) or how quickly the efficiency will rise. We therefore have a two parameter
function for the efficiency:

\[ \epsilon(E|A, E_0) = \frac{1}{1 + e^{-A(E - E_0)}} \quad (5.5) \]

We now analyze the data in Exercise 9 from Chapter 2, which is reproduced here:
To perform a Bayesian Analysis and extract information on the parameters in our function, we need to specify priors. For the offset parameter, we see from the data that our efficiency is about 50% at $E = 2$. Given our functional form, we see that we have $\epsilon(E|E_0, A) = 0.5$ when $E = E_0$. We choose for $E_0$ a Gaussian prior centered on this value, since we have some reasonably good information concerning its best value:

$$P_0(E_0) = \mathcal{G}(E_0|\mu = 2.0, \sigma = 0.3).$$

We can also use the fact that the efficiency changes by about 40% when we move away from $E = 2$ by roughly 0.5 units. Let’s get the slope of the efficiency with the energy:

$$\frac{d\epsilon}{dE} = \frac{Ae^{-A(E-E_0)}}{(1 + e^{-A(E-E_0)})^2} = \frac{A}{4} \quad \text{for} \quad E = E_0.$$

We therefore estimate $A$ with

$$0.5 \cdot A/4 \approx 0.4 \implies A \approx 3.$$

We choose for $A$ a Gaussian prior centered on this value, since we have some reasonably good information concerning its best value$^1$:

$$P_0(A) = \mathcal{G}(A|\mu = 3.0, \sigma = 0.5).$$

---

$^1$You may be wondering if it is ‘legal’ to use the data twice in this way - once to get a prior and then again in the fit. Strictly speaking, the answer is no. However, since we are only interested in getting a good parametrization, then we don’t worry too much about this double use of the data.
The probability of the data given our model is

\[
P(\{r\}|\{N\}, E_0, A) = \prod_{i=1}^{8} \left( \frac{N_i}{r_i} \right) \epsilon(E_i|A, E_0)^{r_i} (1 - \epsilon(E_i|A, E_0))^{N_i-r_i} \tag{5.6}\]

where the $E_i, N_i, r_i$ are the values in the table and $\epsilon$ is given by Eq. (5.5). At this point, we could make a 2-Dimensional grid in $(E_0, A)$ space and evaluate our posterior probability density at each point, but we will instead turn to a numerical algorithm to make progress. We use the Bayesian Analysis Toolkit (BAT), which can be found at the following url: https://www.mpp.mpg.de/bat/ to map out the full posterior probability density function in $E_0, A$. The fitting program gives for the mode of the two-dimensional posterior probability distribution $(A^*, E_0^*) = (3.56, 1.88)$, which is reasonably close to our rough estimates. The resulting sigmoid function is plotted with the data in Fig. 5.2.

Figure 5.2: The efficiency curve resulting from the fit to the data in the table using a sigmoid function. The observed ratio (successes/trials) is shown in the figure with a black marker.

Since we have the full posterior probability density, we can examine in more detail the fit results and extract uncertainties for our parameters. Figure 5.3 gives
a contour plot of the posterior probability density. As we see in the figure, there is some correlation between our two variables (there is a tilt in the contours). Larger values of $E_0$ go together with larger values of $A$. We also get a sense of how well constrained our parameters are. The green shaded area contains 68% of the posterior probability and yields a rather limited range for our parameters. The full information is available in the two-dimensional probability density, and further analysis should use this if possible. However, we might also want to give a simple summary. For this, we usually look at the marginalized distributions, where all parameters except the one of interest are integrated out. In our example, there are two marginalized distributions. These are shown in Figure 5.4

![Probability contours in the $(A, E_0)$ parameter space. The color coding is given in the legend above.](image)

We can quote probable ranges for our parameters from these marginalized probability densities. Using the smallest interval definition, we find that $3.4 < A < 3.76$ and $1.84 < E_0 < 1.94$.

What have we achieved? We have found the optimal parameter values to describe our data assuming a fixed functional form. But is the result acceptable?
What would we consider an acceptable result? Looking by eye, we see that the function with the best-values of the fitted parameters is quite close to the measurements, so this is definitely good. We can try to make a quantitative statement by looking to see if the measured data is in a probable range for the fitted curve. This is a so-called posterior-predictive test. We use the best-fit values from the posterior probability density, i.e., after having used the data to optimize parameter values, and see if the data would have been probable given this outcome. This is by no means a trivial test - if we had a poor functional form, then the probability of our data would be low. Let’s see what happens in our case. We can find the smallest 68% range of outcomes for \( r \) given the model with parameters at the posterior mode, and compare to the actual values of \( r \). The results are given in the table:

We see that 5 of the 8 measurements are in the 68% smallest interval, and the three others are just outside. This is certainly acceptable as a statistical outcome, and we are therefore happy with the results. We cannot expect that any other parametrization would do a better job. It would of course be possible to have all results fall with our probability window if we picked a more complicated functional form, but this would not really be justified since our simple form reproduces the data within the expected fluctuations. We will have much more to say.
5.2 Frequentist Analysis

As we have seen, in the frequentist approach we need to use the probability distribution for possible data under assumed values for our parameters, define sets of outcomes satisfying a probability criteria, and then find values of the parameters where the observed data is a member of the defined set of outcomes. In the Binomial problem, we would therefore first need to define a probability content $1 - \alpha$ and a probability criteria (we have so far used the central, the smallest, and the Feldman-Cousins rankings for summing probabilities to determine which outcomes are included in our sets.) We consider that we have made a series of trials $\{N_i\}$ with possible outcomes $\{r_i; \ r_i = 0, \ldots, N\}$. For given success probabilities $\{p_i(\lambda); \ 0 \leq p_i \leq 1\}$, where $\lambda$ are the parameter(s) of interest, we have to find all sets of $\{r_i\}$ satisfying our probability criteria for our given definition of ranking for which we sum the probabilities. Let’s call these sets of possible results $O_{1-\alpha}^{freq}\{\{r_i\}\}$. We then look at the observed data, and find all values of $\lambda$ for which the observed data is a member of $O_{1-\alpha}^{freq}$; i.e., $\{r_i^{\text{Data}}\} \in O_{1-\alpha}^{freq}$. This yields the $1 - \alpha$ confidence level values of the parameters. The data is usually condensed into a ‘test-statistic’ to carry out this analysis.

5.2.1 Test-statistics

A test-statistic is a scalar function of our data which is used to simplify the analysis. An example for our Binomial situation would be the product of the individual
5.2. FREQUENTIST ANALYSIS

probabilities. Let’s take \( k \) as the number of measurements. Then, a possible test-statistic would be

\[
\xi(\{r_i\}; \lambda) \equiv \prod_{i=1}^{k} \binom{N_i}{r_i} p_i(\lambda)^{r_i}(1 - p_i(\lambda))^{N_i - r_i}.
\]

(5.7)

We can now consider the probability distribution of \( \xi \) for fixed \( \lambda \). Different possible data sets \( \{r_i\} \) will yield different values for \( \xi \). The probability distribution for \( \xi \) is

\[
P(\xi|\lambda) = \sum_{\{r_i\}} P(\{r_i\}) I(\xi(\{r_i\}; \lambda))
\]

(5.8)

where \( I(\cdot) \) is the indicator function and is 1 if the argument is satisfied, else 0.

Once we have the probability distribution for our test statistic, we can then proceed as before using our chosen ranking of outcomes to find the confidence intervals for \( \lambda \).

5.2.2 Example

We consider the same example as above where we want to find values of the parameters \( (A, E_0) \) that fall within our confidence intervals. We will choose \( 1 - \alpha = 0.68, 0.90, 0.95 \) as values and a region (in 2D parameter space) containing this probability by ranking results in order of decreasing \( \xi \). This choice is related to our smallest interval, since \( \xi \) is itself the probability of the data\(^2\).

We need to find the probability distribution for our test-statistic for possible values of \( (A, E_0) \). To this end, we first make a two-dimensional grid in this parameter space as shown in Fig. 5.5, letting \( A \) range from \( 2 \leq A \leq 4 \) in steps of 0.1 and \( E_0 \) range from \( 1 \leq E_0 \leq 3 \) in steps of 0.1. For each point on the grid, rather than calculating the value of our test statistic and its probability for each possible set of results\(^3\) we generate 1000 experiments to simulate the probability distribution of \( \xi \) as follows:

1. We fix the value of \( A \) and \( E_0 \) at one of our grid points, and calculate the success probability for our eight different energies according to Eq. (5.5);

2. For each \( E_i \) where \( i = 1 \ldots 8 \) we randomly generate an \( r_i \) based on the probability distribution \( P(r_i|N_i, p_i = \epsilon(E_i|A, E_0)) \);

\(^2\)Note that we will work with \( \log \xi \) since this is numerically much better. The ranking in \( \log(\xi) \) is identical to the ranking in \( \xi \).

\(^3\)In our example, there are \( \prod_{i=1}^{k} N_i = 10^{19} \) possible \( \{r_i\} \) for each grid point.
Figure 5.5: The gridding used for the frequentist analysis of the efficiency data. The solid points show the values of the parameter \((A, E_0)\) used in the analysis. The analysis is first performed in a coarse grid (top) and then on a finer grid (bottom). The colored points give the values of \((A, E_0)\) where the data is within the 95% most likely values of \(\xi\) (red), the 90% most likely values of \(\xi\) (yellow), and the 68% most likely values of \(\xi\) (green).
3. We then calculate
\[ \xi(\{r_i\}; A, E_0) = \prod_{i=1}^{k} \binom{N_i}{r_i} p_i^{r_i}(1 - p_i)^{N_i - r_i} \]

4. We store the values of \( \xi \) in an array in order of decreasing values of \( \xi \), and note the value of \( \xi \) for which 68, 90, 95 % of experiments are above this value.

5. We check whether \( \xi^{\text{Data}} \) is included in the accepted range; if it is, then the values of \( A, E_0 \) are in our \( 1 - \alpha \) % confidence level interval.

To make this more concrete, we look at some of the distributions we have produced. Figure 5.6 shows the distributions of \( r \) for our eight different energy points for the values \( A = 3 \) and \( E_0 = 1.6 \). The expectation values for \( r \), \( E[r] = Np \), are given in the plots. The distribution of \( \log(\xi) \) for \( A = 3.5 \) and \( E_0 = 1.8 \) (which is in the 68 % CL region for the parameters) is shown in Fig. 5.7. The values of \( \log(\xi) \) which include (68, 90, 95) % of the experiments are indicated in color in the plot, as well as the value of \( \xi^{\text{Data}} \). Since \( \xi^{\text{Data}} \) is in the included set of values, the value of \( (A, E_0) \) are then in our confidence interval for the parameters. The region in \( (A, E_0) \) space that is included in our (68, 90, and 95) % confidence intervals are shown in Fig. 5.5.

Now that we have a good idea for which values of \( (A, E_0) \) we have a high confidence value, we make a finer grid and generate more experiments on the computer. This allows us to produce a finer resolution on the accepted values of the parameters. The result for the finer gridding is also shown in Fig. 5.5.

### 5.3 Goodness-of-fit testing

In the previous sections, we have seen how we can extract information about the parameters in our model. What if we wanted to judge whether the model we selected was appropriate to represent the data - how would we do that? Here we note a significant difference between the Bayesian extraction of credible regions for our model parameters and the Frequentist confidence levels for the parameters. In the Bayesian approach, the credible regions are the one where the parameters are most likely assuming the model is correct. I.e., we only compare more and less favored regions of the parameter space assuming the correct model. This procedure will produce credible regions whether or not the model is a good predictor.
Figure 5.6: The distributions of $r$ for the 8 different energy settings for the parameter values $A = 3.0$ and $E_0 = 1.6$. The expectation value is given for each of the distributions. The energies increase from top left (0.5) to bottom right (4.0).

of the data or not. In the frequentist case, we also assume the model is correct, but then we ask whether the observed data is within the expected outcomes in the model. If the model is a poor representation of the data, then our data will not be within our $1 - \alpha$ sets for any of the possible values of the model parameters, and we could end up with empty confidence intervals. This is a generalization of what we have seen in previous chapters where we found that it was possible to produce empty sets if the data were unlikely (in those cases, even though we had the correct model). We make these points on the same example data considered in the previous sections, but now using a parametrization that does not represent
Figure 5.7: The distribution of $\log(\xi)$ for $A = 3.5$ and $E_0 = 1.8$. The (green, yellow, red) regions show the values that would be included in the top $(60, 90, 95)\%$ CL intervals. The value observed for the data for these parameter values is shown by the arrow.

the data well. We will then come back to our discussion of how to judge model adequacy.

### 5.3.1 Bayes Analysis - poor model

Instead of the sigmoid function chosen above, we consider the following function

$$
\epsilon(E) = \begin{cases} 
1 - e^{-\lambda(E - E_1)} & E \geq E_1 \\
0 & \text{otherwise}
\end{cases}
$$

(5.9)

(5.10)

Since $r_{\text{Data}} > 0$ in our data for $E = 1$, we know that $E_1 \leq 1$. On the other hand, $r_{\text{Data}} = 0$ at $E = 0.5$ so the efficiency there must be very small. We choose as prior for $E_1$

$$
P_0(E_1) = \begin{cases} 
2 & 0.5 \leq E_1 \leq 1 \\
0 & \text{else}
\end{cases}
$$

For $\lambda$, we see from the data that the efficiency rises to values close to 1 over an $E$ range of 2 unit. We therefore anticipate that $\lambda$ will be a value close to 2.
We choose as prior a Gauss probability density function $P_0(\lambda) = G(\lambda|1.5, 1.0)$. The results of the fit (using BAT) are shown in Fig. 5.8. As we see, we can find credible regions for the parameters as expected. However, when we compare the resulting parametrization, setting the parameter values at the mode of the posterior probability density ($E_1 = 0.985, \lambda = 1.74$), we see that the efficiency curve badly misses the data.

![Figure 5.8: Left: The results from BAT for the credible regions for the parameters $E_1, \lambda$ for the parametrization given in Eq. (5.9). Right: the efficiency curve with the parameter values at the mode of the posterior probability density compared to the data.](image)

**5.3.2 Frequentist Analysis - poor model**

We now redo the frequentist parameter analysis for the poor model. We define a grid in the parameter region $(0.5 < E_1 < 1, 1 < \lambda < 2)$ and again generate many pseudo experiments for each of the grid points. We use the same test statistic as before but use the efficiency as defined in Eq. (5.9). This time, we find that for no values of the parameters is our data within the expected sets, even at $1 - \alpha = 0.95$. The expected distribution of $\log(\xi)$ for the best fit parameters is shown in Fig. 5.9. For these parameters, we find $\log(\xi^{\text{Data}}) = -209$. This tells us that the data result is extremely unlikely given our assumed model and parameters.
5.4. P-VALUES

Figure 5.9: The distribution of $\log(\xi)$ for $\lambda = 1.74$ and $E_1 = 0.985$. The (green, yellow, red) regions show the values that would be included in the top (60, 90, 95) % CL intervals. The value observed for the data for these parameter values is $-299$.

5.4 p-values

From the previous example, we glean that using the probability of the data given the model and best fit parameters can be used to determine whether our model yields an adequate explanation of the data. We quantify this a bit by introducing $p$-values - 'tail-area' probabilities that a result as extreme or more extreme than the one seen in the data could have occurred under the assumption that our model was correct. We will spend the next few pages defining precisely what is meant by a $p$-value, and examples of its use. Since $p$-values are highly controversial\footnote{Their use has actually been banned by some journals since authors used them to draw incorrect conclusions.} we will spend some time on this topic, understanding what they are and what they are not, and give examples where they are useful and where they are not useful.
5.4.1 Definition of \( p \)-values

We assume a test statistic has been defined for our data, and will use the symbol \( \xi \) to represent its value. In the simplest case, it can be just an observed number of events. A more complicated example for a Binomial analysis problem was given above. We can evaluate the test statistic for our current data set, resulting in \( \xi_{\text{Data}} \).

Now imagine that we have a model that should ‘explain’ our data, and that this model depends on some parameters \( \lambda \). For every choice of the parameters, we need to find a frequency distribution for \( \xi \). In the simplest cases, we can do this analytically, but more generally this needs to be done simulating the experiment on the computer (as was done above). Let us suppose we have generated such a frequency distribution from computer simulations. We get the probability distribution for \( \xi \) by normalizing the frequency distribution. For example, we would take the frequency distribution shown in Fig. 5.7 and divide by 10000 since that was number of experiments used to generate this distribution. This gives us a binned estimate of the probability density for \( \xi \) for the chosen parameter values.

The \( p \)-value for a given experiment (simulated or the actual data) is the value of the cumulative probability density for the test statistic evaluated at the value seen in the experiment, \( \xi_{\text{exp}} \):

\[
p = F(\xi_{\text{exp}}) = \int_{\xi_{\text{min}}}^{\xi_{\text{exp}}} P(\xi|\lambda) d\xi \quad \text{or} \quad p = 1 - F(\xi_{\text{exp}}) = \int_{\xi_{\text{exp}}}^{\xi_{\text{max}}} P(\xi|\lambda) d\xi \quad (5.11)
\]

where \( P(\xi|\lambda) \) is the probability density for \( \xi \) for the fixed parameter values \( \lambda \). This is shown graphically in Fig. 5.10, where the probability density for the test statistic is shown. The shaded area shows the region in probability density that is integrated to get the \( p \)-value. Since the \( p \)-value is based on tails areas of the probability density, it is a number between \([0, 1]\). We now get the probability density for \( p \), assuming our model is correct.

5.4.2 Probability density for \( p \)

We note first that the \( p \)-value can be viewed as a random variable. Different values of our test statistic will yield different \( p \)-values. In order to make a judgment based on a \( p \) value, we need to know its probability density. Here is the result:

\[
P(p|\lambda) \frac{dp}{d\xi} = P(\xi|\lambda) \quad (5.12)
\]
5.4. P-VALUES

why do we reject the model for small p-values if all are equally likely?

Figure 5.10: Graphical explanation of a p-value. The test statistic is denoted here with $T$, and the left plot shows the probability density for the test statistic under the assumed model. The $p$-value is the tail-area probability for a given measurement - in this case $T = 30.5$. The right plot shows the expected distribution for $p$ if the model is correct. An example value of $1 - \alpha = 0.05$ is shown corresponding to accepting only values with $p < 0.05$.

with

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

so that

$$P(p|\lambda) = 1.$$ 

I.e., the expected $p$-value distribution is flat between $[0.1]$ and is shown in Fig. ??.

All values of $p$ are equally likely! So, how can we use this to decide whether our model is an adequate description of the data? Any single value of $p$ has the same probability density as any other value. For the correct model, large values are not favored over small values, as is often wrongly assumed. We can make an arbitrary decision and say we will only keep models or parametrizations of our data for which $p > 0.05$, say, but this will have us reject the correct model 5% of the time, exactly the same as if we said we would only accept $p$ values below 0.95. The reason $p$-values are useful in practice is rather based on what we expect for $p$-value distributions if the model or parametrization is not the correct one to describe the data. This argument follows Bayesian logic.
5.4.3 Bayesian argument for \( p \)-values

In the example we considered above, we have seen that we found a very small value for our test statistic when choosing a parametrization that did not follow the data. The \( p \)-value\(^5\) is therefore very close to zero. This will generally be true - if our chosen model or parametrization does poorly in representing the data, then the data will yield a very small \( p \)-value.

The use of \( p \)-value for evaluation of models is essentially Bayesian in character. Following the arguments given above, the \( p \)-value probability density for a good model is flat (for the following, we assume the data is generated according to \( M_0 \), so that \( M_0 \) represents a good model),

\[
P(p|M_0) = 1
\]

and that for a poor model can be represented by

\[
P(p|M_i) \sim \lambda_i e^{-\lambda_i p}
\]

where \( \lambda_i \gg 1 \) so that the distribution is strongly peaked at 0 and approximately normalized to 1. The degree-of-belief assigned to model \( M \) after finding a \( p \)-value \( p \) is then

\[
P(M|p) = \frac{P(p|M)P_0(M)}{\sum_{i=0}^{N} P(p|M_i)P_0(M_i)} .
\]

(5.13)

We now want to see what happens to our degree-of-belief that we have chosen a correct model; i.e., that \( M = M_0 \), after analyzing the data and finding a value of \( p \). If we take all models to have similar prior degrees-of-belief, then

\[
P(M = M_0|p) \approx \frac{P(p|M_0)}{\sum_{i=0}^{N} P(p|M_i)} .
\]

If we have \( M \neq M_0 \), then the limit \( p \to 0 \) is assumed to apply, and we have

\[
P(M = M_0|p) \approx \frac{1}{1 + \sum_{i=1}^{N} \lambda_i} \ll 1 .
\]

while if \( M = M_0 \), then often \( p \gg 1/\min\{\lambda_i\} \),

\[
P(M = M_0|p) \approx 1 .
\]

\(^5\)Assuming we could calculate it - here we can only say it is extremely small because we did not find any values of our test statistic in our numerical simulations that were that small.
Therefore, seeing a large \( p \) value means that we have likely picked a model for which the \( p \)-values are reasonably flat; i.e., a reasonable model.

We should note a couple points here:

- We assigned similar priors to the different possible models to draw our conclusion. If we have strong belief that our model \( M = M_0 \), then this should be taken into account;

- Since the argumentation relies on the expected \( p \)-value distribution for unknown models, it is necessarily very vague and not intended to be quantitative.

Using \( p \)-values for goodness-of-fit determination is therefore highly subjective. If we cannot enumerate a full set of models and the probability distribution of the data in these models, then the only quantitative statement that can be made is how often we will reject the correct model with our \( p \)-value cut.

### 5.5 The \( \chi^2 \) test statistic

The best known choice of a test statistic is the weighted sum of the squared distances between the model prediction and the data, called \( \chi^2 \):

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

(5.14)

Here \( f(x|\lambda) \) is our model predicting a value for the data at \( x \) given parameter values \( \lambda \), and the \( y \) are the measured values. In the efficiency example discussed in detail above, \( x \rightarrow E \) and \( f(x|\lambda) \rightarrow N \epsilon(E|A, E_0) \), where \( N \) is the number of trials at energy \( E \). Then \( y_i \rightarrow r_i \), the number of events measured at \( E_i \). We can choose different values of the weights \( 1/w_i^2 \). A common choice would be to use the expected variance of the measurements:

\[
w_i^2 = N_i \epsilon_i (1 - \epsilon_i).
\]

We could use this test statistic and redo the frequentist analysis described above by generating probability distributions for \( \chi^2 \) for different values of the parameters and find the regions in parameter space where \( \chi^2, \text{Data} \) is within our \( 1 - \alpha \) allowed interval. We leave this for the exercises.

Let’s consider what happens if we can assume our data follow a Gaussian probability density function. In this case, the probability distribution for \( \chi^2 \) can be worked out and has a simple behavior.
5.5.1 $\chi^2$ for Gauss distributed data

We assume we have a functional form $f(x|\lambda)$ that predicts data values at measuring points $x$, and that the measurements will follow a Gauss probability density function:

$$P(y) = G(y|f(x|\lambda), \sigma(x|\lambda)) .$$  \hfill (5.15)

Let’s work out the probability distribution for $\chi^2$ for a single measurement:

$$P(\chi^2)\frac{d\chi^2}{dy} = 2P(y) \quad y \geq f(x|\lambda)$$

$$\frac{d\chi^2}{dy} = \frac{2(y - f(x|\lambda))}{\sigma^2}$$

$$\left|\frac{d\chi^2}{dy}\right| = \frac{2\sqrt{\chi^2}}{\sigma}$$

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

$$P(\chi^2) = \frac{1}{\sqrt{2\pi}\chi^2} e^{-\chi^2/2}$$ \hfill (5.16)

We find that the probability distribution for $\chi^2$ does not depend on any function parameters - it is a universal function. Here are some properties of this distribution:

1. The expectation value $E[\chi^2] = 1$;

2. The variance is $Var[\chi^2] = 2$;

3. The mode is $\chi^2^* = 0$;

If we have multiple data points, each following a Gaussian pdf, we just sample many times from the $\chi^2$ pdf. Let’s derive the probability density function for $n$ data points. Each data point will have the same distribution. Let’s start with two

---

6 The factor 2 in the 1st line of the derivation is there because both positive and negative values of $(y - f(x|\lambda)$ can contribute to the same $\chi^2$ (which can only be positive).
data points: we want \( P(Z) \) for \( Z = X_1 + X_2 \) and the \( X \)'s are \( \chi^2 \) distributed.

\[
P(Z) = \int_0^Z P(X_2) P(X_1 = Z - X_2) dX_2
\]

\[
= \int_0^Z \frac{1}{\sqrt{2\pi X_2}} e^{-X_2/2} \frac{1}{\sqrt{2\pi(Z - X_2)}} e^{-(Z-X_2)/2} dX_2
\]

\[
= \frac{1}{2\pi} e^{-Z/2} \int_0^Z \frac{1}{X_2(Z - X_2)} dX_2
\]

\[
= \frac{1}{2\pi} e^{-Z/2} \left[ 2\tan^{-1}\left( \sqrt{\frac{X_2}{Z - X_2}} \right) \right]_0^Z
\]

\[
= \frac{1}{2} e^{-Z/2}.
\]

We have \( P(\chi^2|n = 2) = \frac{1}{2} e^{-\chi^2/2} \). Adding the third measurement is straightforward, and yields

\[
P(\chi^2|n = 3) = \frac{\sqrt{\chi^2}}{\sqrt{2\pi}} e^{-\chi^2/2}
\]

Repeating the integration, we end up with the simple result:

\[
P(\chi^2|n) = \frac{(\chi^2)^{n/2-1}}{2^{n/2}\Gamma(n/2)} e^{-\chi^2/2} \tag{5.17}
\]

where the Gamma function is simply

\[
\Gamma(1/2) = \sqrt{\pi}
\]

\[
\Gamma(1) = 1
\]

\[
\Gamma(r + 1) = r \cdot \Gamma(r) \quad r \in \mathbb{R}^+
\]

\[
\Gamma(n) = (n - 1)! \quad n \in \mathbb{Z}
\]

\[
\Gamma(n + 1/2) = \frac{(2n)!}{4^n n!} \sqrt{\pi} \quad n \in \mathbb{Z}
\]
5.5.2 Alternate derivation

With the definition
\[ q_i = \frac{y_i - f(x_i|\lambda)}{\sigma_i} \]
we can write
\[ \chi^2 = \sum_{i=1}^{n} q_i^2 . \]

I.e., in using the shifted and scaled values of the data, we see that \( \chi^2 \) can be viewed as a squared radius in the \( n \)-dimensional space of the \( q_i \). All choices of \( q_i \) resulting in the same radius with have the same \( \chi^2 \). To find the probability density \( P(\chi^2) \), we evaluate the probability in a thin shell of radius \( d\chi^2 \): Given that our measurements are independent, we have

\[ P(q_1, q_2, \ldots, q_n) = \left( \frac{1}{2\pi} \right)^{n/2} e^{-\sum_{i=1}^{n} q_i^2/2} = \left( \frac{1}{2\pi} \right)^{n/2} e^{-\chi^2/2} . \]

The probability in a thin shell \( d\chi^2 \) is then

\[ P(\chi^2) d\chi^2 = \int_{\text{shell}} d\chi^2 \left( \frac{1}{2\pi} \right)^{n/2} e^{-\chi^2/2} dq_1 dq_2 \ldots dq_n \]
\[ = \left( \frac{1}{2\pi} \right)^{n/2} e^{-\chi^2/2} \int_{\text{shell}} d\chi^2 dq_1 dq_2 \ldots dq_n \]
\[ = \left( \frac{1}{2\pi} \right)^{n/2} e^{-\chi^2/2} A\chi^{N-1} d\chi \]
\[ = C e^{-\chi^2/2}\chi^{N-2} d\chi \]

and we find the constant \( C \) from normalization and find the same result as reported in Eq. 5.17. This version of the derivation will be useful for later considerations.

5.5.3 Use of the \( \chi^2 \) distribution

The \( \chi^2 \) probability distribution for Gauss distributed data is shown in Fig. 5.11 for several values of \( n \), the number of data points. We will call this distribution the ‘canonical’ \( \chi^2 \) distribution. The reason to give it this name is that many \( \chi^2 \) test-statistics have been defined in the literature and applied to non-Gauss distributed data. The \( \chi^2 \) probability distributions for those quantities will differ from the
5.5. THE $\chi^2$ TEST STATISTIC

canonical distribution (although will usually approach it asymptotically as $n \to \infty$).

We note that we expect the $\chi^2$ probability distribution to approach a Gaussian distribution for large enough $n$ from the Central Limit Theorem, since we have repeated sampling from the same distribution and the distribution has finite moments. Since the variance of the canonical $\chi^2$ distribution for one measurement is 2, the variance for $n$ measurements is $2n$. We compare the $\chi^2$ distribution with a Gauss distribution of variance $2n$ in Fig. 5.11 for several cases and see that it is indeed a good approximation for $n$ large enough (say $\geq 50$). This is quite useful since it gives us an indication of what a ‘reasonable’ value of $\chi^2$ is, where reasonable means here in the central part of the expected distribution. E.g., for $n$ data points, we expect $\chi^2$ to be in the range $n \pm \sqrt{2n}$ 68% of the time when $n$ is large.

As for any test-statistic, we can define the $p$-value associated with $\chi^2$ by taking the cumulative of the probability distribution. Since larger values of $\chi^2$ means a larger deviation of the data from the expectation, the $p$-value is defined as

$$
p_{\chi^2,D} = \int_{\chi^2,D}^{\infty} P(\chi^2|n) d\chi^2
= 1 - F(\chi^2,D|n)
= 1 - \frac{\gamma\left(\frac{n}{2}, \frac{\chi^2,D}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.
$$

(5.18)

(5.19)

(5.20)

where $\gamma(s,t)$ is the lower-incomplete gamma function defined as

$$
\gamma(s,t) = \sum_{i=0}^{\infty} \frac{t^i e^{-t} s(s+1)\ldots(s+i)}{s(s+1)\ldots(s+i)}
$$

So the $p$-value gives the probability that $\chi^2$ could have been larger than the value observed in our data $\chi^2,D$ in a large number of experiments. Again, the probability distribution for $p$ is flat between $[0, 1]$ for the correct model, but as discussed above, we believe that for the wrong model, typically small values of $p$ will result so that we are suspicious when we find a small $p$-value.

**Simple Example**

Imagine that we are trying to determine the value of a quantity, $\mu$, and that we perform several measurements where we assume that the measurements are Gaussian
Figure 5.11: Probability distribution for the canonical $\chi^2$ distribution (Gauss distributed data) for different numbers of data points is shown as the black solid curve. $n = 1$ upper left; $n = 10$ upper right; $n = 20$ lower left; $n = 50$ lower right. Shown in red is a Gauss probability distribution function with the mean $\mu = n$ and variance $\sigma^2 = 2n$.

distributed about the true value but with possibly varying resolutions $\sigma_i$ (which are however kept fixed: i.e., they do not depend on $\mu$). This is the same case we treated in Sec. 4.11. We have

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma_i^2}$$
where we have \( n \) measurement values \( y_i \). For any given value of \( \mu \), we can calculate \( \chi^2 \) and find the probability of this \( \chi^2 \). Suppose the task is to find an estimate for \( \mu \). We can use the value of \( \mu \) that maximizes the probability of the data (maximum likelihood estimator):

\[
P(\{y\} | \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{(y_i - \mu)^2}{2\sigma_i^2}} \tag{5.21}
\]

\[
= \left[ \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} \right] e^{-1/2\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma_i^2}} \tag{5.22}
\]

\[
= \left[ \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} \right] e^{-\chi^2/2} \tag{5.23}
\]

so we see that maximizing the probability of the data means minimizing the value of \( \chi^2 \). Since the probability of the data viewed as a function of the parameters is the likelihood, minimizing the \( \chi^2 \) for this simple example corresponds to maximizing the likelihood. And, if we have flat priors for the parameters, then this is the same as maximizing the posterior probability density. Let us see what result we get with this procedure. We need to find the minimum value of \( \chi^2 \), so we set the derivative relative to \( \mu \) to zero:

\[
\left. \frac{d\chi^2}{d\mu} \right|_{\hat{\mu}} = 0 \implies \hat{\mu} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} = \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}
\]

This leads to

\[
\hat{\mu} = \frac{\sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}
\]

which is exactly the same result that we found in the Bayesian analysis of the same case. The results are the same because we assumed a flat prior in the Bayesian analysis and minimizing \( \chi^2 \) in this case is the same as maximizing the posterior probability density.

There are different approaches that we can consider to produce an uncertainty on our value of \( \hat{\mu} \). In the frequentist approach, we would find the values of \( \mu \) for which our test statistic is within an interval containing \( 1 - \alpha \) probability. We discuss this further on since new concepts need to be introduced to carry this out.

\[ ^7 \text{beware - this is not in general true.} \]
In the Bayesian approach, assuming a flat prior on \( \mu \), we have

\[
P(\mu|D) = \frac{1}{\sqrt{2\pi}\sigma_{\mu}} e^{-\frac{(\mu-\hat{\mu})^2}{2\sigma_{\mu}^2}} \propto e^{-\chi^2(\mu)/2}
\]

with \( \frac{1}{\sigma_{\mu}^2} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \). For 68 \% probability, we have the interval

\[
\mu \in [\hat{\mu} - \sigma_\mu, \hat{\mu} + \sigma_\mu].
\]

Which corresponds to:

\[
\chi^2 \in [\chi^2_{\text{min}}, \chi^2_{\text{min}} + 1].
\]

In general, we have

\[
\chi^2 \in [\chi^2_{\text{min}}, \chi^2_{\text{min}} + N^2].
\]

where \( N \) represents the number of standard deviations desired for the interval \([\hat{\mu} - N\sigma_\mu, \hat{\mu} + N\sigma_\mu]\).

In summary, we see that performing a \( \chi^2 \) minimization for this special case corresponds exactly to a Bayesian analysis with a flat prior. To carry out the frequentist analysis using \( \chi^2 \) as our test statistic, we need to find the probability distribution for \( \chi^2 \) in this setting. This will not be the same as that found in Eq. 5.17 because we now allow parameter values to change in order to optimize the resulting value of \( \chi^2 \).

### 5.5.4 General case for linear dependence on parameters and fixed Gaussian uncertainties

We assume that our prediction depends linearly on the parameters

\[
f(x|\vec{\lambda}) = \sum_{k=1}^{K} \lambda_k g_k(x)
\]

where we have \( K \) parameters and the functions \( g_k(x) \) do not depend on the parameters. We also assume that the \( \sigma_i \) do not depend on the parameters\(^8\). We measure

---

\(^8\)We note that this is a very strong assumption, usually not met in practice. E.g., if we are dealing with event counts, then the expected value would be \( \nu(\vec{\lambda}) \) and \( \sigma^2 = \nu(\vec{x}) \). You should critically investigate whether your data uncertainties are indeed independent of the parameters.
5.5. THE $\chi^2$ TEST STATISTIC

a set of values \{y_i\} at locations \{x_i\}. As discussed in the previous chapter (actually, not yet included in the script, so take this on faith for now), the resulting probability distribution will be a Gaussian probability distribution of the form

$$P(\vec{\lambda}|D) \propto e^{-\chi^2(\vec{\lambda})}.$$  

To find the maximum of the posterior probability distribution, we take derivatives wrt the parameters and set these to zero:

$$\left. \frac{\partial \chi^2}{\partial \lambda_k} \right|_{\hat{\lambda}_k} = -\sum_{i=1}^{n} \frac{2(y_i - f(x_i|\vec{\lambda})))}{\sigma_i^2} \frac{\partial f(x_i|\vec{\lambda})}{\partial \lambda_k} \bigg|_{\hat{\lambda}_k} = 0 \ \forall k .$$

There is typically no analytic solution to this set of equations and a numerical evaluation is necessary. However, we can deal with some simple cases such as the one considered here. (Note: if the $\sigma_i$ depend on the parameters, then the expression above is not correct. We assume here this is not the case). For our special case

$$\left. \frac{\partial f(x_i|\vec{\lambda})}{\partial \lambda_k} \right|_{\hat{\lambda}_k} = C_{ik} = g_k(x_i)$$

is just a number depending on the value of $x_i$. We can rewrite our set of equations as follows

$$\sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} C_{ik} = \sum_{i=1}^{n} \frac{f(x_i|\vec{\lambda})}{\sigma_i^2} C_{ik} = \sum_{i=1}^{n} \sum_{l=1}^{K} \frac{\lambda_l g_l(x_i)}{\sigma_i^2} C_{ik} = \sum_{i=1}^{n} \sum_{l=1}^{K} \frac{\lambda_l C_{il}}{\sigma_i^2} C_{ik}$$

which can be rewritten as a matrix equation

$$\vec{Y} = M \vec{\lambda}$$

with

$$Y_k = \sum_{i=1}^{n} \frac{C_{ik}}{\sigma_i^2} y_i$$
\[ M_{kl} = \sum_{i=1}^{n} \frac{C_{ik}C_{il}}{\sigma_{i}^2}. \]

So, we can find the values of the parameters that give the maximum posterior probability density via

\[ \hat{\lambda} = M^{-1} \bar{Y}. \]

Matrix inversion is often very difficult in practice and the \( \chi^2 \) minimization usually has to be done numerically. We give this result for completeness since it has historically played an important role in data analysis. However, given the advent of powerful computing, there is no need to use these expressions which are often a crude approximation.

We have spent many pages discussing the \( \chi^2 \) distribution in great detail because of its historical (and continued) importance. A great factor in its continued use is 1) widespread familiarity with this distribution and 2) the fact that the Gaussian probability distribution function is almost universally used. It is therefore important that you know its origin, under what conditions it applies, but also its limitations. It should not be used in place of a full-fledged frequentist or Bayesian analysis unless you are convinced that the conditions under which it applies are fulfilled.

**Numerical example**

We give here a simple example to show the use of the formulas derived above. We consider a quadratic function to be fit to some data:

\[ f(x|a_0, a_1, a_2) = a_0 + a_1 \cdot x + a_2 \cdot x^2 \]

so that \( C_{i1} = 1, C_{i2} = x_i, C_{i3} = x_i^2 \). Imagine that we have the following data

<table>
<thead>
<tr>
<th>( i )</th>
<th>( x )</th>
<th>( y )</th>
<th>( \sigma_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>1.4</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>3.8</td>
<td>1.0</td>
</tr>
<tr>
<td>4</td>
<td>4.0</td>
<td>5.2</td>
<td>1.0</td>
</tr>
</tbody>
</table>
5.5. THE $\chi^2$ TEST STATISTIC

We find $Y_1 = 11.4, Y_2 = 36, Y_3 = 124$ and for the matrix $M$

$$M = \begin{bmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{bmatrix}$$

For the inverse, we have

$$M^{-1} = \begin{bmatrix} 7.75 & -6.75 & 1.25 \\ -6.75 & 6.45 & -1.25 \\ 1.25 & -1.25 & 0.25 \end{bmatrix}$$

which then yields for our parameters: $a_0 = 0.35, a_1 = 0.25, a_2 = 0.25$. The results are shown graphically in Fig. 5.12. We discuss how to assess uncertainties next.

Figure 5.12: The data from the numerical example in the text and the best fit function from the $\chi^2$ minimization.
5.5.5 Parameter uncertainties for the linear model

The probability of the data (likelihood) assuming our Gauss distributed data is

\[
P(D|\vec{\lambda}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left( -\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - f(x_i|\vec{\lambda}))^2}{\sigma_i^2} \right).
\]

Taking flat priors for the parameters, this will lead to a multivariate Gauss probability distribution function for the parameters:

\[
P(\vec{\lambda}|D) = \frac{1}{(2\pi)^{K/2} |\Sigma|} \exp \left( -\frac{1}{2}(\vec{\lambda} - \hat{\vec{\lambda}})^t \Sigma^{-1}(\vec{\lambda} - \hat{\vec{\lambda}}) \right)
\]

with \(\Sigma\) the covariance matrix of the parameters and \(|\Sigma|\) the determinant of the matrix and \(\hat{\vec{\lambda}}\) the expectation values for the parameters. We see that we can write

\[(\Sigma^{-1})_{ij} = -\frac{\partial^2 \ln P(\vec{\lambda}|D)}{\partial \lambda_i \partial \lambda_j}\]

or in terms of \(\chi^2\)

\[(\Sigma^{-1})_{ij} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial \lambda_i \partial \lambda_j} = M_{ij}\]

where we have again assumed that the \(\sigma_i^2\) do not depend on the parameters and that we have a linear dependence on the parameters.

Applying this to the numerical example above, we have:

\[
\sigma_{a0} = 2.78; \sigma_{a1} = 2.54; \sigma_{a2} = 0.50
\]

and we also find some correlations amongst the parameters. As we see, the uncertainties on the parameters are very large when compared to their values. We note that it is much fast in practice to solve this type of problem by coding the probability of the data into a program such as BAT and solving for the posterior pdf numerically. The result from BAT for this example is shown in Fig. 5.13. The numerical results are the same as those found analytically to high precision, and in addition we have graphical output that allows for quick comprehension of the results.
5.5. THE $\chi^2$ TEST STATISTIC

Figure 5.13: The results from BAT for the quadratic polynomial fit to the data in the text. One and two-dimensional marginalized probability distributions are shown.

5.5.6 $\chi^2$ distribution after fitting parameters

In deriving Eq. 5.17, we assumed that the parameter values were fixed, and we evaluated the probability distribution that we expected for $\chi^2$ assuming we have
measurements assumed to follow Gaussian distributions around the expected value from our fit function. What happens if we modify the parameters in order to get 'best fit' values - what distribution do we expect from $\chi^2$? This distribution is needed if we want to perform a frequentist analysis of the results. As noted above, the distribution is not needed for the Bayesian analysis. An analytic form can be derived under the following special conditions:

- the weights $1/w_i^2$ in the $\chi^2$ definition are not dependent on the parameters;
- the functional form for our model depends linearly on the parameters.

Writing for this special case

$$f(x_i | \lambda) = \sum_{k=1}^{K} a_k \lambda_k$$

where $K$ is the number of parameters, we have

$$\chi^2 = \sum_i \left( \frac{y_i - \sum_{k=1}^{K} a_k \lambda_k}{w_i^2} \right)^2.$$  \hspace{1cm} (5.25)

Minimizing $\chi^2$ will yield $K$ constraint equations which will translate into a reduction of the dimensionality of our integral by $K$. We introduce these constraint equations in our derivation of the probability distribution for $\chi^2$ schematically via Dirac delta functions:

$$P(\chi^2)d\chi^2 \propto \int_{\text{shell}} e^{-\chi^2/2} \delta(1) \ldots \delta(K) dq_1 dq_2 \ldots dq_n$$

$$\propto e^{-\chi^2/2} \int_{\text{shell}} dq_1 dq_2 \ldots dq_{n-K}$$

$$\propto e^{-\chi^2/2} \chi^{n-K-2} d\chi^2$$

so that the $\chi^2$ probability distribution function is that for $n - K$ 'degrees-of-freedom.' We remind again that this derivation is valid under the special conditions mentioned above (fixed uncertainties and linear dependence on the parameters).
5.5. THE $\chi^2$ TEST STATISTIC

5.5.7 $\chi^2$ for Poisson distributed data

The fact that we have a known probability distribution for $\chi^2$ for Gauss distributed data makes its use very appealing since the computationally intensive step of finding the probability distribution for our test statistic is avoided. The question is then how good an approximation the canonical $\chi^2$ distribution is for the test statistic at hand. The usual answer is that for an asymptotically large expectation, distributions become Gaussian and the canonical distribution is warranted. The following $\chi^2$ definitions have been defined for use in event counting settings (Poisson distribution of event counts).

**Neyman $\chi^2$**

The Neyman $\chi^2$ definition is

$$\chi^2_{\text{Neyman}} = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{n_i}$$

(5.26)

where $\nu_i$ is the expectation for measurement $i$ and $n_i$ is the observed number of events. $N$ is the total number of measurements. This definition is clearly problematic for small $n_i$; e.g., for $n_i = 0$ it is not defined. In the limit of large $\nu_i$, the Poisson distribution approaches a Gauss distribution and the Neyman $\chi^2$ approaches the $\chi^2$ probability distribution for Gauss distributed data. For a Poisson distribution in the limit of large $\nu$, we have that

$$\frac{|\nu - n|}{\nu} \sim \frac{1}{\sqrt{\nu}} \to 0$$

so that we can use $n$ as a good approximation for $\nu$, and furthermore $\text{Var}(n) = \nu$ so that $\sigma^2 = \nu \approx n$. So the Neyman $\chi^2$ distribution approaches the Gaussian $\chi^2$ distribution for asymptotically large values of $\nu$.

**Pearson $\chi^2$**

The Pearson $\chi^2$ definition is

$$\chi^2_{\text{Pearson}} = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

(5.27)
where $\nu_i$ is the expectation for measurement $i$ and $n_i$ is the observed number of events. $N$ is the total number of measurements. Since $\nu_i \neq 0$ in practical applications, this definition does not have the problematic behavior for $\nu = 0$. However, the probability distribution of $\chi^2_{\text{Pearson}}$ only approaches the canonical $\chi^2$ distribution in the limit $\nu \to \infty$.

A comparison of the Pearson, Neyman and canonical $\chi^2$ distributions is left for the exercises. We just point out here that using the canonical $\chi^2$ distribution as an approximation for the probability distribution of either the Pearson or Neyman $\chi^2$ is hardly justified today when the required numerical calculations can be performed on ordinary computers.

### 5.5.8 General $\chi^2$ Analysis for Gauss distributed data

In the general case, there are no simple formulas for finding the values of the parameters that give the smallest $\chi^2$ and we need to use an iterative approach. We start with a guess of the best values of the parameters, $\bar{\lambda} = \bar{\lambda}_0$, and we calculate the gradient of $\chi^2$ around this point:

$$
\frac{\partial \chi^2}{\partial \lambda_k} \bigg|_{\bar{\lambda}_0} = g_k(\bar{\lambda}_0) = \sum_{i=1}^{n} -2 \left[ \frac{y_i - f(x_i|\bar{\lambda}_0)}{\sigma_i^2} \right] \frac{\partial f(x_i|\bar{\lambda}_0)}{\partial \lambda_k}.
$$

To find an extremum, we want to solve

$$
g_k(\bar{\lambda}_0 + \delta \bar{\lambda}) = \frac{\partial \chi^2}{\partial \lambda_k} \bigg|_{\bar{\lambda}_0 + \delta \bar{\lambda}} = 0 \quad \forall k .
$$

We expand in a Taylor series

$$
g_k(\bar{\lambda}_0 + \delta \bar{\lambda}) \approx g_k(\bar{\lambda}_0) + \sum_l \frac{\partial g_k}{\partial \lambda_l} \bigg|_{\bar{\lambda}_0} \delta \lambda_l = g_k(\bar{\lambda}_0) + \sum_l \frac{\partial^2 \chi^2}{\partial \lambda_k \partial \lambda_l} \bigg|_{\bar{\lambda}_0} \delta \lambda_l
$$

We want the term on the LHS of the equation to be zero, so we have

$$
g_k(\bar{\lambda}_0) = - \sum_l \frac{\partial^2 \chi^2}{\partial \lambda_k \partial \lambda_l} \bigg|_{\bar{\lambda}_0} \delta \lambda_l
$$

We can write this in matrix notation as

$$
G_0 \delta \bar{\lambda} = \bar{g}_0
$$
5.5. THE $\chi^2$ TEST STATISTIC

or

$$\delta \bar{\lambda} = G_0^{-1} \bar{g}_0$$

with

$$(G_0)_{kl} = \frac{\partial^2 \chi^2}{\partial \lambda_k \partial \lambda_l} \bigg| \bar{\lambda}_0.$$  

This gives us an iterative procedure to approach an extremum of $\chi^2$. We apply the shift to the parameters, and reevaluate the derivatives of $\chi^2$ around this new point, producing $\bar{g}_1$ and $G_1$ and a new shift in the parameters. We iterate until the change in the parameters and the change in $\chi^2$ becomes small enough. We note that the starting point can be critical, as well as choosing the step size for the parameters, in particular if the $\chi^2$ distribution does not have a simple unimodal shape.

5.5.9 $\chi^2$ for Bayesian and Frequentist Analysis

Frequentist use of $\chi^2$

As we have noted, $\chi^2$ is a test statistic. In the general case, for a frequentist analysis we need to find the probability distribution expected for $\chi^2$ for the particular data model at hand for possible values of the parameters of the model, and then find the range of parameter values where the value of $\chi^2$ for the data is within the $1 - \alpha$ range of $\chi^2$ values expected in the model. If we can treat the data as Gaussian distributed around the model expectation, then we can use the canonical $\chi^2$ distributions and we know the probability distribution of our test statistic - it is given by Eq. (5.17) and we can use the cumulative, or $p$-value, given in Eq. (5.18) to find the values of the parameters in our $1 - \alpha$ confidence interval. We require that the test statistic is within a $1 - \alpha$ probability range, where in this case we order $\chi^2$ from smallest to highest value, then we need to find the parameter values for which

$$p(\bar{\lambda}) > \alpha .$$

We note that the parameter region included in the $1 - \alpha$ Confidence Level region depends on the absolute value of $\chi^2_{Data}$ and is possibly empty. If the model chosen does not do a good job of representing the data for any values of the parameters, then even the best-fit values could be excluded.

Bayesian use of $\chi^2$

If we can model the probability of the data as a production of Gaussian probability distributions, and we can take the variance of the Gauss distributions as not de-
pending on the parameters, and we use flat priors for the parameters of the model, then we have seen that the resulting posterior probability distribution will be

$$P(\vec{\lambda}) \propto e^{-\chi^2(\vec{\lambda})/2}.$$ 

In this case, the $1 - \alpha$ credibility region for the parameters depends only on

$$\chi^2(\vec{\lambda}) - \chi^2_{\text{min}}$$

and we will always have accepted values for the parameters.

### 5.6 Maximum Likelihood analysis

We have pointed out that the probability of the data, viewed as a function of the parameters, is called the likelihood. and have mentioned that performing a maximum likelihood analysis is effectively the same as using the maximum of the posterior probability in a Bayesian analysis with flat priors. It is also possible to start with the statement that the best estimate for the parameters is the values of the parameters that maximize the probability of the data. This is taken by many as the starting point for an analysis.

Recalling the equivalence to a Bayesian analysis, we start with Bayes Theorem:

$$P(\lambda|D) = \frac{P(D|\lambda)P_0(\lambda)}{\int P(D|\lambda)P_0(\lambda) d\lambda}$$

and using a constant prior, we have

$$P(\lambda|D) \propto P(D|\lambda)$$

since the integral in the denominator just provides a normalization. We use the symbol $\mathcal{L}$ to represent the likelihood:

$$\mathcal{L}(\lambda) = P(D|\lambda) \propto P(\lambda|D).$$

It is therefore clear that the value of $\lambda$ that maximizes $\mathcal{L}(\lambda)$ also maximizes $P(\lambda|D)$. There is therefore a close relationship of a likelihood analysis to a Bayesian analysis. The main point of similarity is that the result of the analysis only uses the data that was acquired. This is in contrast to a frequentist analysis, which relies for its results on all possible experimental results, not just the data acquired. There are however also significant differences to a Bayesian analysis.
5.6. MAXIMUM LIKELIHOOD ANALYSIS

The first point to realize is that a likelihood is a function of the parameters and is not a probability or probability density, since it is not normalized. The second point is that the maximum of the likelihood is independent of the parameters used in the analysis. This is not generally the case for a Bayesian analysis, since a prior flat in one parameter is usually not flat after reparametrization. As an example of this, we consider parametrization a Poisson process by a rate and again by the mean time between events:

\[ L(R) = \frac{(RT)^n e^{-RT}}{n!} \]

which has a maximum at \( R^* = n/T \) or using \( \tau = 1/R \)

\[ L(\tau) = \frac{(T/\tau)^n e^{-T/\tau}}{n!} \]

which has a maximum at \( \tau^* = T/n \). We have that \( R^* = 1/\tau^* \).

However, if we assume that the prior is flat in \( R \), then it is not flat in \( \tau \), since

\[ P_0(\tau) = P_0(R) \left| \frac{dR}{d\tau} \right| = P_0(R)1/\tau^2 \]

and

\[ P(\tau|n) = T(T/\tau)^{n+2} e^{-T/\tau} \]  \hspace{1cm} (5.28)

yielding

\[ \tau^* = \frac{T}{n + 2} \, . \]

We no longer have that \( R^* = 1/\tau^* \). The reason is that in a Bayesian analysis, we have to be consistent with our prior definitions and in this case a flat prior in \( R \) leads to a non-flat prior in \( \tau \).

5.6.1 Parameter uncertainties in likelihood analysis

Suppose the maximum of the likelihood function (mode) is at \( \lambda^* \). If the likelihood function is approximately Gaussian, there are standard ways to quote results. One way to describe the experimental results is by quoting \( \lambda^* \) and the standard error

\[ \Delta \lambda = \left[ \frac{\int (\lambda - \lambda^*)^2 L(\lambda) d\lambda}{\int L(\lambda) d\lambda} \right]^{1/2} \]  \hspace{1cm} (5.29)
We note that if the likelihood is indeed Gaussian in shape, then \( \lambda^* = E[\lambda] \) and furthermore \( L(\lambda) / \int L(\lambda)d\lambda \) is normalized so that \( \Delta \lambda \) is the standard deviation of the normalized Gaussian function.

Again assuming the Gaussian shape for the likelihood

\[
L(\lambda) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\lambda - \lambda^*)^2}{2\sigma^2}},
\]

we can rewrite the standard error as follows:

\[
\frac{d}{d\lambda} \ln L(\lambda) = -\frac{\lambda - \lambda^*}{\sigma^2}
\]

and

\[
\frac{d^2}{d\lambda^2} \ln L(\lambda) = -\frac{1}{\sigma^2}
\]

so that we have

\[
\Delta \lambda = \sigma = \left[ -\frac{d^2}{d\lambda^2} \ln L(\lambda) \right]^{-1/2}.
\] (5.30)

This form is easier to use in practice since we do not need to perform an integral and just need to map out the likelihood near its maximum so that we can calculate the second derivative. Of course, all of this relies on the assumption that the likelihood is Gaussian in shape.

**Example**

Suppose we are trying to determine the lifetime of an unstable particle, and that we have measured several decay times. Our model of the probability for a particle to have decayed in the time interval \( t \rightarrow t + dt \) is

\[
P(t)dt = \frac{1}{\tau} e^{-t/\tau} dt
\]

where \( \tau \) is the ‘lifetime’ of the particle. Our measured decay times are \( \{t\} \) and we assume we have \( N \) measurements. We have different options for our probability of the data, or likelihood. E.g., we can define the likelihood as a product of the individual probability densities

\[
L(\tau) = P(\{t\} | \tau) = \prod_{i=1}^{N} \frac{1}{\tau} e^{-t_i/\tau}.
\]
This goes under the name of ‘unbinned likelihood’. Another option is to make bins in time and use Poisson statistics

\[ \mathcal{L}(\tau) = \prod_{j=1}^{M} \frac{n_j^{\nu_j}}{\nu_j!} e^{-\nu_j} \]

where we have made \( M \) time bins, \( n_j \) is the number of decays measured in the \( j \)th time bin and

\[ \nu_j = \frac{T}{\tau} \int_{\Delta t_j} \frac{1}{\tau} e^{-t/\tau} dt \]

is the expected number of events in time bin \( j \). This would then be a ‘binned likelihood’ analysis.

We consider here the unbinned likelihood case in more detail. We can rewrite the likelihood as

\[ \ln \mathcal{L}(\tau) = -N \ln \tau - \frac{1}{\tau} \sum_{i=1}^{N} t_i . \]

It is usually simpler to work with the logarithm of the likelihood (the mode is at the same value of the parameter):

\[ \ln \mathcal{L}(\tau) = -N \ln \tau - \frac{1}{\tau} \sum_{i=1}^{N} t_i . \]

Taking the derivative to find the mode, we have

\[ \frac{d}{d\tau} \ln \mathcal{L}(\tau) = -\frac{N}{\tau} + \frac{\sum_{i=1}^{N} t_i}{\tau^2} = 0 \quad \Rightarrow \quad \tau^* = \frac{1}{N} \left( \sum_{i=1}^{N} t_i \right) \]

so that the mode is at the average of the measured times, \( \bar{t} \). To calculate the standard error, we need the second derivative

\[ \frac{d^2}{d\tau^2} \ln \mathcal{L}(\tau) = \frac{N}{\tau^2} - 2 \frac{\sum_{i=1}^{N} t_i}{\tau^3} \]

and applying Eq. 5.30 we have

\[ \Delta \tau = \left( -\frac{N}{\tau^2} + 2 \frac{\sum_{i=1}^{N} t_i}{\tau^3} \right)^{-1/2} . \]
We can evaluate this at any value of $\tau$. If the likelihood is indeed Gaussian, we will always get the same value. Let’s use $\tau = \tau^*$. This yields

$$\Delta \tau = \frac{\tau^*}{\sqrt{N}}$$

and we would report

$$\hat{\tau} = \bar{t} \pm \frac{\bar{t}}{\sqrt{N}}.$$

Is this a reasonable result? It depends very much on whether the likelihood is indeed Gaussian in shape. Example likelihoods are plotted in Fig. 5.14. We see that for $N = 2$, we have a rather non-Gaussian distribution. For $N = 10$, it is somewhat more Gaussian. In these cases, the standard error do not correspond to 68% probability intervals (see exercises).

![Figure 5.14: Example likelihood functions for $N = 2$ (left) and $N = 10$ each for two different values of $\tau^*$.](image)

We note that if the likelihood is indeed Gaussian in shape, then the

### 5.6.2 Likelihood ratio as test statistic

Likelihood ratios are often used as test statistics (at least in particle physics, astronomy and astrophysics). The likelihood function in this case is normalized to the maximum of the likelihood:

$$\xi(\lambda) = \frac{\mathcal{L}(\lambda)}{\mathcal{L}(\lambda^*)}$$  \quad (5.31)
so that we have a number ranging between 0 → 1 with increasing values corresponding to higher likelihoods. If the distribution of the likelihood ratio for our model can be explicitly determined then it can directly be used to form confidence level regions for \( \lambda \). In most cases, however, the exact distribution of the likelihood ratio is very difficult to determine analytically and numerical simulations must be performed. An asymptotic result\(^9\) is that, as the sample size approaches \( \infty \), the test statistic \(-2 \ln(\xi)\) will be \( \chi^2 \)-distributed with the number of degrees of freedom corresponding to the dimension of \( \lambda \).

**Example - frequentist analysis using likelihood ratio**

We try this out on the example of the measurement of a set of decay times where our model is that the time distribution should be exponentially distributed. In this case, we know the distribution for the sum of the times (it is the same as the probability for the waiting time for the \( N^{th} \) event in a Poisson process -you worked it out in exercise 3.3). It is

\[
P(T = \sum_{i=1}^{N} t_i) = R \left( \frac{RT}{(N-1)!} \right) e^{-RT} \tag{5.32}
\]

with \( R = 1/\tau \). We can use this to find the probability distribution for the \( \xi' = -2 \ln(\xi(\tau)) \) as follows:

\[
\xi' = -2 \ln \xi \\
= -2 \ln \left[ \frac{\mathcal{L}(\tau)}{\mathcal{L}(\tau^*)} \right] \\
= -2 \ln \left[ \left( \frac{\tau^*}{\tau} \right)^N e^{N(\bar{t}/\tau^* - \bar{t}/\tau)} \right] \\
= -2 \ln \left[ \left( \frac{\bar{t}}{\tau} \right)^N e^{N(1-\bar{t}/\tau)} \right] \\
= -2N \left[ \ln \bar{t} - \ln \tau + (1 - \bar{t}/\tau) \right] \\
= -2N \left[ \ln \frac{T}{N} - \ln \tau + (1 - \frac{T}{N\tau}) \right].
\]

It is difficult (impossible?) to get an analytic form for $P(\xi')$, so we proceed with a numerical algorithm.

We proceed as follows:

1. Define a grid of values for $\tau$

2. For each $\tau$, sample from Eq. 5.32 many times, each time using the expression above to find $\xi'$, and from this build the distribution for $\xi'$

3. For each $\tau$, we find the $p$-value for our data (where we have a fixed $T_{\text{Data}}$). The $p$-value is calculated by finding the cumulative of $P(\xi')$ starting from 0 (which corresponds to max likelihood) and defining

$$p = 1 - F(\xi').$$

4. We accept the values of $\tau$ for which the $p > \alpha$. This yields a $1 - \alpha$ confidence level range for $\tau$.

We try this for some concrete values $\{t\} = \{1.2, 2.5, 0.2, 0.8, 0.5, 1.7\}$. For this data, we have $T_{\text{Data}} = 6.9$ and $\bar{t} = 1.15$. We anticipate that reasonable values of $\tau$ will be in the range $0.5 - 2.4$, so we start by defining a grid in this range with spacing 0.1. The 20 resulting distributions for $T$ and $\xi'$ are shown in Figs. 5.15-5.16. We observe from the plots that for the smaller and larger values of $\tau$, that the data value is in the tails of the distributions. For the values of $\tau$ near $\tau = 1$, the observed value in the data is in the bulk of the distribution, indicating that the $p$ value is reasonably high. Indeed, from the analysis, we find that $p > 0.32$ for $\tau \in [0.79, 1.77]$ so that this would be the 68% confidence level range of values for $\tau$. We used $\xi'$ as a test statistic for this analysis, but could easily well have used $T$. This is saved for the exercises.

We can also check to see how well Wilk’s Theorem applies in this case. We expect that the distribution of $\xi'$ should be distributed according to a $\chi^2$ distribution with one degree of freedom in the limit of $N \to \infty$. In our case, $N = 6$, but we find that the distribution of $\xi'$ is nevertheless very well reproduced by the $\chi^2$ distribution. An example is shown in Fig. 5.17 for $\tau = 1.0$. We could therefore use the cumulative of the $\chi^2$ distribution to find our accepted values of $\tau$ at $1 - \alpha$ confidence. For 60% CL, we require that $\chi^2 \leq 1$. The values of $\xi'_{\text{Data}}(\tau)$ are shown in Fig. 5.17, and we see indeed that we can get the 68% CL range for $\tau$ to good approximation using Wilk’s Theorem.
5.6. MAXIMUM LIKELIHOOD ANALYSIS

![Graph showing distribution of T = \( \sum_i t_i \) for different \( \tau \) for \( N = 6 \) measurements. The values of \( \tau \) range from 0.5 in the upper left plot to 2.4 in the lower right in steps of 0.1. The value observed in the data is shown with the vertical red line.]

5.6.3 Likelihood ratio for Poisson distributed data

Another test statistic option for a Poisson model is based on the (log of) the likelihood ratio \[11\]

\[
R = 2 \log \frac{P(\tilde{D}|\nu_i = m_i)}{P(\tilde{D}|\nu_i = \nu_i(\lambda))} = 2 \sum_{i=1}^{N} \left[ \nu_i - m_i + m_i \log \frac{m_i}{\nu_i} \right].
\] (5.33)
Figure 5.16: Distributions of $\xi'$ for different $\tau$ for $N = 6$ measurements. The values of $\tau$ range from 0.5 in the upper left plot to 2.4 in the lower right in steps of 0.1. The value observed in the data is shown with the dashed vertical line.

In bins where $m_i = 0$, the last term is set to 0. Again, rather than using the probability distribution of this discrepancy variable directly, since $R$ has a distri-
5.7. MODEL SELECTION

There are three main schools of thought associated with Fisher, Jeffreys and Neyman. Fischer argued for the use of p-values as ‘evidence’ against a hypothesis; Jeffreys argued for Bayes factors, and Neyman argued for predefined data probabilities and accepting Type I and Type II errors. We describe each in a little detail here with a simple example:
5.7.1 Fisher’s significance testing

We suppose there is a model we want to test where a parameter takes on a fixed value \( \lambda_0 \). The model makes the prediction

\[ y \sim P(y|\lambda) \]

and we define the hypothesis \( H_0 \) as 'The data is distributed according to \( P(y|\lambda_0) \). Then, define a test statistic \( T(y) \) where large values of \( T \) reflect evidence against \( H_0 \) (e.g., \( \chi^2 \)). Then compute a \( p \)-value

\[ p = P(T(y) \geq T(y^{\text{Data}}) | \lambda_0) \]

and reject \( H_0 \) if \( p \) is small. Obviously this only gives negative evidence against \( H_0 \) and not positive evidence for another hypothesis. The use of \( p \)-values as a measure of support has been shown to not be logically coherent (see [3]). While this procedure is in wide use in particle physics, I would argue against its use.

5.7.2 Neyman-Pearson hypothesis testing

Neyman argued that only model comparisons make sense and that there should be an alternative hypothesis that we test against. Let’s call the alternative hypothesis \( H_1 \), and this is the hypothesis that 'The data is distributed according to \( f(x|\lambda_1) \). The procedure advocated by Neyman is to reject \( H_0 \) if \( T(y^{\text{Data}}) \geq c \) assuming \( \lambda = \lambda_0 \) and accept it otherwise. Here \( c \) is a ‘critical value’ that was chosen before the data is taken. The Type I and Type II error probabilities are

- Type I \( \alpha = P(T(y^{\text{Data}}) | \lambda_0) \geq c \) is the probability of rejecting model hypothesis I when it is in fact correct;

- Type II \( \beta = P(T(y^{\text{Data}}) | \lambda_1) < c \) is the probability of accepting model hypothesis I when it is in fact incorrect.

The argument for this approach is that it obeys the frequentist principle that if one had a very large number of trials, the average of the reported error rates would be the correct error rates. It is debatable how interesting this long-term behavior of error rates is in practice.
5.7.3 Bayesian posterior probability test

Jeffrey’s approach is based on Bayes’ Theorem. We define the Bayes factor

\[ B = \frac{P(y|\lambda_0)}{P(y|\lambda_1)} \]

The prior odds are

\[ O_o = \frac{P_0(\lambda_0)}{P_1(\lambda_1)} \]

and the posterior odds are

\[ O = B \frac{P_0(\lambda_0)}{P_1(\lambda_1)} \]

The positive decision for a model can be based on its posterior probability or on the posterior odds (in case there are only two hypotheses). Schervish has shown that this is the only coherent approach [3] to building support for a hypothesis, and this is the approach that this author favors.

5.8 Test of \( p \)-value definitions, from Beaujean et al.

In the following, we test the usefulness of different \( p \)-values by looking at their distributions for specific examples motivated from common situations faced in experimental physics. We first consider a data set which consists of a background known to be smoothly rising and, in addition to the background, a possible signal. This could correspond for example to an enhancement in a mass spectrum from the presence of a new resonance. The width of the resonance is not known, so that a wide range of widths must be allowed for. Also, the shape of the background is not well known. We assume we do not have an exhaustive set of models to compare and want to look at \( p \)-values for models individually to make decisions. In this example, we will first assume that distributions of the data relative to expectations are modeled with Gaussian distributions. We then consider the same problem with small event numbers, where Poisson statistics are appropriate, and again test our different \( p \)-value definitions. These examples were also discussed in [17]. Finally, we consider the case of testing an exponential decay law on a sample of measured decay times.


5.8.1 Data with Gaussian uncertainties

Definition of the data

A typical data set is shown in Fig. 5.18. It is generated from the function

\[ f(x_i) = A + B x_i + C x_i^2 + \frac{D}{\sigma \sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}. \]

(5.34)

with parameter values \((A = 0, B = 0.5, C = 0.02, D = 15, \sigma = 0.5, \mu = 5.0)\). The \(y_i\) are generated from \(f(x_i)\) as

\[ y_i = f(x_i) + r_i \cdot 4 \]

where \(r_i\) is sampled according to \(G(0, 1)\).

Figure 5.18: Example data set for the case \(N = 25\) with Gaussian fluctuations. The fits of the four models are superposed on the data.

The \(x\) domain is \([0, 20]\) and two cases were considered: \(N = 25\) data points evenly sampled in the range, and \(N = 100\) data points evenly sampled in the range. The experimental resolution (Gaussian with width 4) is assumed known and correct. Ensembles consisting of 10,000 data sets with \(N = 25\) or \(N = 100\)
data points each were generated and four different models were fitted to the data. Table 5.1 summarizes the parameters available in each model and the range over which they are allowed to vary. In all models, flat priors were assumed for all parameters for ease of comparison between results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Par</th>
<th>Small Range</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>I.</td>
<td>$A_I + B_I x_i + C_I x_i^2$</td>
<td>$A_I$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$B_I$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$C_I$</td>
<td>-0.1</td>
</tr>
<tr>
<td>II.</td>
<td>$A_{II} + \frac{D_{II}}{\sigma_{II}} e^{-\frac{(x_i-\mu_{II})^2}{2\sigma_{II}^2}}$</td>
<td>$A_{II}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$B_{II}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mu_{II}$</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sigma_{II}$</td>
<td>0.2</td>
</tr>
<tr>
<td>III.</td>
<td>$A_{III} + B_{III} x_i + \frac{D_{III}}{\sigma_{III}} e^{-\frac{(x_i-\mu_{III})^2}{2\sigma_{III}^2}}$</td>
<td>$A_{III}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$B_{III}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$D_{III}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mu_{III}$</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sigma_{III}$</td>
<td>0.2</td>
</tr>
<tr>
<td>IV.</td>
<td>$A_{IV} + B_{IV} x_i + C_{IV} x_i^2 + \frac{D_{IV}}{\sigma_{IV}} e^{-\frac{(x_i-\mu_{IV})^2}{2\sigma_{IV}^2}}$</td>
<td>$A_{IV}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$B_{IV}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$C_{IV}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$D_{IV}$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\mu_{IV}$</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sigma_{IV}$</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 5.1: Summary of the models fitted to the data, along with the ranges allowed for the parameters.

The models were fitted one at a time. Two different fitting approaches were used:

1) The fitting was performed using the gradient based fitting package MIGRAD from the MINUIT package [18] accessed from within BAT [17]. The starting values for the parameters were chosen at the center of the allowed ranges given in Table 5.1.
2) The Markov Chain Monte Carlo (MCMC) implemented in BAT was run with its default settings first, and then MIGRAD was used to find the mode of the posterior using the parameters at the mode found by the MCMC as starting point.

Since it is known that the distributions of the data are Gaussian and flat priors were used for the parameters, the maximum of the posterior probability corresponds to the minimum of \( \chi^2 \). The \( p \)-values were therefore extracted for each model fit using the \( \chi^2 \) probability distribution.

**Comparison of \( p \)-values**

The \( p \)-value distributions for the four models are given in Fig. 5.19 for the case \( N = 25 \). Two different histograms are shown for each model, corresponding to the two fitting techniques described above. The distributions for models I and II are peaked at small values of \( p \) for MIGRAD only and for MCMC+MIGRAD, and the models would be disfavored in most experiments. For models III and IV, there is a significant difference in the \( p \)-value distribution found from fitting using only MIGRAD, or MCMC+MIGRAD, with the \( p \)-value distribution in the latter case being much flatter. An investigation into the source of this behavior showed that the \( p \)-value distribution from MIGRAD is very sensitive to the range allowed for the parameters. Even in this rather simple fitting problem, it is seen that the use of the MCMC can make a significant difference in the fitting results.

**5.8.2 Example with Poisson uncertainties**

**Definition of the data**

We again use the function (5.34), but now generate data sets with fluctuations from a Poisson distribution. The function is used to give the expected number of events in a bin covering an \( x \) range, and the observed number of events follows a Poisson distribution with this mean. For \( N = 25 \) bins, e.g., the third bin is centered at \( x_3 = 2.0 \) and extends over \([1.6, 2.4]\). The expected number of events in this bin is defined as \( \nu_3 = \int_{1.6}^{2.4} \frac{N}{20} f(x) \, dx \).

**Comparison of \( p \)-value distributions**

Different fits were performed for the different discrepancy variable definitions considered. For the \( \chi^2 \) discrepancy variables, \( \chi^2 \) minimization was performed us-
Figure 5.19: p-value distributions from $\chi^2$ fits to $N = 25$ data points for the four models discussed in Table 5.1 and allowing parameter values in the small range. The p-values correspond to $N - n$ degrees of freedom, where $n$ is the number of fitted parameters.
CHAPTER 5. MODEL FITTING AND MODEL SELECTION

Figure 5.20: The $p$-values are from $\chi^2$ fits to $N = 100$ data points and correspond to $N - n$ degrees of freedom.
ing either the expected or observed number of events as weight (Pearson or Neyman $\chi^2$). For the likelihood ratio test, a maximum likelihood fit was performed. The likelihood was defined as

$$L(\vec{\lambda}) = \prod_{i=1}^{N} \frac{e^{-\nu_i} n_i^{\nu_i}}{n_i!}$$

(5.35)

where $n_i$ is the observed number of events in a bin and $\nu_i = \nu_i(\vec{\lambda})$. Since we use flat priors for the parameters, the same results were used for the case where the discrepancy variable is the probability of the data. A typical likelihood fit result is shown in Fig. 5.21 together with the data, while the $p$-value distributions are given in Figs. 5.22-5.23.

![Figure 5.21: Example fits to one data set for the four models defined in Table 5.1 for $N = 25$ and the small parameter range. The lines shows the fit functions evaluated using the best-fit parameter values.](image)

In general, the $p$-value is not uniformly distributed for any of the discrepancy variables, even using the true value of the parameters, except in the case where we use the probability of the data. This is because in all other cases, approximations are used for the probability distributions of the discrepancy variables.
For all definitions of \( p \)-values considered here, models I,II would generally lead to low degree-of-beliefs, while models III and IV would generally have high degree-of-beliefs. However, there are differences in the behavior of the \( p \)-values which we now discuss in more detail.

The lower left panel in Figs. 5.22-5.23 show the \( p \)-value distribution taking the true model and the true parameters, and can be used as a gauge of the biases introduced by the \( p \)-value definition. There is a bias for all discrepancy variables shown in Fig. 5.22 because approximations are used for the probability distribution of the discrepancy variable (using \( P(\chi^2|N) \) rather than \( P(R = \chi^2_p|N) \) or \( P(R = \chi^2_N|N) \)). This bias is usually small, except in the case where the Neyman \( \chi^2 \) discrepancy variable is used. In this case, very small \( p \)-values can occur even with the true parameter values. The \( p \)-value distribution using the probability of the data, on the other hand, is flat when the true parameters are used as seen in Fig. 5.23, so that this choice would be optimal in cases where no parameters are fit.

**Pearson's \( \chi^2 \)**

Comparison of the behavior of the Pearson \( \chi^2 \) to the Neyman \( \chi^2 \), Fig. 5.22, clearly shows that using the expected number of events as weight is superior. The spike at 0 in the \( p \)-value distribution when using the observed number of events indicates that this quantity does not behave as expected for a \( \chi^2 \) distribution when dealing with small numbers of events, and will lead to undesirable conclusions more often than anticipated. The behavior of the Pearson \( \chi^2 \) using the expected number of events for the different models is quite satisfactory, and this quantity makes a good discrepancy variable even in this example with small numbers of events in the bins.

**Likelihood Ratio**

While models (I,II) are strongly peaked at small \( p \)-values, for this definition of a \( p \)-value models III,IV also have a falling \( p \)-value distribution. This is somewhat worrisome. In assigning a degree-of-belief to models based on this \( p \)-value, this bias towards smaller \( p \)-values should be considered, otherwise good models will be assigned too low a degree-of-belief. The behavior of the \( p \)-value extracted from Pearson's \( \chi^2 \) is preferable to the likelihood ratio, as it will lead to value judgements more in line with expectations.
5.8. TEST OF $P$-VALUE DEFINITIONS, FROM BEAUMEAN ET AL. 193

Figure 5.22: $p$-value distributions for ensembles of 10000 experiments generated as described in the text, for $N = 25$ bins. The $p$-values are from Pearson’s $\chi^2$ (blue histogram), Neyman’s $\chi^2$ (red dotted histogram), and from the likelihood ratio test (green dashed histogram). The $p$-values are calculated for $N - n$ degrees of freedom.
Probability of the Data

Using this definition, the $p$-value distribution is flat for the true model, since this is the only case where the correct probability distribution of the discrepancy variable is employed. For the fitted models, two $p$-value distributions are shown - one uncorrected for the number of fitted parameters, and one corrected for the fitted parameters (by transforming back-and-forth to a $\chi^2$ probability). Using the uncorrected $p$-values, both models III and IV show $p$-value distributions peaking at $p = 1$, so that both models would typically be kept. As expected, the correction for the number of fitted parameters pushes $p$ to lower values, and produces a nearly flat distribution for model IV. This ad-hoc correction works well for this example. In general, this $p$-value definition has similar properties to the Pearson $\chi^2$ statistic, but with the advantage of a flat $p$-value distribution when the true model was used and no parameters were fit.

5.8.3 Exponential Decay

When dealing with event samples with continuous probability distributions for the measurement variables, it is common practice when determining the parameters of a model to use a product of event probabilities (unbinned likelihood):

$$P(\vec{x}|\vec{\lambda}, M) = \prod_{i=1}^{N} P(x_i|\vec{\lambda}, M).$$

If the model chosen is correct, then this definition for the probability of the data (or likelihood) can be used successfully to determine appropriate ranges for the parameters of the model. However, $P(\vec{x}|\vec{\lambda}, M)$ defined in this way has no sensitivity to the overall shape of the distribution and can lead to unexpected results if this quantity is used in a goodness-of-fit test of the model in question. We use a common example, exponential decay, to illustrate this point (see also discussion in [19] and [7]).

Our model is that the data follows an exponential decay law. We measure a set of event times, $\vec{t}$, and analyze these data to extract the lifetime parameter $\tau$. We define two different probabilities of the data $D = \{t_i|i = 1 \ldots N\}$

1. Unbinned likelihood

$$P(\vec{t}|\tau) = \prod_{i=1}^{N} \frac{1}{\tau} e^{-t_i/\tau},$$
Figure 5.23: \( p \)-value distributions for ensembles of 10000 experiments generated as described in the text, for \( N = 25 \) bins. The \( p \)-values are from the data probability, \( P(\hat{D}|\hat{X}^*, M) \). The blue histograms are for \( N \) degrees of freedom, and the red histograms are for corrections to \( N - n \) degrees of freedom of freedom (see text).
2. Binned likelihood

\[ P(\vec{t} | \tau) = M \prod_{j=1}^{M} \nu_j^{n_j} \frac{n_j!}{\nu_j^{n_j}} e^{-\nu_j}, \quad \nu_j(\tau) = \int_{t_j}^{t_j+\Delta t} dt \frac{N}{\tau} e^{-t/\tau}. \]

In the first case, the probability density is a product of the densities for the individual events, while in the second case the events are counted in time intervals and Poisson probabilities are calculated in each bin. The expected number of events is normalized to the total number of observed events. We consider time intervals with a width \( \Delta t = 1 \) unit and time measurements ranging from \( t = 0 \) to \( t = 20 \). The overall probability is the product of the bin probabilities. For each of these two cases, we consider the \( p \)-value determined from the distribution of \( R = P(\vec{t} | \tau) \).

In order to make the point about the importance of the choice for the discrepancy variable for goodness-of-fit tests, we generated data which do not follow an exponential distribution. The data is generated according to a linearly rising function, and a typical data set is shown in Fig. 5.24.

**Product of exponentials**

If the data are fitted using a flat prior probability for the lifetime parameter, then we can solve for the mode of the posterior probability of \( \tau \) analytically, and get the well-known result

\[ \tau^* = \frac{1}{N} \sum t_i. \]

Defining \( \xi \equiv \sum t_i \), so \( \tau^* = \xi/N \), we can also solve analytically for the \( p \)-value:

\[ p = \int_{\sum t_i>\xi} dt' \int dt'_2 \cdots (\tau^*)^{-N} e^{-\sum t_i/\tau^*} \]

and the result is

\[ p = 1 - P(N,N) \]

with the regularized incomplete Gamma function

\[ P(s,x) = \frac{\gamma(s,x)}{\Gamma(s)} = \frac{\int_0^x t^{s-1}e^{-t}dt}{\int_0^\infty t^{s-1}e^{-t}dt}. \]
Figure 5.24: Typical data set used for the fitting of the exponential model. The individual event times are shown, as well as the binned contents as defined in the text. The best fit exponential from the unbinned likelihood is also shown on the plot, normalized to the total number of fitted events.
Surprisingly, as \( N \) increases \( p \) is approximately constant with a value \( p \approx 0.5 \). Regardless of the actual data, \( p \) is never small and depends only on \( N \). The best fit exponential is compared to the data in Fig. 5.24, and yields a large \( p \)-value although the data is not exponentially distributed. This \( p \)-value definition is clearly useless, and the ‘unbinned likelihood’ is seen to not be appropriate as a goodness-of-fit statistic.

The definition of \( \chi^2 \) in the Gaussian data uncertainties case also results from a product of probability densities, so the question as to what is different in this case needs clarification. The point is that in the case of fitting a curve to a data set with Gaussian uncertainties, the data are in the form of pairs, \( \{(x_i, y_i)\} \) where the measured value of \( y \) is assigned to a particular value of \( x \), and the discrepancy between \( y \) and \( f(x) \) is what is tested. I.e., the value of the function is tested at different \( x \) so the shape is important. In the case considered here, the data are in the form \( \{x_i\} \), and there is no measurement of the value of the function \( f(x) \) at a particular \( x \). The orderings of the \( x_i \) is irrelevant, and there is no sensitivity to the shape of the function.

**Product of Poisson probabilities**

In this case, the \( p \)-value for the model is determined from the product of Poisson probabilities using event counting in bins as described above. Since the expected number of events now includes an integration of the probability density and cover a wide range of expectations, it is sensitive to the distribution of the events and gives a valuable \( p \)-value for goodness-of-fit. The \( p \)-value distribution from the Pearson \( \chi^2 \) for the data shown in Fig. 5.24 is \( p = 0 \) within the precision of the calculation. The exponential model assumption is now easily ruled out.

In comparison to the unbinned likelihood case, the data are now in the form \( \vec{n} \), where the \( n_i \) now refer to the number of events in a well-defined bin \( i \) which defines an \( x \) range. I.e., we have now a measurement of the height of the function \( f(x) \) at particular values of \( x \), and are therefore sensitive to the shape of the function.

As should be clear from this example, the choice of discrepancy variable can be very important in goodness-of-fit decisions.
Evidence of the effect on the results of using different definitions of $p$-values.

**Literature**

CHAPTER 5. MODEL FITTING AND MODEL SELECTION

Exercises

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

<table>
<thead>
<tr>
<th>Energy ($E_i$)</th>
<th>Trials ($N_i$)</th>
<th>Successes ($r_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>1.0</td>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>1.5</td>
<td>100</td>
<td>22</td>
</tr>
<tr>
<td>2.0</td>
<td>100</td>
<td>55</td>
</tr>
<tr>
<td>2.5</td>
<td>100</td>
<td>80</td>
</tr>
<tr>
<td>3.0</td>
<td>100</td>
<td>97</td>
</tr>
<tr>
<td>3.5</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td>4.0</td>
<td>100</td>
<td>99</td>
</tr>
</tbody>
</table>

(a) Find the posterior probability distribution for the parameters ($A$, $E_0$)
(b) Find the 68 % CL region for ($A$, $E_0$)

2. 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

3. Derive the mean, variance and mode for the $\chi^2$ distribution for one data point.

4. Using the Central Limit Theorem, argue that the $\chi^2$ distribution for large enough number of data points approaches a Gauss probability distribution. What are the parameters of the Gauss distribution?

5. Use $\chi^2$ as defined in Section 5.5 as a test-statistic to find the allowed regions in the parameter space of ($A$, $E_0$) at $1 - \alpha = 0.68$ for the efficiency example of Section 5.1.2
6. Generate 10 values of \( n \) according to a Poisson distribution and compare the Pearson \( \chi^2 \) distribution with the Neyman \( \chi^2 \) distribution and the \( \chi^2 \) distribution for Gauss distributed data with \( \sigma^2 = \nu \) for the following values:

(a) \( \nu = 1 \)
(b) \( \nu = 5 \)
(c) \( \nu = 20 \).

7. Perform a straight line fit of the function \( f(x|m, b) = mx + b \) using \( \chi^2 \) minimization to the following data.

<table>
<thead>
<tr>
<th>i</th>
<th>x</th>
<th>y</th>
<th>( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>4.5</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>8.5</td>
<td>1.2</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>12.9</td>
<td>1.4</td>
</tr>
<tr>
<td>4</td>
<td>4.0</td>
<td>15.5</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td>5.0</td>
<td>21.2</td>
<td>1.0</td>
</tr>
<tr>
<td>6</td>
<td>6.0</td>
<td>25.5</td>
<td>1.2</td>
</tr>
<tr>
<td>7</td>
<td>7.0</td>
<td>31.2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

(a) Find the best-fit values of the parameters and the covariance matrix.
(b) What is the \( \chi^2 \) value for the fit?
(c) Make a plot of the results.
(d) Find the uncertainty in the predicted expectation value for \( y \) at an arbitrary \( x \) using

\[
\sigma_f^2 = \left( \frac{\partial f}{\partial m} \right)^2 \sigma_m^2 + \left( \frac{\partial f}{\partial b} \right)^2 \sigma_b^2 + 2 \left( \frac{\partial f}{\partial m} \right) \left( \frac{\partial f}{\partial b} \right) \sigma_m \sigma_b
\]

(e) plot this uncertainty.

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 - which is a better fit?

(a) quadratic, representing background only:

\[
f(x|A, B, C) = A + Bx + Cx^2
\]
(b) quadratic + Breit-Wigner representing background+signal:

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

Find the best value of the parameters. Try different priors for the parameters to see what happens.

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

9. Derive Eq. 5.28 and show that \( \tau^* = \frac{T}{n+2} \)

10. Consider the lifetime measurement example Section 5.6.1. We have the following measurement:

(a) \( N = 2 \) \( t_1 = 1, t_2 = 2 \)

\[ \mathcal{L}(\tau) = \frac{1}{\tau^2} e^{-3/\tau} \implies \hat{\tau} = 1.5 \pm 1.1 \]

(b) \( N = 10 \) \( t_i = i \)

\[ \mathcal{L}(\tau) = \frac{1}{\tau^{10}} e^{-55/\tau} \implies \hat{\tau} = 5.5 \pm 1.7 \]

(a) Confirm the results for \( \hat{\tau} \) for the two data sets.

(b) Perform a Bayesian analysis, using a flat prior for \( \tau \). Find the mode and smallest 68% probability interval for \( \tau \) and compare to the results from the likelihood analysis.

11. Repeat the analysis in Section 5.6.2 using \( T \) as your test statistic.

12. (a) Verify Eq. 5.33

(b) Try it out for the following model:

\[ \nu(\lambda) = \lambda x \]

and the following data:
(c) Show that the best fit value of $\lambda$ using a maximum likelihood approach is

$$\lambda^* = \frac{\sum_i n_i}{\sum_i x_i}$$

(d) What $p$-value do you get using the approximate formula in the text?
Appendix-A

We use the same notation to describe probabilities and probability densities: $P(x)$ is the probability (or probability density) of $x$. We use the general definition of probability; i.e., a ‘degree-of-belief’. A conditional probability (or density) is written as $P(x|y)$ means the probability (or probability density) of $x$ given $y$. Usually the context will make it obvious whether we are dealing with probabilities or probability densities. Joint probability densities are written as $P(\vec{x})$ and joint conditional probability densities as $P(\vec{x}|\vec{y})$.

A probability (conditional or not) must satisfy the following conditions:

$$0 \leq P(x) \leq 1$$

$$\sum_{x_{\text{min}}}^{x_{\text{max}}} P(x) = 1$$

For probability densities, the conditions are

$$0 \leq P(x)$$

$$\int_{x_{\text{min}}}^{x_{\text{max}}} P(x)dx = 1.$$ 

For multidimensional probability densities, we have

$$0 \leq P(\vec{x})$$

$$\int P(\vec{x})d\vec{x} = 1.$$ 

The marginal probability densities are

$$P(x_i) = \int P(\vec{x}) \prod_{j \neq i} dx_j.$$
The mode of a function is defined as the parameter values where the function attains its maximum. Function can be unimodal, multimodal or have an infinite number of modes. We use the notation $^*$ to denote the mode. E.g.,

$$x^* = \mu = \arg\max_x N(x|\mu, \sigma)$$

where $N(x|\mu, \sigma)$ is a normal (Gauss) distribution centered at $\mu$ and with standard deviation $\sigma$. The expectation value
Bibliography


