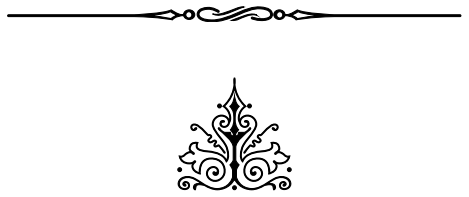
Statistics for Astronomy

Lecture notes

2nd year undergraduate course,
Semester 1a, 2024-2025

Harish K. Vedantham

Netherlands Institute for Radio Astronomy
and
University of Groningen, Netherlands



Contents

Chapter 1

Plausible reasoning and Bayes' theorem

Here is a basic truth about most situations you are going to encounter in life— you won't have enough information to assert that something is *definitively* true or false. The best you can do is to say that something is likely true or probably true and assign to it a measure of 'definiteness' or 'belief'. This measure of belief is called probability. If the answers to most problems we will encounter are probabilistic, then surely we all must strive to be good at calculating it. This course will set you on that path. The first thing you need to learn on this path is the 'language' of probability and the basic properties of probability— the topic of this chapter. Let's get started!

1.1 Deductive and Inductive logic

Broadly speaking, there are two directions in which we use logical reasoning— deduction and induction. Deductive reasoning means we first assume that some general principle is true and use logic to work out its particular consequences. For instance, we can start with Newton's inverse square law of gravitation as the general principle and work out one of its particular consequence— that the planets must go around the Sun in elliptical orbits.

Inductive reasoning goes in the opposite direction. You start with particular observations and then try to get at the general theory that gave rise to the observations. In our example, if you started with the measurements of the positions of the planets and use those to propose the general principle of the inverse square law of gravitation, then you are using inductive logic.

As a scientist, you will use deduction to test your theories/hypotheses. You do this by constructing a chain a logic going from one of more general hypotheses/postulates to a particular prediction that is a consequence of the hypotheses. You then test that prediction with experiment. If the experiments shows that the prediction is false then you have rejected the original hypothesis. Of course for this scheme to work properly you need to construct your logical chain of reasoning such that if the prediction is proven false then at least one of the general hypotheses/postulates is false. Let us understand this with an example.

- Postulate-1: Newton's gravitational theory is the correct theory of gravity.
- Postulate-2: The Sun and the known solar system planets are the only sources of gravitational force on Mercury.
- Prediction: The gravitational tug of solar system planets must cause Mercury's perihelion to precess at a rate of 8.85 arc minutes per century.

Let us say that the experimental observation is as follows: Mercury's perihelion precesses at a rate of 9.55 arc minutes per century. The experiment showed that the predicted precession was wrong. So you conclude that both of premises cannot be true. Either Newton's theory is not the correct theory of gravity and/or there is an undiscovered planet whose gravitational influence is causing an additional precession. In fact, this anomalous precession of Mercury's perihelion really puzzled astronomers for a long time. Le Verrier suggested that there must be an undiscovered planet inside the orbit of Mercury. The planet was never found. The second postulate actually turned out to be correct. It was the first which had to be abandoned

and the matter was laid to rest by Einstein's theory of gravity that predicts the correct value for Mercury's perihelion precession.

1.2 Plausible reasoning

Anyone who has done measurements knows that the there is always some error or uncertainty in the measurements. This can stops us from making definitive statements. Let us add a hypothetical twist to our example of Mercury's perihelion precession. What if the experimenter who measured the precession of mercury's perihelion told you that the error in her measurement was 0.25 arc minutes per century? Maybe both premises are correct and you happened to measure 8.85 arc-min/century instead of 9.55 arc-min/century purely due to measurement error. After all, the discrepancy of 0.7 arc-min/century is just 2.8 times the measurement uncertainty. How sure can we now be of rejecting at least one of the premises as false? 90 per-cent? 99 per-cent? Now we are engaging in what is called plausible reasoning; we are dealing in chances of something being true or false. The measurement error leaves no room for definitive statements. Your conclusions are necessarily probabilistic. You will learn how to assign probabilities to such conclusions in later chapters.

1.2.1 Estimation

As a scientist you are usually trying to measure something and comparing the measurement with a theoretical prediction. The parameter you measure could be, for example, the precession rate of Mercury's perihelion, the Hubble parameter, the masses of the neutrinos etc. That is you are *estimating* some physical parameter. And because you have incomplete information and because there are experimental measurement errors you must learn the craft of plausible reasoning. Let us start with a simple example.

What are the chances of getting 5 heads and 8 tails out of 13 tosses?

Most of you might have done such calculations when you learn permutations are combination. The answer is, or course ${}^{13}C_5 p^8(1-p)^5$ where p is the bias in the coin (specifically it is the probability of getting tails). If the coin is unbiased then $p = 0.5$ and the probability of getting the experimental outcome is 0.157 or around 16%.

OK that is reasonably straightforward. But let us say you want to use the experimental outcome to test the hypothesis that the coin is a fair one (i.e. $p = 50\%$) and your experiment shows that out of 13 tosses the coin gave 8 tails. How sure are you now that the coin is unbiased? As you will see, this is actually a harder problem to solve!

What is the bias of a coin that yields 8 tails out of 13 tosses?

Now it is tempting to say that the bias of the coin towards tails is $p = 8/13$. But is that an accurate statement? You just calculated that even an unbiased coin ($p = 0.5$) *can* give the outcome: 8 tails out of 13 tosses, 16 times out of 100. In fact unless $p = 0$ or $p = 1$ exactly, the observation (8 tails out of 13 tosses) *can* happen. That means any value of $p \in (0, 1)$ *is* possible. It is just that not every value of p in this range is equally likely. Because if p were very small then the odds of getting tails would be so low that it would be unlikely to get 8 tails out of 13. That means that plausible reasoning in this case is a matter of finding the probability of p being some value. In this course you will learn how to properly calculate that probability and use that to test hypotheses. But first, we must get used to the grammar and vocabulary of this new language of probability.

1.3 The language of probability

Before we start doing problems let us come to an agreement on how we will do our calculations. That is, let us agree on the names we give to concepts so that we can understand each other. Let us also give some rules to quantify our belief in something being true or the probability of something being true.

For simplicity let us say that probability always lies between 0 and 1. 0 means that we fully believe something is false and 1 means that we fully believe something is true and of course, values in between mean we are not totally sure one way or another but we do have some idea.

And now we are ready to make our first statement in this language: Prob(rain today) = 0.2. That is, we believe there is a 20% chance that it rains today. What is the chance that it does not rain today? That's easy: 80%. So we can say that Prob(no rain today) = 0.8. We notice here the axiom: $\text{Prob}(X) = 1 - \text{Prob}(\text{not } X)$. We will write 'not X' as $\bar{X}$. So we have

$$\text{Prob}(X) = 1 - \text{Prob}(\bar{X}).$$

Good! Let us now see if out new language can handle more complex situations. What is the probability that it will rain given that it is overcast. Notice what I did there, I made the probability *conditional* .... conditional on it being overcast. So now you have total belief that it is overcast (you see it with your own eyes) and then ask yourself, what is the probability that it rains. Let us choose to write this conditional probability as Prob(rain|overcast). You can now see that our first axiom works even with

conditional probabilities:

$$\text{Prob}(X|Y) = 1 - \text{Prob}(\bar{X}|Y).$$

Good, we are making progress in building up our new language's vocabulary and grammar.

Let us play around a bit more with our new language. We already said that $\text{Prob}(X|Y) = 1 - \text{Prob}(\bar{X}|Y)$. Can we likewise say $\text{Prob}(X|Y) = 1 - \text{Prob}(X|\bar{Y})$?

No! That statement is wrong! It helps sometimes to go back to concrete examples. That statement said $\text{Prob}(X|Y) + \text{Prob}(X|\bar{Y}) = 1$ which is like saying $\text{Prob}(\text{rain}|\text{overcast}) + \text{Prob}(\text{rain}|\text{not}-\text{overcast}) = 1$ which is clearly wrong. In fact $\text{Prob}(\text{rain}|\text{overcast}) + \text{Prob}(\text{rain}|\text{not}-\text{overcast}) = \text{Prob}(\text{rain})$. So it should be $\text{Prob}(X|Y) + \text{Prob}(X|\bar{Y}) = \text{Prob}(X)$. In fact you can generalize this as follows:

$$\text{Prob}(X) = \sum_{i=0}^{i=N-1} \text{Prob}(X|Y_i)\text{Prob}(Y_i); \text{ if } Y_i \text{ are independent and exhaustive}$$

Here independent means that each case Y_i is independent of the other so that there is no double counting (more on that below) and exhaustive means that the range $i \in [0, N-1]$ covers all possibilities. A special case of this of course is $Y_0 = Y, Y_1 = \bar{Y}$. This statement is called 'marginalization' of probabilities where you marginalize over the variable Y. You use marginalization when you want to calculate the probability of some variable X but it turns out that you only know the probability of X conditional on some other variable Y. For example, say NL is playing DE in the world cup final and you want to know the probability of an NL win. The problem is you only have a good idea of an NL win if the game does not go to penalties: $\text{Prob}(\text{NL}-\text{win}|\text{no}-\text{pen}) = 0.5$ and likewise if it went to penalties: $\text{Prob}(\text{NL}-\text{win}|\text{pen}) = 0.3$[1]. Because 'pen' and 'no-pen' are independent and exhaustive possibilities you can say $\text{Prob}(\text{NL}-\text{win}) = 0.3\text{Prob}(\text{pen}) + 0.5\text{Prob}(\text{no}-\text{pen})$.

1.4 Adding and multiplying probabilities

Let me invite you to a game of chance. Suppose I say "We will toss a coin once and roll a die once. If you get heads *or* you get a roll greater than 2 you win." What is the probability that you will win? OK so we have

[1]Lower because its a young side and may not handle the nerves well; not as bad as ENG but still.

Prob(head) = 0.5 and Prob(> 2) = 4/6. Because either outcome will let you win can we say that the probability of winning is the sum of the two?

No! That would be a huge mistake. In fact the sum exceeds 1 which is absurd. What went wrong? Well, what went wrong is that we have double counted! Let us write down all the possibilities. We have (H,1), (H,2), (H,3), (H,4), (H,5), (H,6), (T,1), (T,2), (T,3), (T,4), (T,5), and (T,6). Prob(head) is satisfied by the first 6 of these possibilities and Prob(> 2) is satisfied by (H,3), (H,4), (H,5), (H,6) and (T,3), (T,4), (T,5), (T,6). When you added the two probabilities you ended up double counting (H,3) to (H,6) because they appear in both lists. So you should have subtracted the double counted possibilities to get the right answer. So we should actually write Prob(head or > 2) = Prob(head) + Prob(> 2) − Prob(head and > 2). In our language this statement becomes Prob(A or B) = Prob(A) + Prob(B) − Prob(A and B). You can think of this graphically in the form of a Venn diagram.

How do we evaluate Prob(A and B)? Is it not just Prob(A) × Prob(B)? Well, in this case it is, but not in general as we will now see.

1.5 Algebra with conditional probabilities

So we have so far learnt how to do some simple algebra on probabilities (adding, multiplying) and also learnt the concept of conditional probabilities. What if we combined the two? Would not that be more fun? Let us again start with an example. But this time, the example will be more complicated. You see in the previous example, I deliberately choose two actions (rolling die and tossing coin) that were totally independent of one another. That is the outcome of one had no influence on the outcome of the other so there was nothing conditional in that problem. In fact if we wanted to use our conditional probability vocabulary in that problem, we would write Prob(head$|\ > 2$) = Prob(head). That is we would have the case Prob($X|Y$) = Prob(X), Prob($Y|X$) = Prob(Y).

We will now use an example to show that our statement Prob(X and Y) = Prob(X)Prob(Y) really should be Prob(X and Y) = Prob(X)Prob($Y|X$) = Prob(Y)Prob($X|Y$).

Suppose you have a friend who did not take this course, graduated and went on to crunch numbers for a flood insurance company. The company specializes in farm insurance. But they also give additional house insurance to farmers whose houses are on the farm. They are getting worried about what their modelers are saying about the flooding risk with new climate models. Apparently they expect a 1% chance that a farm floods in a given year. Maybe they can pay out the resulting insurance claims, but what if

the farms *and* farm-houses flood in such large numbers? That would be financial disaster. So they turn to your friend to calculate the probability that a farm and the farm-house flood.

Your friend says that easy: $\text{Prob}(\text{house} - \text{claim}\,\text{and}\,\text{farm} - \text{claim}) = \text{Prob}(\text{house} - \text{claim}) \times \text{Prob}(\text{farm} - \text{claim})$. The climate modelers tell him that $\text{Prob}(\text{farm} - \text{claim}) = 0.01$ and $\text{Prob}(\text{house} - \text{claim}) = 0.005$ (smaller because Dutch Farm houses are build on a raised mound!). So the probability of both claims coming in is 0.00005, small enough. All is well, your friend concludes. *Big mistake!*

What went wrong? Your friend did not take this course is what went wrong. You see, a farm flooding and a house on the farm flooding are not independent at all. If it is raining excessively and the house flooded, the farm has already flooded in all likelihood (as it is usually lower). Similarly if the farm floods then the chances of house also flooding has gone up massively. What your friend should have calculated was

$$\text{Prob}(\text{house} - \text{claim}\,\text{and}\,\text{farm} - \text{claim}) =$$

$$\text{Prob}(\text{house} - \text{claim}|\text{farm} - \text{claim})\text{Prob}(\text{farm} - \text{claim}).$$

He should have gone back to the modelers and asked for the conditional probability and averted financial disaster.

To avert disasters of this sort, let us etch the following into our minds

$$\text{Prob}(A\,\text{and}B) \equiv \text{Prob}(A, B) = \text{Prob}(A|B) \times \text{Prob}(B)$$

$$= \text{Prob}(B|A) \times \text{Prob}(A).$$

Finally, let us also recognise here that you can take a statement like the one above and make all quantities conditional upon the same new variable. For instance, in our flood insurance case, everything is conditional upon the climate models being correct. So we may write

$$\text{Prob}(A, B|I) = \text{Prob}(A|B, I) \times \text{Prob}(B|I) = \text{Prob}(B|A, I) \times \text{Prob}(A|I).$$

1.6 Bayes' theorem

You are now ready to state one of the most consequential theorems in probability, statistics, modeling, inference and a whole lot. It underpins a lot of modern science, finance, medicine, you name it! Notice that you had two equal expressions for $\text{Prob}(A\,\text{and}B)$ given by $\text{prob}(A|B)\text{Prob}(B)$ and $\text{Prob}(B|A)\text{Prob}(A)$? You rearrange the terms on the LHS and RHS and that's it, that's Bayes' theorem.

$$\text{Prob}(A|B) = \frac{\text{Prob}(B|A)\text{Prob(A)}}{\text{Prob}(B)}$$

"That's it?" I hear you saying, "I though Bayes' was a big deal!" It *is* a big deal and it is trivially simple to derive. Both things can be true at once. Actually the big deal is in realizing what each of the terms *really* means.

To see how amazing Bayes' theorem is, let us get back to our 8 tails out of 13 tosses example. We want to know the value of the coins bias, let us call it b with the understanding that $b = 0$ means a coin that always gives heads, and $b = 1$ is a coin that always gives tails and $b = 0.5$ is an unbiased coin. We want to find out what is the probability that the bias is some value b given the experiment we have just conducted (8 tails out of 13 tosses). We want to know Prob(bias $= b|E$) where E is the experimental outcome: 8 tails out of 13 tosses. Let us now writes Bayes' theorem for this problem of plausible reasoning.

$$\text{Prob}(b|E) = \frac{\text{Prob}(E|b)\text{Prob}(b)}{\text{Prob(E)}}$$

OK let us try to understand the terms on the RHS. Consider the term Prob$(E|b)$. Hey that's the easy bit— given bias b what is the probability of 8 tails out of 13 tosses? It is ${}^{13}C_8 b^8(1-b)^{13-8}$. What about Prob$(b)$? That is the probability that the bias is some value *regardless* of the outcome of the experiment you have done. That is your *prior* belief in the value of b even before you started doing the experiment. Let us say for now that you did not have any idea what b should be before the experiment. So you assign it a uniform probability: Prob$(b) = 1$; $b \in (0, 1)$. Now to the denominator where you find Prob(E). That is the probability of getting 8 tails out of 13 tosses with no knowledge of the bias b. Since every value of b is mutually exclusive of every other value you will have to add up the probability of the experimental outcome over all possible values of b, i.e. marginalize over b. In other words we have

$$\text{Prob}(E) = \int \text{dbProb}(E|b)\text{prob}(b)$$

Basically this term is just the integral of the numerator over all possible values of b. In fact it is there to normalize the RHS so that the probabilities come out between 0 and 1. That's it! For every value of b you can compute the RHS in a computer and get a probability distribution function for b. You have solved a parameter estimation problem.

Because Bayes' theorem formalizes the method to solve plausible reasoning problems, the different terms have obtained their own names over the years:

- Prob($b|E$) is the probability of the parameter you are estimating having some value given the experimental outcome. It is called the posterior.
- Prob($E|b$) is the probability that your experiment gave you the outcome you observed given some value of the parameter. This is called the likelihood.
- Prob(b) is your belief in the value of the parameter (from some prior knowledge/intuition) even before you did the experiment. This is called the prior.
- Prob(E) is just there to normalize the probabilities. It is the probability of our experiment giving the outcome you observed. This is called the evidence.

That is basically what I have for this chapter. But because out new language is such fun, let us get a bit more practice with some examples.

Consider the following statements:

- A: Maya brought an umbrella to work.
- B: It will rain today.

Suppose that you live in a pretty dry place, say southern California, where it rains on only 1% of the days. You also notice that whenever it rains, Maya has an umbrella 90% of the time. She is a very conscientious girl! Now suppose that you see Maya walk into work with an umbrella. What is the probability that it will rain today?

Your first instinct is to guess that it is highly likely to rain today. After all you have noticed that Maya walking in with an umbrella goes hand-in-hand with it raining. Correct?

Let us do the math. You want to evaluate Prob($B|A$). From Bayes' theorem you know that

$$\text{Prob}(B|A) = \frac{\text{Prob}(A|B)\text{Prob}(B)}{\text{Prob}A} = 0.9 \times 0.1/\text{Prob}(A) = 0.09/\text{Prob}(A)$$

Aha! You now see that you are missing a key piece of information: Prob(A). How often does Maya bring an umbrella to work, rain or shine? Suppose she brings an umbrella to work 90% of the time, rain or not. Then you have Prob($B|A$) = 0.09/0.9 = 0.01, just 1%, so not likely at all.

You may think this is a banal example. But it is precisely what a lot of people, including myself, struggled to get their head around during the COVID pandemic and why Bayes' theorem should be mandatory learning. Let me explain with another example.

Suppose it is the height of the pandemic and you want to attend the wedding of a childhood friend. You take a COVID test at home before the wedding just to be sure and yikes! It shows a positive result! You have taken this course so you start doing the numbers with a cool head. What are the odds that you actually have COVID? That must depends on how good the test is. The COVID test information sheet says that it has a sensitivity of 95% and specificity of 98%. What on Earth are these numbers? You dig more online and find the following definitions: Sensitivity is the true positive rate; that is the probability that the test yields a positive result if you actually have the disease. Specificity is the true negative rate; that is the probability that the test yields a negative result if you do not have the disease. OK you write down two statements:

- A: I have COVID
- B: The COVID test yields a positive result

You now want to calculate $\text{Prob}(A|B)$. The information sheet told you that $\text{Prob}(B|A) = 0.95$ and $\text{Prob}(\bar{B}|\bar{A}) = 0.98$. Bayes' theorem says

$$\text{Prob}(A|B) = \frac{\text{Prob}(B|A)\text{Prob}(A)}{\text{Prob}(B)} = 0.95\text{Prob}(A)/\text{Prob}(B)$$

What is $\text{Prob}(A)$? It is the chance that you have COVID regardless of the test. That must be the current incidence of COVID in your area. You go on the government's information portal and find that one in a thousand people in your area have COVID. So $\text{Prob}(A) = 0.001$. What is $\text{Prob}(B)$? It is the chance that the test yields a positive result regardless of whether you have COVID. So now you need to marginalize on the two independent and exhaustive possibilities: you have COVID and you do not have COVID. So you write

$$\text{Prob}(B) = \text{Prob}(B|A)\text{Prob}(A) + \text{Prob}(B|\bar{A})\text{Prob}(\bar{A})$$

$$\text{Prob}(B) = \text{Prob}(B|A)\text{Prob}(A) + [1 - \text{Prob}(\bar{B}|\bar{A})][1 - \text{Prob}(A)] = 0.021$$

So now you have everything you need:

$$\text{Prob}(A|B) = 0.95 \times 0.001/0.021 = 0.0454$$

There is less than 5% chance that you have COVID despite the positive test result! Wait... what? The test is quite sensitive: 95% in fact. What

is going on here? I will let you struggle with this conundrum. And when it finally dawns on you, you would have truly understood Bayes' theorem and the virtues of a Bayesian view of the world.

1.6.1 Frequentist and Bayesian world views

The COVID example above illustrates the difference between what is called a 'frequentist' view of the world and a 'Bayesian' view of the world. A frequentist view assigns probabilities to 'data' whereas a Bayesian view assigns probabilities to 'hypotheses'. You see, in the frequentist view, the data takes center stage. Here the data is 'a positive COVID test result' and the frequentist says what are the odds that I observed this data if I had COVID and the answer is of course the sensitivity of the test, so 95%, and a hasty frequentist may wrongly conclude that there must be an 95% chance that I have the disease. But as we saw, while this may be what we all would do intuitively, it is not the correct way to view the situation. The correct way is to always start with a hypothesis that you wish to test with the data and ask the correct question: 'what is the probability that my hypothesis is true given the data I have got.' It is *very* important to ask the right question!

1.7 Summary

- Deductive logic: General postulates to particular consequences. Used to make predictions of a theory for experimental test.
- Inductive logic: Particular observations to general postulates/theory. Used to construct new theories or hypotheses.
- Plausible reasoning: Used to assign a measure of belief or probability to statements when you have incomplete information about the world.
- Properties of probability:
  - $0 \leq \text{Prob} \leq 1$
  - $\text{Prob}(X) + \text{Prob}(\bar{X}) = 1$
  - $\text{Prob}(X \text{ or } Y) = \text{Prob}(X) + \text{Prob}(Y) - \text{Prob}(X \text{ and } Y)$
  - $\text{Prob}(X \text{ and } Y) = \text{Prob}(X|Y)\text{Prob}(Y) = \text{Prob}(Y|X)\text{Prob}(X)$
  - $\text{Prob}(X|Y) = \text{Prob}(Y|X)\text{Prob}(X)/\text{Prob}(Y)$ (Bayes' theorem)
  - $\text{Prob}(X) = \sum_i \text{Prob}(X|Y_i)\text{Prob}(Y_i)$ for Y_i independent and exhaustive (marginalization)

- Be very careful with 'frequentist' ways of doing things. It can lead you down the wrong way. Always calculate the probability of a hypothesis given the data, a.k.a the Bayesian way.

Chapter 2

Parameter estimation

Much of life's decision making involves estimating the value of some parameter. A scientist may estimate the mass of an neutron, a stock broker may estimate the value of some stock, a restaurant owner must estimate the number of customer she expects on a given evening, a football stadium must estimate the number of fans they expect to buy tickets at their price point and so on. Most people go with gut-instinct of previous experience in estimating values of parameters. But you need to be much more rigorous when you are in uncharted territory or when a mistake can prove too costly to bear. In this chapter, we will learn the basics of parameter estimation using simple examples. We will learn how to not just estimate the value of a parameter buy also how to assign a measure of confidence to your estimation.

2.1 Posterior distribution

Remember the coin toss example from Chapter 1? We wanted to estimate the bias of a coin, b using data from an experiment (8 tails out of 13 tosses). We already learnt that the answer is not deterministic but rather probabilistic. Let us explore that ideal further.

We wrote en expression for the probability of b having some value using Bayes' theorem as

$$\text{Prob}(b|E) = \frac{\text{Prob}(E|b)\text{Prob}(b)}{\text{Prob(E)}}$$

Figure 2.1 is the output of a program that does the calculation for the posterior probability on a computer.

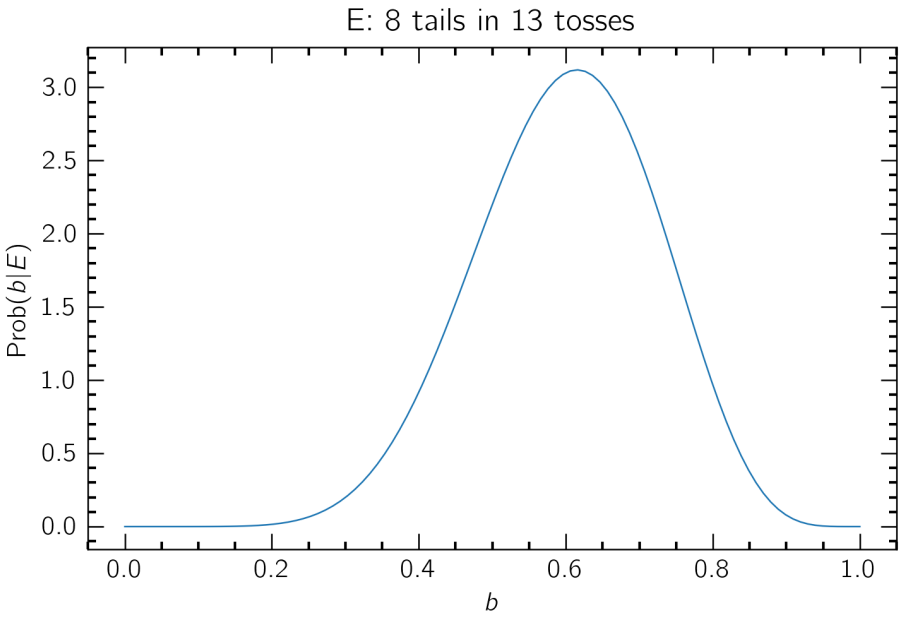


Figure 2.1: Posterior distribution of the bias of a coin (towards tails) that gave 8 tails out of 13 tosses.

Let us try to understand what this posterior probability curve is telling us. It is solving an estimation problem by telling us something about the value of the parameter b given the knowledge/data we have, E. Obviously, the higher the y-axis value of the curve, the more probable the x-axis value is. You can see that the curve has a peak around a bias of 0.6 which makes sense because we obtained more tails than heads in the experiment

so the coin is likely to be biased towards tails. But the curve is also quite broad which tells us that the bias towards tails is not established with a high degree of certainty. In fact, the y-axis value of the curve at $b = 0.5$ (unbiased coin) is also quite high. So perhaps we cannot rule out the possibility that the coin is biased just yet. There is a limit to what we can asset with the available experiment. OK so we have learnt the first important aspect of posterior curves: that the location of the peak and the width of the curve around the peak tell us about the likely value of the parameter and its uncertainty, respectively.

What if we ran the experiment much longer? What if say we did 100 tosses and got 63 tails? Figure 2.2 is what the posterior curve would look like. Now we see that the curve is more narrowly peaked around 0.6. This

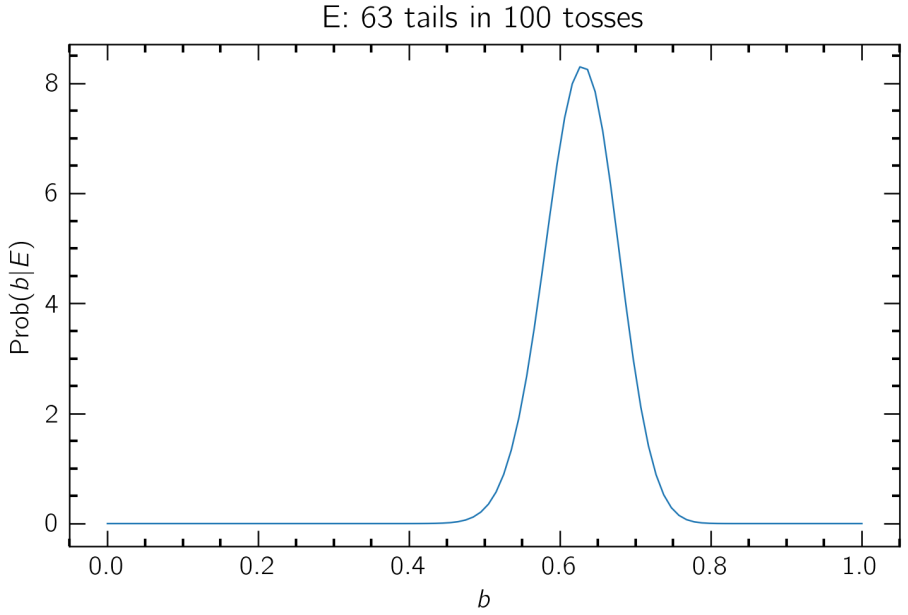


Figure 2.2: Same as Fig. 2.1 but for the case of 63 tails in 100 tosses.

makes sense as it was possible for an unbiased coin to give you 8 tails in 13 tosses (instead of 6 or 7) just out of random chance, but is very unlikely for an unbiased coin to keep giving an excess of tails again and again for 100 tosses just out of random chance. OK we have learnt another aspect of posterior curves: the more data we have the more certain we become of our estimate.

So far we have only made qualitative statements, which have very limited use. Can we be more quantitative about our statements on the likely

value of the parameter— let us call this the central value, and our measure of uncertainty of this value— let us call this the dispersion or just uncertainty?

2.1.1 Probability density functions

In order to make quantitative statements about the central value and uncertainty, we will have to condense all the information in the probability curve into a few key numbers. Before we learn to do that, let us first appreciate a few properties of the probability curve.

The curve such as the ones in Figure 2.1 and Figure 2.2 are called probability density functions, or PDFs. They are called *density* functions because they show us the probability of a continuous variable. In this case, the variable b, can take any value on the real number line between 0 and 1. In fact, the number of possible values of b is infinitely large so it makes no sense to assign each value of b a probability. The only thing that makes sense is to assign a probability for a finite range of values of b. The value of the function PDF(b) at some value b_0 should therefore be interpreted as the probability that b lies within an infinitesimal range db around the value b_0. With this definition, the probability that the value of b lies in the interval (b_1, b_2) is simply the integral

$$\text{Prob}(b_1 < b < b_2) = \int_{b_1}^{b_2} \mathrm{d}b \ \text{PDF}(b)$$

And because b must lie between 0 and 1, the integral for $b_1 = 0$ and $b_2 = 1$ must be equal to unity.

2.1.2 Measure of central value, dispersion and confidence interval

Suppose, someone wants to know the answer to the question: 'what is the probability that the coin is biased towards tails?' You may show this person the curve in the figure but will be met with "Cute plot. Is this coin biased towards tails or not?"

Let us work the numbers out. The question really is asking, 'what is the probability that b lies in the interval [0.5, 1]. That is simply

$$\text{Prob}(\text{tails} - \text{bias} = \int_{0.5}^{1} \mathrm{d}b \ \text{PDF}(b)$$

For the case of the curve in Figure 2.1 this works out to be 0.9954. So your reply is 'I am more than 99% certain that this coin is biased towards tails.' For the curve in Figure 2.1 this value works out to be only 0.78801. So with

just 13 tosses we can only be around 79% certain that the coin is biased towards tails. What we just did is formally called assigning 'confidence levels' to our claim that the coin is biased towards tails.

The next question you are likely to get is "I had a hunch it was biased. But by how much is it biased?" You now have to find the central or most representative value. You could just pick out the peak of the curve. This is called the 'most likely' value.

Or, you could calculate the 'average' value of the bias— analogous to the average of a population. When you calculate the average of a bunch of numbers, you add up all the numbers and divide by the size of the population. The same can be done with PDFs. If we divided the PDF curve into N equal intervals between (b_0, b_1), (b_1, b_2) and so on till (b_{n-1}, b_n) and assign the constant values of $\mathrm{PDF}(b_0)$, $\mathrm{PDF}(b_1)$ and so on until $PDF(b_{n-1})$ to the value of the PDF within these intervals, then we can calculate the average value of b as

$$\mathrm{avg}(b) = \frac{1}{N} \sum_i \mathrm{PDF}(b_i).$$

You can see that the above equation is just the way to take the weighted average of a bunch of numbers. In this case the weights are given by the probability.

The above equation can be written in integral form and the resulting value is called the 'expected value' of b.

$$\mathrm{avg}(b) \equiv \langle v \rangle \equiv \int \mathrm{d}b\, \mathrm{PDF}(b)$$

As an interesting aside, a common way to specify the properties of the function $\mathrm{PDF}(b)$ is the value of the so-called moments of the function. The n^{th} moment of the function is defined as

$$M^n(\mathrm{PDF}) \equiv \langle b^n \rangle \equiv \int \mathrm{d}b\, b^n \mathrm{PDF}(b)$$

It is clear that the 0^{th} moment of any probability density function must be equal to 1. The first moment is the mean or expected value of the parameter that the function describes. We will get back to what the higher moments mean in a while. Let us, for now, get back to our discussion of the central value.

The most likely and expected value are good enough for probability curves that have a nice symmetric peak. Sometimes we are dealing with curves that are very lopsided, that is, they run down the peak abruptly on one side and very gradually on the other. A commonplace example of this

is the PDF of incomes of households as shown in Fig. 2.3. As you can see there is so much area under the curve right of the peak that choosing the likley value or even the expected value may give a misleading impression on the central value of the income. In other words, if there are a small number of extraordinarily rich people, then they will drag the mean value upwards but that does not make me feel rich! This is why you almost never hear

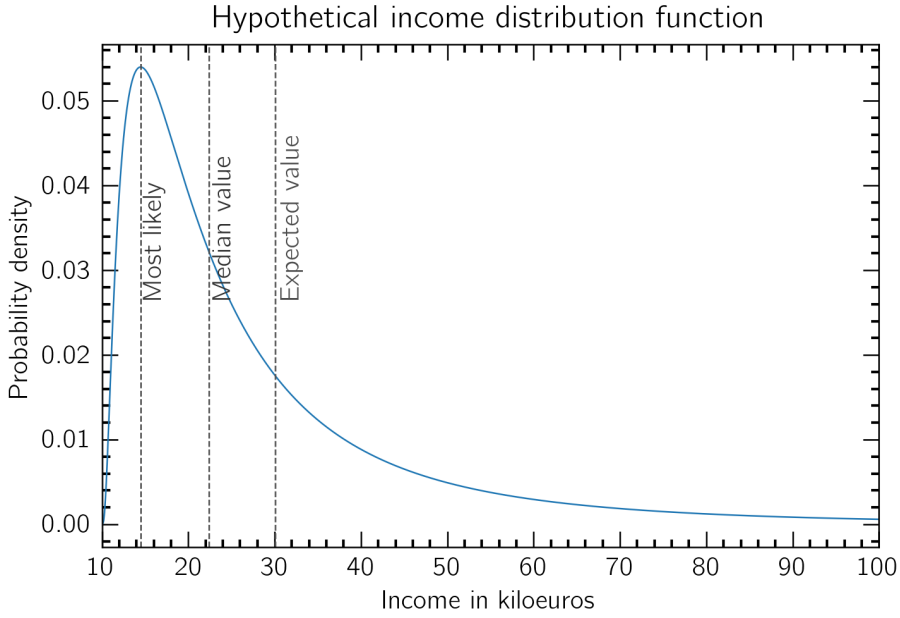


Figure 2.3: PDF of a hypothetical income distribution with the likely, mean and median values indicated.

the terms 'most likely' income or 'expected' or 'mean' income on the news. What do you usually hear? It is the *median* income. Median is simply the 50^{th} percentile by which we mean that half of the people have income above the median (and half below, of course). For a PDF, we mean to say that half of the area under the curve is above the median (and half is below).

It makes sense to use median in this case not only because the curve is lopsided but also because humans subjectively evaluate their financial position in comparison to the *number* of people around them who are richer or poorer than them and not based on the number of people weighted by their respective incomes. It is not enough in life to do the numbers properly, the *choice* of what number you evaluate is equally important. You must pay close attention to the choice of statistic someone is quoting, especially in advertising and politics where people have an axe to grind!

OK, let us return to our interlocutor who is asking us questions about the bias. Suppose the next question is"What is the range within which you are fairly certain that the bias lies?" We would then evaluate so-called 'confidence intervals'. This is usually the interval around the most likely value of b within which we are say 99% certain that the value of b lies. What is the most likely value of b? Well suppose we take that to be the location of the peak— the most likely value. In case of the curve in Figure 2.2, this value is about 0.63. Now let us find a range of b centred around this value such that the integral of the curve in this range is 0.99. This would give us a range within which we are 99% sure that the value of b lies. For the case of the curve in Fig. 2.2 this range works out to be between 0.506507 and 0.752753. We call this the 99% confidence interval for the parameter b.

The confidence interval tells you something around the uncertainty in the estimate of b's value. Another common way to quote this uncertainty is via the width of the curve. This is because the sharper the peak in the curve, more our certainty in the estimate of b. How to measure this width? The crudest way is to find the x-axis values at which the curve reaches half of its peak value and call the interval between the two points as the width. A better way to calculate the width of the 'peakedness' of the curve is by computing the standard deviation of b. You might remember standard deviation from high school statistics as a measure of dispersion in data values. Just like we defined the mean value above as a weighted average, we can write the standard deviation to be

$$\text{std}(b) = \frac{1}{N}\sqrt{\sum_i [b - \text{avg}(b)]^2 \ \text{PDF}(b)}$$

which in integral form is

$$\text{std}(b) \equiv \sigma(b) \equiv \sqrt{\int \text{d}b \ [b - \langle b \rangle]^2 \ \text{PDF}(b)}$$

Hey, that looks a bit like the expression for the second moment of b expect for that $-\text{avg}(b)$ term. Indeed, the above integral is called the second *central* moment. Finally, the square of the standard deviation is called the variance.

Let us now summarize all the ways we can quantify the PDF function of a parameter we have just estimated.

- **Central value measures**:
  - Most likely value — location of the peak of the PDF.
  - Expected value — First moment.

  - Median value — value above which the area under the curve is half.

- **Dispersion**:

  - Confidence level— Fraction of the area under the PDF that lies in a given range.
  - X% Confidence interval — Interval around the central value within which X% of the area under the PDF lies.
  - Variance — second central moment of the PDF.
  - standard deviation — square root of the variance.

Let us end with a list of common ways to crunch down a posterior PDF of a parameter estimate into simple numbers/statements.

- "We place the constraint $b > 0.5$ with a confidence of 99.54%."
- "The most likely value of b is 0.6296."
- "The expected value and standard deviation of b are 0.627 and 0.002 respectively. "
- "We estimate the bias to be $b = 0.627 \pm 0.002$."
- "We estimate the bias to be $b = 0.627(2)$."
- "The 99% confidence interval of b is $[0.5065, 0.7527]$."
- "Here is a machine readable table of the PDF of b. Knock yourself out!" Please don't do this! Provide the table, sure, but definitely make a statement such as the ones above depending on what the exact question you wanted to answer was.

2.2 Designing an estimation experiment

OK that was a lot of theory and definitions. Maybe a bit of a let-down after I told you that parameter estimation is the essence of living an examined life. Let us study a couple of realistic, yet, hypothetical examples so you can appreciate the value of parameter estimation.

2.2.1 Example: A new intervention

Suppose you are meeting up with a friend who works as a doctor at a large hospital. She is very excited about a new medical intervention for some scary-sounding illness. The standard intervention for the illness has been practiced for many decades but it only has a 87% chance of success. Your friend has tried the new intervention on 10 volunteer patients and they all recovered! It is a 100% success rate! The old procedure had 87% odds of success, so out of the 10 volunteer patients, on average, 1 or 2 would not be cured of the ailment. The new procedure is clearly better, she asserts. What would your reply be to her?

Let us look at this as an estimation problem. You are trying to estimate the success rate of an intervention, say b. The outcome is binary: either the patents are cured (success) or not (failure). b must lie between 0 (intervention always fails) and 1 (intervention always succeeds). The situation is mathematically identical to estimating the bias of a coin. You know how to do that!

Let us write down the posterior of b, the probability of the new procedure succeeding in curing the illness given the volunteer experiment, E.

$$prob(b|E) = \frac{\text{Prob}(E|b)\text{Prob}(b)}{\text{Prob}(E)}$$

You know from the coin-toss example that $\text{Prob}(E|b) =^{13} C_0 b^{13} = b^{13}$. What is the prior, $\text{Prob}(b)$?. Well it is a new intervention, so you have no prior knowledge of b which means, the only choice is to assign a uniform distribution as the prior:

$$\text{Prob}(b) = 1\, b \in (0, 1];\ = 0, \text{ otherwise}$$

Remember that the denominator, the evidence is just the integral of the numerator taken over the variable b. You code all of this in a computer and get the black curve in Fig. 2.4

You now have to calculate the probability that the new intervention is an improvement over the standard, that is $\text{Prob}(b > 0.87)$. You just need to integrate the curve between 0.87 and 1 to get this value. It turns out to be 0.79. So there is a 79% chance that your friend's hunch is correct and a 21% chance that it is incorrect. These are not great odds to change the whole treatment protocol. You really need to be more sure. Your friend agrees. In fact you both agree that you really need to know if the new protocol is better with 99% confidence before changes can be made. But how to improve this confidence?

One way is obvious from the coin-toss example: sign up more patients. Your friend says that she could get 40 more volunteers. Would that be

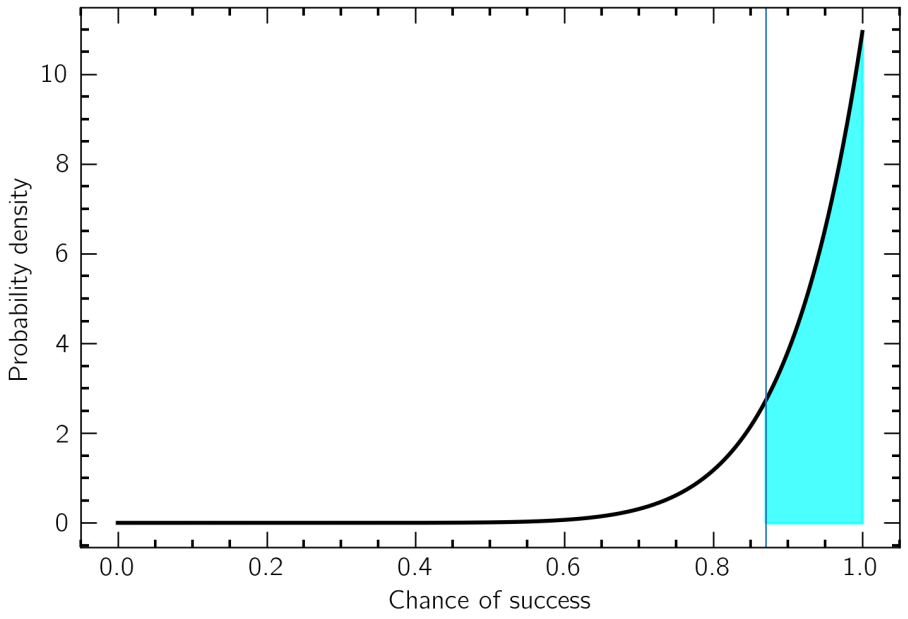


Figure 2.4: Posterior distribution of the chance of success of the new intervention (black) and the area under the curve where the new intervention is better than the standard intervention (cyan).

enough? Let us see. Suppose 38 of the new volunteers recover using the new intervention; that makes it 48 out of 50. The new posterior is shown in Fig. 2.5 and the area under the curve above 0.87 is 97%.

Nice! But you have not done the experiment yet. What if only 47 recover. What then? Well you already have the code, so you can give your friend a table of what she can expect with a total of 50 patients.

Number that recover	*Confidence that intervention is better*
50	99.9%
49	99.3%
48	97%
47	91.4%
46	81.3%
45	67%

As you can see, the confidence depends on the outcome of the experiment. This is the because the area under the posterior curve to the right of 0.87 depends on what the *real* chance of success of the new intervention

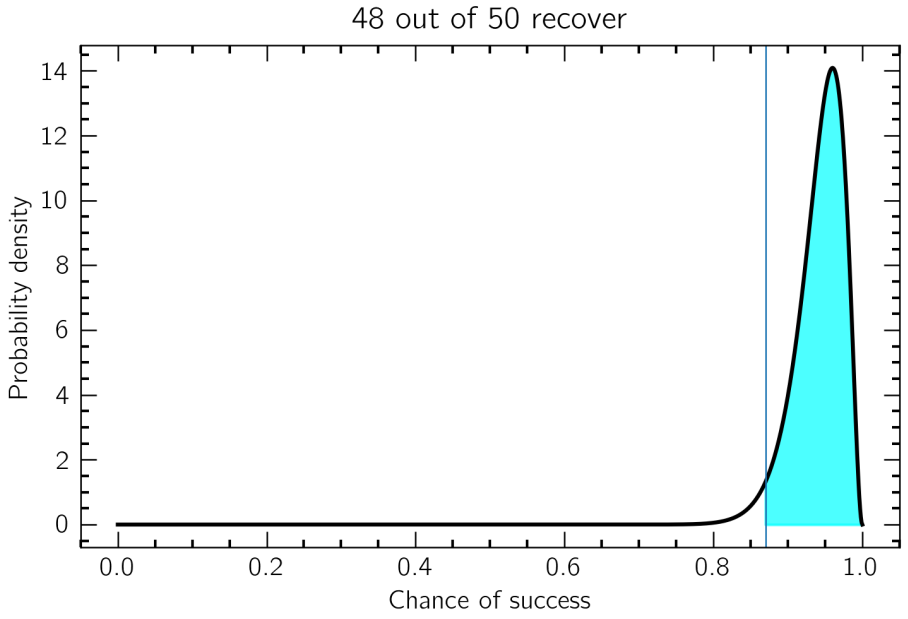


Figure 2.5: Same as Fig. 2.4 but assuming 48 out of 50 recover.

is. If the real chance of success if close to that of the old intervention, it will take many more trials to statistical separate their potency in curing the illness. This makes sense because to separate their potency values, the width of the posterior much be much narrower which means 50 trials may not be sufficient.

There is an interesting aside here. We have assumed that we know the success rate of the old intervention with zero uncertainty. This is cant be right. The uncertainty may be small because the estimate is based on a lot of historical data but not zero. Suppose, the posterior for the old intervention was known. How would you now estimate the odds of the new intervention being better?

2.2.2 Example: size of a population

Protracted wars can be won and lost on intelligence and industrial production capabilities. This was on the mind of the Allied war cabinet during WW2 when they faced a hard-to-answer question. How many tanks were the Germans producing per month? Of course, wartime secrecy meant they could not just ask; they had to be clever. British and American intelligence folks were quoting scarily large numbers! Yikes!

The statisticians came up with an ingenious solution to this estimation problem. You see, every time a German tank was captured in battle, the allied soldiers could read off the serial number on the gearbox of the tank. By looking at the serial numbers of the captured tank, perhaps they could estimate the maximum serial number, N?

Well one thing is obvious, N must be larger than the maximum serial number of all the captured tanks. But how much larger? Let us assume that each tank is equally likely to be captured. If say you captured three tanks and they had serial numbers: 2, 4, 7 then you know that N cant be very large. Because, if N were 1000 as the intelligence folks claim, why are all the captured serial numbers so small? So you *can* say something more than the obvious $N \geq 7$ after all.

If you do not like the examples involving depressing topics such as wars, don't worry. Mathematically the problem is the same as asking how many taxis are there in a city given a list of serial numbers you saw on the streets. How many phones are produced by some company? I took the example of WW2 to impress upon you what a difference knowledge of statistics can make.

Ok, here is the mathematical statement of the problem: Suppose k numbers are drawn at random from the set 1,2,3...N. If the maximum of the k numbers is n then estimate the PDF of N.

Of course we will use our favourite Bayes' theorem. Let us start with the prior. The only information you have is $N \geq n$ and say you choose to ignore the intelligence-types and say N can be arbitrarily large. So you have

$$\text{Prob}(N) = (W - n)^{-1};\ N \geq n$$

and 0 otherwise. Here W is some large number (we will get back to the choice of W later).

What is the likelihood? It is the probability, given N, of finding the maximum number n in k random draws. The number of all possible draws is ${}^{N}C_{k}$. Of these, how many satisfy the observation? The number n must be in the draws that satisfy the observations. So we are only left with choosing $k-1$ numbers out of the $n-1$ numbers that are smaller than n. This can be done in ${}^{(n-1)}C_{(k-1)}$ ways. OK so we have the likelihood:

$$\text{Prob}(N|E) = \frac{{}^{(n-1)}C_{(k-1)}}{{}^{N}C_{k}} = \frac{k^2(n-1)!(N-k)!}{(n-k)!N!}$$

which is

$$\text{Prob}(N|E) = \frac{k(n-1)(n-2)....(n-k+1)}{N(N-1)(N-2)....(N-k+1)}$$

The evidence is just the summation of the numerator over all possible values of N. So we have everything we need to calculate the posterior. I have done this on a computer for the following case: $k = 3$ and the serial numbers are $43, 87, 112$, so $n = 112$. The posterior is shown in the top panel of Fig. 2.6.

The 50^{th} and 90^{th} percentile values work out to be 159 and 351 respectively. If say the peak serial number were the same but we had $k = 20$ captured tanks, then the posterior is given in the bottom panel of the figure. Now we have the 50^{th} and 90^{th} percentile values of 117 and 127. Even with just 3 tanks captured, it is amazing how well the total number of tanks is bounded.

What did I choose for W? This is hard to know apriori. You see that the distribution converges to zero as N is increased so you only have to choose W such that you capture most of the area under the curve. I experimented with increasingly larger values of W and settled when increasing W further did not change the percentile values by more than unity.

This story wont be complete if I did not tell you what really happened during WW2. Obviously, I have greatly simplified matters here and taken hypothetical values for k and n. But the proper analysis was done by the Allies' statisticians. Here are three median estimates for three different months of the war[1]: 169, 244, and 327 tanks. After the war, with access to German records, the real numbers were revealed to the Allies. For these three months, they were: 122, 271, 342. And what were the best intelligence estimates during the war, you may ask? 1000, 1550 and 1550 respectively. If this is not a mic-drop moment for a statistician, what else is?

[1]I have taken these figures from *An Empirical Approach to Economic Intelligence in World War II* by Ruggles and Brodie, published in Journal of the American Statistical Association, 42(237), pp. 72–91.

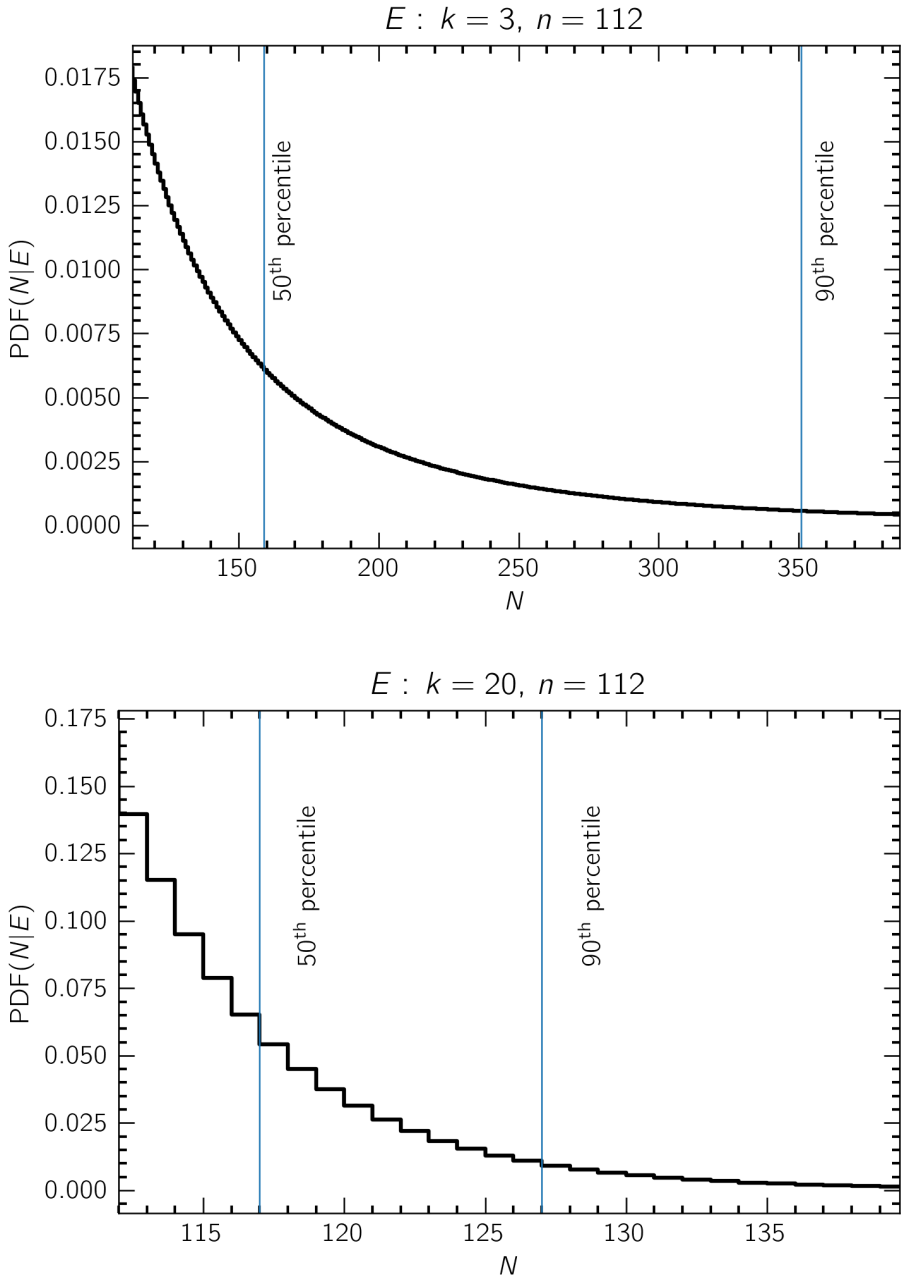


Figure 2.6: Posterior distribution of the number of tanks in the German tank problem. Top panel is for the case of $k = 3$, $n = 112$ and the bottom panel is for the case of $k = 20$, $n = 112$. Note the different x-axis extents of the two plots.

Chapter 3

Probability distribution functions

In Chapter 2 you learnt how to estimate the PDF of some parameter whose value you are interested in. You learnt this via examples. But did you notice that in all the examples, we always used the binomial distribution to calculate the probabilities? There are many real-life situations where the binomial distribution is impractical to calculate, or is just not the right distribution to use. In this chapter we will learn about new distributions that you are likely to encounter. You also saw in previous chapters that different situations that arise in life have the same underlying mathematics. This means that there are a small number of distributions that must capture most of the situations that arise. The aim of this chapter is to study the properties of these common or famous distributions and get comfortable with the algebra of transforming one distribution into another.

3.1 The law of small numbers

Suppose you decide to start a new website to generate some advertisement revenue. Obviously, you care very much about the number of people visiting the site per day. You collect the numbers over a period of time and make a histogram that looks like Figure 3.1. There are good days with 130 hits and

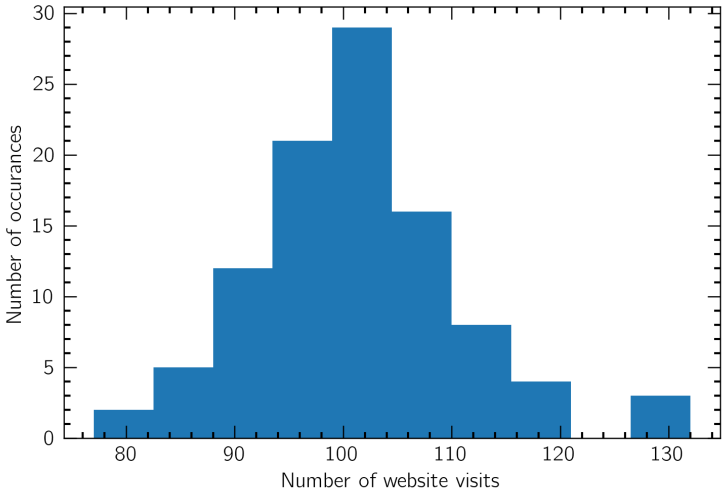


Figure 3.1: Histogram of number of visitors to the website per day

some bad days with just 80. Is this because of something you are doing, or your competition is doing? Or are these random statistical fluctuations?

This is exactly the same sort of question asked by a restaurant owner about the number of customers, a cellphone company about the number of phones sold, an insurance company about the number of claims filed, a factory manager about the number of defective items produced and so on. The list is endless. It is also timeless; the same sort of question asked by Bortkiewicz, a Polish-Russian statistician in the 19^{th} century. Bortkiewicz was looking at Prussian cavalry records of soldiers who had died by horse-kick. It was rare enough, sure, but some cavalry units had more deaths than others. Was this a random fluctuation or lack of discipline in these units?

Let us return to your new website and see if we can use what we know, the binomial distribution, to solve this problem. How to represent a 'visit' to the website with a binomial experiment that can only yield true or false outcomes? Well, we could divide up time into bins and ask whether there was a visit in each time bin. That would make it akin to a coin toss: visit=heads, no-visit=tails. The problem then transforms to the probability

of getting k heads in n tosses. What if there were two or more visits in a time bin? The binomial cannot handle that by definition. OK, then let us break down time into bins so fine that there is negligible chance of having two visits; its either zero (most of the time) or one (rarely). If p is the probability of a visit in a small bin δt then the probability of k visits in time $t = N\delta t$ is

$$\text{Prob}(p, N, k) = {}^{N}C_k \, p^k \, (1-p)^{N-k}$$

Suppose $\delta t = 1\,\text{second}$. Over a period of a day, $N = 86400$. And because there are on average 100 visits per day, p should be around $100/86400 \sim 10^{-3}$. We are in a regime where N Is very large and p is very small. In this regime, it is impractical to calculate with the above expression. Why? Because try calculating 86400! and you will know why!

Surely there must be a way to reduce the expression on pen-and-paper in the limit of large N and small p for circumstances like this before rushing to write a code? This is what French mathematician Siméon Denis Poisson did and the resulting distribution bears his name.

Let us expand ${}^{N}C_r$. We will also set $\lambda = pn$ which is the average number of visits, a.k.a the rate parameter

$$\text{Prob}(\lambda, N, k) = \frac{N(N-1)(N-2)...(N-k+1)}{k!} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-k}$$

$$\text{Prob}(\lambda, N, k) = \frac{1(1-1/N)(1-2/N)....(1-(k-1)/N)}{k!} \lambda^k \left(1 - \frac{\lambda}{N}\right)^{N-k}$$

Now let us take the limit as $N \to \infty$ of the different terms. In this limit, we have

$$\lim_{N\to\infty} \left(1 - \frac{\lambda}{N}\right)^{N-k} = \frac{\lim_{N\to\infty} \left(1 - \frac{\lambda}{N}\right)^{N}}{\lim_{N\to\infty} \left(1 - \frac{\lambda}{N}\right)^{k}} = \lim_{N\to\infty} \left(1 - \frac{\lambda}{N}\right)^{N},$$

and

$$\lim_{N\to\infty} 1(1-1/N)(1-2/N)....(1-(k-1)/N) = 1$$

So we get

$$\lim_{N\to\infty} \text{Prob}(p, N, k) = \frac{\lambda^k}{k!} \lim_{N\to\infty} \left(1 - \frac{\lambda}{N}\right)^{N}$$

Finally, we use the definition of the exponential function, or more precisely, the definition of Euler's constant e:

$$e^x = \lim_{n\to\infty} (1 + x/n)^n$$

to get the final expression for the Poisson distribution:

$$\text{Prob}(\lambda, k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Let us, henceforth, denote the Poisson distribution function with $\mathcal{P}$. So in the website example, suppose the average number of visits per day was $\lambda = Np = 100$, then the probability of getting 80 or fewer visits per day is visits per day is

$$\Sigma_{k=1}^{k=80}\mathcal{P}(100, k) = 0.023$$

So at least 2 in 100 days could have visits less than or equal to 80. So it is not unusual that you have a few days out of the 100 where you had 80 or fewer hits. There is nothing weird going on with your website or your competitors, these are just statistical fluctuations!

As an aside, note that I did not simply calculate $\mathcal{P}(80, k)$ because there is nothing special about 80 visits per day. You would have asked the same question if it were 81 visits or 79 visits. Your question of 'Is the low number of visits a statistical anomaly?' requires us to calculate the probability for a 'range of values' for the visits which you would have considered 'low' and asked the question in the first place.

Let us take a step back and state the general scenario where the Poisson distribution is the right model to use. We must use the distribution when

- Evaluating the probability of a discrete number of 'events' occurring when
- Each event occurs independently of the others, AND
- Two events cannot occur at the same time. In other words in a sufficiently small time interval, either one event occurs or none at all (this is the limit of large N and small p.)

In astronomy, the Poisson distribution appears in many circumstances. Here are some examples

- You are counting the number of sources of a particular type to understand how common they are.
- You are counting the number of photons received from a source to understand how luminous the source is.
- You discovered a new type of source ($k = 1$) and you want to know how many you will discover if you did a larger survey to find more.

The last example really shows the power of the Poisson distribution in modelling small numbers, just as Bortkiewicz did with the horse-kicks. Suppose you have discovered the new phenomenon of Fast Radio Bursts (FRB). You found humanities first FRB by surveying a patch of the sky that is $1\,\mathrm{deg}^2$ for 100 hours. You are now planning a survey with the same sensitivity and time-on-sky but with a much larger field of view, say $10\,\mathrm{deg}^2$. How many FRB do you expect to detect?

FRBs sources don't co-ordinate with one another, so the events appear independently of one another and the event rate is small enough that you can easily define a time-interval in which the odds of two or more FRBs arriving are negligible. We can use Poisson's statistics here. Good.

Our first instinct is to say that we expect 10 FRBs as it is 10 times the survey footprint (in deg^2-hours). Well maybe that is the mean or expected value, but remember that you are trying to *estimate* the value of λ with just one data point, so you better be more careful and calculate the posterior distribution of λ just like you computed the posterior distribution of the bias of a coin by looking at the outcome of coin-tosses.

Let is define μ as the number of FRBs above your telescope's detection threshold that occur per deg^2 per hour. Our experiment E is '1 FRB in 100 deg^2-hour'. The posterior is (Bayes' theorem)

$$\mathrm{Prob}(\mu|E) = \frac{\mathrm{Prob}(E|\mu)\mathrm{Prob}(\mu)}{\mathrm{Prob}(E)}$$

What shall we choose for the prior? Well, we have no idea that FRBs were even a thing before the experiment, so we must choose a 'non-informative' prior as in a prior that does not introduce any new information on the value of μ. I have not yet taught you how to choose a 'non-informative' prior in different circumstances, so for now let us choose the uniform distribution for μ (all values are equally likely).

What about the likelihood? For the Poisson distribution, this should be

$$\mathrm{Prob}(E|\mu) = \frac{e^{-100\mu}(100\mu)^1}{1!} = 100\mu e^{-100\mu}$$

The evidence is the just the integral of likelihood $\times$ the prior taken over the variable μ. The calculated posterior is shwon in Fig. 3.2

Now onto your new survey. For any given value of μ you can use Poisson's statistics to calculate the survey's yield. But we don't know the exact value of μ, we just have a PDF. This means that we marginalise over all values of μ. So the probability of k FRB detections in the new survey (with 10 times the footprint) is

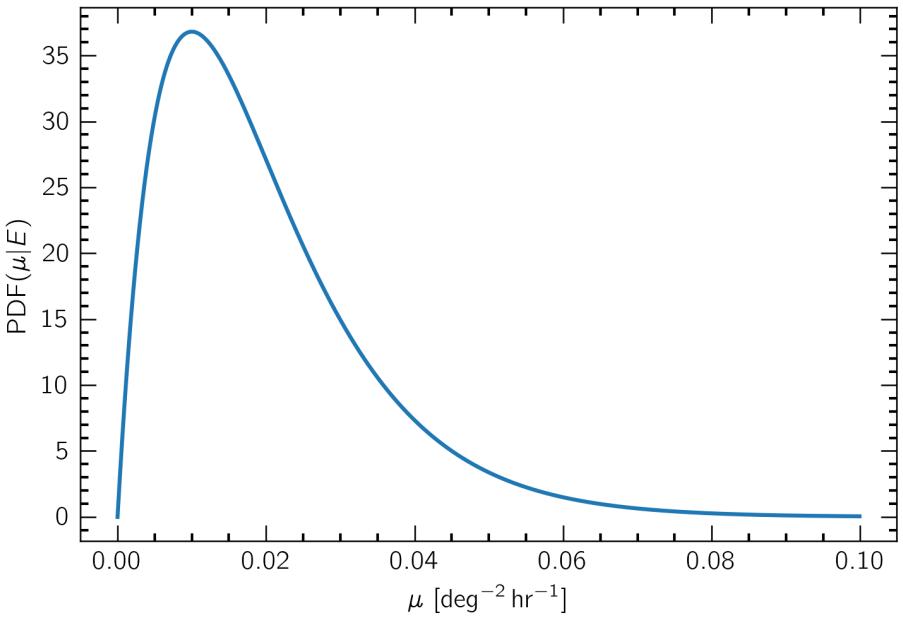


Figure 3.2: Posterior distribution of the FRB rate parameter from a single discovered FRB

$$\text{Prob}(k) = \int_0^\infty \mathrm{d}\mu \; \text{Prob}(k|\mu)\text{Prob}(\mu)$$

which is

$$\text{Prob}(k) = \int_0^\infty \mathrm{d}\mu \; \frac{e^{-1000\mu}(1000\mu)^k}{k!}\text{Prob}(\mu)$$

Here the calculation for different value of k (Fig. 3.3).

We see that there is 0.7% chance that no FRBs are detected; in other words a 99.3% chance that at least one FRB is detected. The distribution also peaks around 10, as anticipated for a ten-fold increase in the survey footprint.

Let us now mathematically work out the expected value of the Poisson distribution. By definition, the expected value of a Poisson variable k is given by

$$\langle k \rangle = \sum_{k=0}^{k=\infty} k\mathcal{P}(\lambda, k) = \sum_{k=0}^{k=\infty} k \, \frac{e^{-\lambda}\lambda^k}{k!}$$

Notice that the $k = 0$ term goes to zero. So we only need to consider

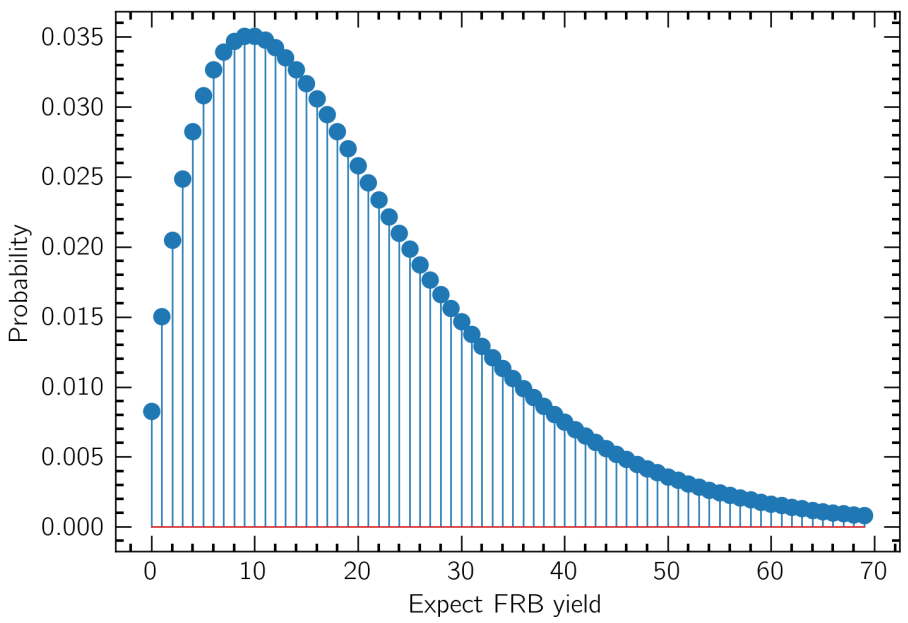


Figure 3.3: PDF of the expected yield in the new FRB survey

terms with $k \geq 1$. This gives

$$\langle k \rangle = e^{-\lambda}\lambda \sum_{k-1=0}^{k-1=\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

From the Taylor expansion of $e^x = 1 + x + x^2/2! + x^3/3!....$ we have the expected value of

$$\langle k \rangle = e^{-\lambda}\lambda e^{\lambda} = \lambda.$$

So the expected value of the Poisson distribution is just λ.

What about the variance of the Poisson distribution. By definition, the variance is given by

$$\sigma^2(k) = \left\langle (k - < k >)^2 \right\rangle = \left\langle k^2 - 2k < k > + < k >^2 \right\rangle = \left\langle k^2 \right\rangle - < k >^2$$

We already know that $< k >= \lambda$. So we now need to evaluate $< k^2 >$. This is given by

$$\left\langle k^2 \right\rangle = \sum_{k=0}^{k=\infty} \frac{e^{-\lambda}\lambda^k k^2}{k!} = e^{-\lambda}\lambda \left[\sum_{k-1=0}^{k-1=\infty} \frac{\lambda^{k-1}(k-1+1)}{(k-1)!} \right].$$

$$\left\langle k^2 \right\rangle = e^{-\lambda}\lambda \left[\lambda \sum_{k-2=0}^{k-2=\infty} \frac{\lambda^{k-2}}{(k-2)!} + \sum_{k-1=0}^{k-1=\infty} \frac{\lambda^{k-1}}{(k-1)!} \right]$$

$$\left\langle k^2 \right\rangle = e^{-\lambda}\lambda \left[\lambda e^{\lambda} + e^{\lambda} \right] = \lambda^2 + \lambda$$

The variance is therefore

$$\sigma^2(k) = \left\langle k^2 \right\rangle - \langle k \rangle^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

In summary, if you have a Poisson process with rate parameter λ then its expected value if λ and its standard deviation is $\sqrt{\lambda}$. Finally, we wont prove it here, but the median of the Poisson distribution is also λ.

3.2 The law of large numbers

An important application of Poisson statistics in astrophysics is the number of photons you receive from a source. In many astrophysical sources, photons are emitted by a large aggregate of electrons in a random fashion. In other words, no two electrons are coordinating photon emission; they are doing their own thing. Hence the arrival time of one photon is independent of the arrival of the next photon, and the one after that and so on. For a source of finite brightness we can always define an infinitesimally small time interval where at most one photon may arrive. So we can sidestep the practical problems with the binomial distribution and just use its small-number-limit, the Poisson distribution.

Suppose we define the rate parameter λ as the number of photons received per unit time at our detector/telescope. If you are operating at X-ray wavelengths, each photon has a large amount of energy so it is not uncommon to detect only a small number of photons: $\lambda = 1\,\mathrm{sec}^{-1}$ or even $\lambda = 10^{-2}\,\mathrm{sec}^{-1}$. What if we are operating at radio wavelengths where the energy of each photon is miniscule. We may end up with a huge value for λ. It therefore makes sense to study the large-number-limit of the Poisson distribution to see if further analytical simplifications can be made.

You may protest by saying that there is a contradiction here. We derived the Poisson distribution as a small-number-limit and now I am asking you to consider its large-number-limit. Well not really, when deriving the Poisson distribution from the Binomial distribution, we said that the number of binomial trials N is very large. We said nothing about the expected value Np (which is also the Poisson rate parameter) or $N(1-p)$. We are now saying that not only is N large, but Np and $N(1-p)$ are also very large.

Let us study this limiting case: $\lambda \gg 1$. Notice that the standard deviation of the distribution (the width of the curve) is $\sqrt{\lambda}$. Therefore if $\lambda \gg 1$ the we only need to consider values of k that are also in the vicinity of λ. Therefore we also have the limit $k \gg 1$. In this limit, there is a very handy approximation for the factorial, called Sterling's approximation, given by

$$k! \approx k^k e^{-k} \sqrt{2\pi k}$$

With this approximation, the distribution becomes

$$f(k \gg 1, \lambda \gg 1) \approx \frac{e^{-\lambda} \lambda^k}{k^k e^{-k} \sqrt{2\pi k}}$$

Now let us use the variable, $x = k - \lambda$ which is the offset from the expected value of the distribution. We get

$$f(k \gg 1, \lambda \gg 1) \approx \frac{e^x}{\sqrt{2\pi(x+\lambda)}} \left(\frac{\lambda}{x+\lambda} \right)^{x+\lambda}$$

We know that $x \ll \lambda$ in our limit, so we can approximate $2\pi(x+\lambda) \approx \lambda$, but because x appears in both the exponent and the fraction in the parenthesis, we are a bit stuck with this term. To get unstuck let us work with the logarithm of f.

$$\log f \approx x - \log \sqrt{2\pi\lambda} - \lambda \left(\frac{x}{\lambda} + 1 \right) \log \left(\frac{x}{\lambda} + 1 \right)$$

Now we can Taylor expand the logarithm in powers of x/λ to second order using $\log(1+x) \approx x - x^2/2$ for $x \ll 1$, to get

$$\log f \approx -\log \sqrt{2\pi\lambda} + \lambda \left[\frac{x}{\lambda} - \left(\frac{x}{\lambda} + 1 \right) \left(\frac{x}{\lambda} - \frac{x^2}{2\lambda^2} \right) \right]$$

Omitting the third order term, we get

$$\log f \approx -\log \sqrt{2\pi\lambda} - \lambda \frac{x^2}{2\lambda^2}$$

Taking back the exponent, and substituting $x = k - \lambda$ we get

$$f(\lambda \gg 1, k \gg 1) \approx \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{(k-\lambda)^2}{2\lambda}}$$

This is called the Gaussian density function or the normal distribution. The mean and variance are usually denoted by μ and σ^2 and let us use the symbol $\mathcal{N}$ to denote the Gaussian distribution:

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

Let us carry though our usual practice of finding out the mean and variance of any new distribution we encounter. The expected value of a Gaussian variable is

$$\langle x \rangle = \int_{-\infty}^{\infty} \mathrm{d}x \mathcal{N}(x|\mu,\sigma) = \int_{\infty}^{\infty} \mathrm{d}x \frac{x}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The integration can be split into two parts using $x = (x-\mu)+(\mu)$. The second term is just $\mu\times$ the area under the density function which gives μ. The first term is an odd function of $x-\mu$ whose integral from $-\infty$ to ∞ must vanish. Hence the expected value of the Gaussian distribution is

$$\langle x \rangle = \mu$$

What about the variance? This is a bit more involved but can be done with the help of a basic table of mathematical functions. The variance is the second central moment, which is

$$\sigma^2(x) = \left\langle (x-\mu)^2 \right\rangle = \int_{-\infty}^{\infty} \mathrm{d}x \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Using the change of variables $x-\mu \to x$ this becomes

$$\sigma^2(x) = \int_{-\infty}^{\infty} \mathrm{d}x \frac{x^2}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

Now using the change of variables $x^2/(2\sigma^2) \to x$. we get

$$\sigma^2(x) = \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{\infty} \mathrm{d}x x^{1/2} e^{-x}$$

These integrals are called Gamma functions defined by

$$\Gamma(z) = \int_0^{\infty} \mathrm{d}x x^{z-1} e^{-x}$$

which have the property $\Gamma(z+1) = z\Gamma(z)$ and $\Gamma(1/2) = \sqrt{\pi}$. Using these properties, we get

$$\sigma^2(x) = \frac{2\sigma^2}{\sqrt{\pi}} \Gamma(3/2) = \sigma^2$$

Hence the Gaussian distribution has a mean of μ and a variance of σ^2.

3.2.1 Central limit theorem

The Gaussian distribution is a special distribution among all the PDFs. Say you have random values drawn from some arbitrary distribution (doesn't matter which). Let us assume that each value you draw is independent of the other values. If you draw N values x_1, x_2, ...x_N and then calculate the arithmetic mean of these values: $\bar{x} = N^{-1} \sum x_i$ then as $N \to \infty$ the mean value $\bar{x}$ tends to a Gaussian random variable. This amazing statement is called the central limit theorem. It is amazing because it is true regardless of which distribution you drew random variables from! We already saw that as the number of events in a Poisson process tends to large values, the distribution reverts to a Gaussian distribution (it is how we derived the form of the Gaussian distribution). The theorem is saying that, it does not matter which distribution you start with. As long as you keep drawing a very large number of values and sum those values, the sum itself is a random variable with Gaussian statistics.

This is why the Gaussian distribution is really the 'law of large numbers', and why it appears in so may places. The general rule of thumb you can use is this: if whatever phenomenon you are observing is the aggregate result of a large number of independent 'things' happening then you can safely use the Gaussian distribution.

As an example, consider the thermal noise generated in electronic circuits. The noise is a result of random motions of electrons in the circuits. Every time an electron 'jiggles' it emits a photon. The noise is basically the electric field you see is the aggregate result of a large number of such photons and is therefore Gaussian distributed. Indeed as long as $h\nu \ll k_B T$ we are always in the large N regime and the detector noise can be considered to be Gaussian distributed.

3.3 Functions of random variables

We have already learnt about a few distribution and feel well equipped. But life always throws challenges at us. Take this for example: suppose you work for Gemeente Groningen and are tasked with setting up a radar that measures the speed of a car, v. Your radar has some thermal noise so your measured velocity is the true velocity v_t plus some random noise, n: $v = v_t + n$. You have good reasons to believe that the central limit theorem applies to the noise. So you know that n is Gaussian distributed. You take several measurements of a stationary car to confirm the properties of the noise. Here you know for sure that $v_t = 0$ so $v = n$ in this case. You make a histogram and it looks like Fig. 3.4. The distribution looks Gaussian. Nice! Also the mean value seems to be very close to zero. This means the

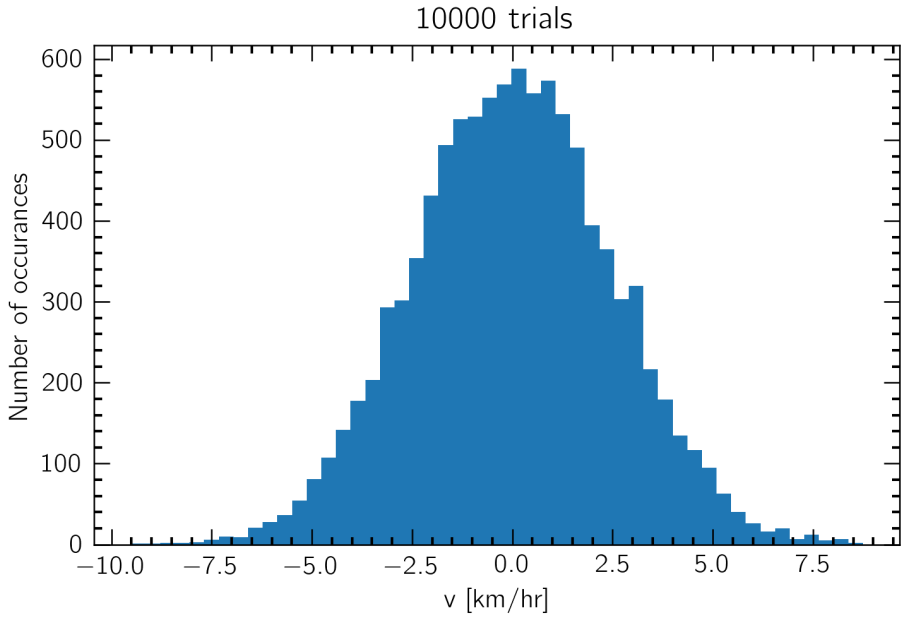


Figure 3.4: Histogram of noise samples from a speed measuring apparatus

detection has no systematic bias, just that it has some noise that generates uncertainty. You can take the variance of the samples in the histogram or fit a Gaussian curve to the histogram to get your best estimates for the mean and variance of the detector. I did this for values in my simple simulation and got 0.01 and 2.5 respectively. So the noise must be a Gaussian variable with $\mu = 0$ and $\sigma = 2.5\,\text{km/hr}$.

Suppose you take this device out to the streets and measure the speed of a car. It reads 57.8 km/hr. You know that the speed is 57.8 ± 2.5 m/s. But what is the probability density function for your estimated speed? You know intuitively that this should be a Gaussian with a mean of 57.8 km/hr and a standard deviation of $\sigma = 2.5\,\text{m/s}$. In fact, you can show this more formally using Bayes' theorem. Your variable to be estimated is the true speed v_t, the experiment E is the reading of 57.8 m/s. The likelihood is the probability of obtaining the experimental value given some true speed v_t. This is exactly equal to the probability of the noise value being $57.8 - v_t$. So the likelihood is

$$\text{Prob}(E|v_t) = \frac{1}{\sqrt{2\pi(2.5^2)}} e^{-\frac{(v_t - 57.8)^2}{2(2.5)^2}}$$

If your prior is uniform then your posterior is just a Gaussian with a mean

of 57.8 km/s and a standard deviation of 2.5 km/hr as anticipated.

OK now suppose your are not really interested in the speed but rather the kinetic energy of the cars: $\mathcal{E} = 1/2 \times mv^2$. Suppose you know the mass apriori what is the posterior density function of the energy $\mathcal{E}$? What I am asking you is this: suppose the PDF of a random variable x is given by $f_x(x)$, what is the PDF, $f_y(y)$ of the random variable y that is related to x by $y = g(x)$?

Let us return to the definition of probability density functions. Consider a small range of values around x_0 given by the interval $[x_0 - \mathrm{d}x/2,\ x_0 + \mathrm{d}x/2]$. Let us assume that there is a monotonic mapping between the variable x and y, that is the function g is such that if x increases, y either always increases or always decreases. The probability that x lies in the interval $[x_0 - \mathrm{d}x/2,\ x_0 + \mathrm{d}x/2]$ is given by $f_x(x_0)\mathrm{d}x$. Because of the monotonic criterion, this must be the probability that y lies in the corresponding range $[g(x_0 - \mathrm{d}x),\ g(x_0 + \mathrm{d}x)]$ which by definition of the PDF is $f_y(y)|g(x_0 - \mathrm{d}x) - g(x_0 + \mathrm{d}x)|$. Note that I have taken the absolute value to handle both monotically increasing and decreasing relationships while keeping the PDF always positive. We now we have the relationship we seek:

$$f_x(x)\mathrm{d}x = f_y(y)|g(x_0 - \mathrm{d}x) - g(x_0 + \mathrm{d}x)|$$

which in differential form is

$$f_x(x) = f_y(y)\left|\frac{\mathrm{d}g(x)}{\mathrm{d}x}\right|$$

or

$$f_y(y) = f_x(x)\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|^{-1}$$

In our example, $x = v$ and $y = \frac{1}{2}mv^2$. So $\frac{\mathrm{d}y}{\mathrm{d}x} = mv$. The PDF of the energy is therefore

$$f_{\mathcal{E}}(\mathcal{E}) = f_v(v)/(mv) = f_v\left(\sqrt{2\mathcal{E}/m}\right)/\sqrt{2\mathcal{E}m}$$

Because f_v is a Gaussian random variable, we have

$$f_{\mathcal{E}}(\mathcal{E}) = \frac{1}{\sqrt{4\pi\sigma^2\mathcal{E}m}}e^{\left[-\frac{\left(\sqrt{2\mathcal{E}/m}-\mu\right)^2}{2\sigma^2}\right]}$$

The special case of this distribution when $\mu = 0$ is called the Exponential distribution. As an astronomer you may encounter this distribution in the noise power in some of your detector circuits. Remember that we said that the electric field due to noise fluctuations was Gaussian distributed? Well

the power is proportional to the square of the electric field which will then be exponentially distributed.

Finally, in case we have a multi-variate function going from variables x, y to l, m of the form $l, m = f(x, y)$ then the differential must be replaced by the determinant of the Jacobian matrix:

$$\text{Prob}(l, m) = \text{Prob}(x, y) \left| \det \begin{bmatrix} \frac{\partial x}{\partial l} & \frac{\partial x}{\partial m} \\ \frac{\partial y}{\partial l} & \frac{\partial y}{\partial m} \end{bmatrix} \right|$$

3.4 Gull's lighthouse problem

We will end this week's discussion with a famous albeit theoretical problem that will introduce you to a new distribution called Cauchy's distribution. The problem is stated as follows. Suppose you are on a shore that is parallel to $y = 0$ line. There is a light house at location $x = \alpha$, $y = \beta$. This is a weird light house. Instead of sending out a rotating beam of light, it sends out light pulses in totally random directions. Suppose you have a linear stretch of photo detectors along the shore along $y = 0$. If a light pulse reaches the shore line at $y = 0$ the detectors measure the x location of the pulse. Using the values of the x-locations, say x_i, can you find the location of the light house?

Now let us see *how* the information on the ticks can even be used here. First of all, we only need to consider light pulses that reach the line of detectors. That is, they have to be fired at a particular elevation angles to reach the line of detectors along the $y = 0$ line. Ok then we are left with one angle θ which is a random variable to denote the direction of the light pulses. If $\theta = 0$ the pulse lands at $x = \alpha$ and only when θ is in the interval $(-pi/2, \pi/2)$, the pulse will intersect the shore line. Because the light house is firing pulses at random directions, we have the uniform density function for θ between $-\pi/2$ and $\pi/2$:

$$f_\theta(\theta) = \frac{1}{\pi},\ -\pi/2 < \theta < \pi/2;\ = 0, \text{otherwise}$$

Ok we are asked to estimate the probability of the light-house being at some location α, β given the outcome of the experiment: the shore-line values x_k. This is the posterior distribution and the parameters are α and β. Shall we evaluate the likelihood? This is the PDF of the shore-line locations given α and β. Notice that the shore line location of one pulse is independent of the locations of other pulses.

$$\text{Prob}(x_1, x_2, ...x_n|\alpha, \beta) = \prod_i \text{Prob}(x_i|\alpha, \beta)$$

So if we can find the probability distribution function of x, we can evaluate the likelihood but we are only given the PDF of θ. We must change variables. A pulse fired at angle θ will reach the shore line at location $x(\theta) = \alpha + \beta \tan\theta$. So now we have to convert the PDF of θ into a PDF of x. We know how to do this. First notice that x is a monotonically increasing function of θ in the interval $\theta \in (-\pi/2, \pi/2)$. So we just need to find the differential:

$$\frac{\mathrm{d}x}{\mathrm{d}\theta} = \beta \mathrm{sec}^2\theta = \beta(1 + \tan^2\theta) = \beta(1 + (x-\alpha)^2/\beta^2)$$

$$\mathrm{Prob}(x_i|\alpha,\beta) = \mathrm{Prob}(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}x} = \frac{\beta}{\pi\left[\beta^2 + (x_i-\alpha)^2\right]}$$

This is called the Cauchy-Lorentz distribution function, or commonly, the Cauchy distribution. It is the same function that describes the intrinsic line profile of spectral lines. But as a distribution, it has some very weird properties: its mean and variance are undefined! This means that if you try to compute the first and second moments, you will end up with an undefined integral— one whose value diverges. Why does this happen? It is because even though the distribution is peaked around $x = \alpha$, its wings are so shallow that there is a theoretically infinite amount of area under the wings. The distribution just does not fall rapidly enough as one moves away from the peak.

Anyways, lets return to the light-house problem. Assuming uniform priors on α and β in some range, we have the posterior

$$\mathrm{Prob}(\alpha,\beta|x_i) \propto \prod_i \mathrm{Prob}(x_i|\alpha,\beta) \propto \prod_i \frac{\beta}{\pi\left[\beta^2 + (x_i-\alpha)^2\right]}$$

We have not really dealt with the intricacies of multi-parameter estimation. So let us for the moment, fix the value of β (using a delta-function prior) and only focus on estimating α. We have the posterior

$$\mathrm{Prob}(\alpha|x) \propto \prod_i \frac{\beta}{\pi\left[\beta^2 + (x_i-\alpha)^2\right]}$$

Of course, like in previous examples, we can run to a computer programme to calculate the posterior, but let us stay analytical a little longer here to learn some new tricks. Suppose I am only interested in the most likely value of α, that is the peak-location of the posterior. We know that at the peak, the derivative w.r.t α must vanish. But taking the derivative of the product will give us a headache. Here is where a nice property of probabilities is helpful. Notice that the probabilities are always positive. So

the value of α that maximizes the PDF should also maximize the logarithm of the PDF. So we can just try to find the peak of the log-PDF by setting derivative-of-log-PDF equal to 0. The logarithm of the posterior is

$$L = \log \text{Prob}(\alpha|x) = \text{constant} - \sum_i \log\left[\beta^2 + (x_i - \alpha)^2\right]$$

where 'constant' sequesters all terms that do not depend on α.

Setting the derivative w.r.t α equal to zero yields

$$\frac{\mathrm{d}L}{\mathrm{d}\alpha} = -\sum_i \frac{x_i - \alpha}{[\beta^2 + (x_i - \alpha)^2]} = 0$$

The solution to this equation should give us the most likely value of α. I have written a computer code to (a) generate the locations of the random light pulses with $\beta = 1$ and $\alpha = 0$, and the log-posterior. The results are shown in Fig. 3.5 where each panel is for different number of pulses received.

You can see the weirdness caused by the mean and variance not being defined for the Cauchy distribution especially for small values of n: the mean of the curve is not inexorably reverting to the true value of 0. In fact if one estimates the mean of the curves in each of the panels, then one finds that the mean values are wildly jumping between random values rather than progressing closer and closer to some 'true' value. The same is true of the standard deviation. Here is a numerical experiment to show the pathological nature of the Cauchy distribution. I drew samples at random from a Gaussian and a Cauchy distribution both centered on zero and with comparable widths. Every time I drew a number, I re-evaluated the sample mean and sample standard deviation using all the numbers I had drawn up until then. Then I plot the mean and standard deviations to see if they converge to the true values as the number of draws becomes large. Figure 3.6 shows the results and it is clear that the Gaussian random variables behave as we would expect them to but the Cauchy random variables are out of control! Neither their mean not standard deviation neatly converge to some 'true' values. The way these quantities are jumping around proves that mean and standard deviation are not meaningful concepts for a Cauchy random variable. The root cause of this property is that the Cauchy distribution has very wide wings, so it is highly probable that some hugely positive of hugely negative number will be drawn soon enough that will wildly swing the values of the sample mean and sample standard deviation.

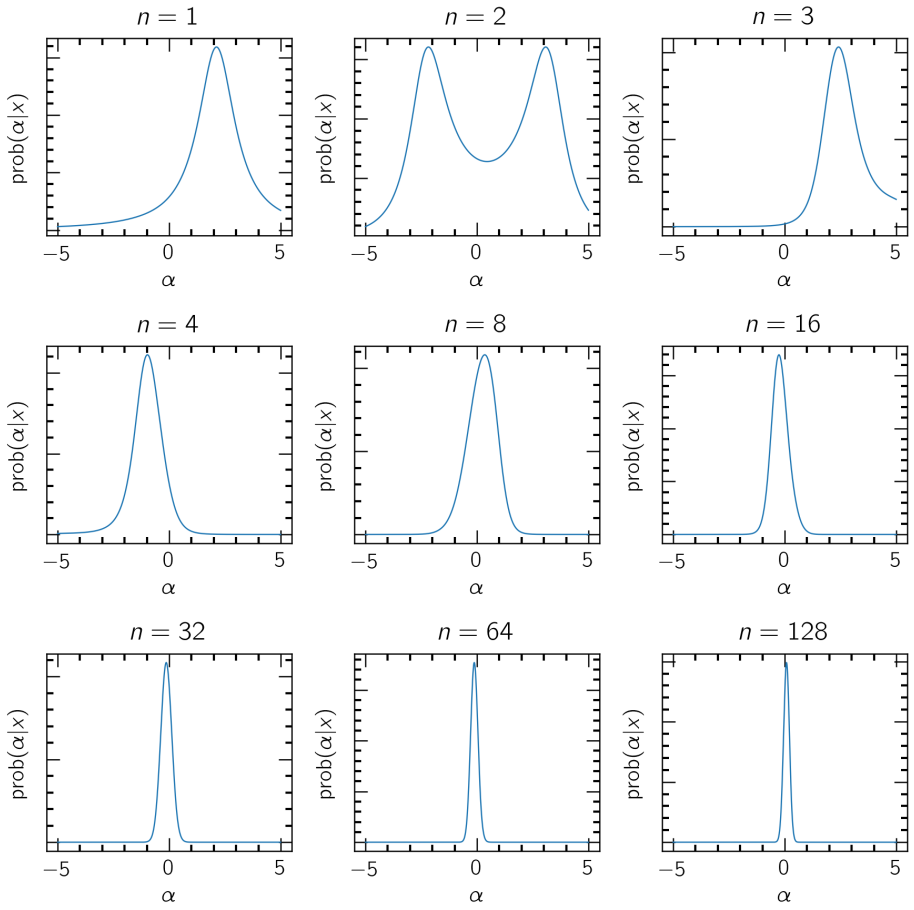


Figure 3.5: The posterior of α for different number of light flashes detected.

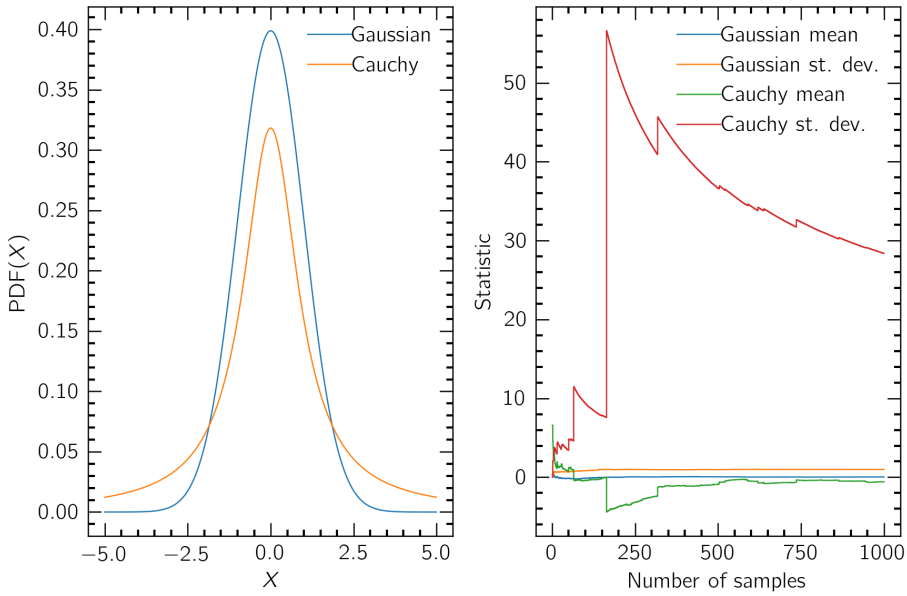


Figure 3.6: Sample mean and sample standard deviation as sample size increases for a Gaussian and Cauchy random variable (right panel). Left panel shows the two distributions.

Chapter 4

Multivariate parameter estimation

So far we have tried to estimate one parameter at a time. We had a sneak peak at the case of multiple parameters being estimated when we dealt with Gull's lighthouse problem. But we simplified matters by fixing one of the parameters. Now you are ready to get to the next level by going to more realistic problems where multiple parameters must be estimated at once. This presents a new set of problems in defining their posterior PDFs and assigning uncertainties to them. For instance what is the right way to deal with circumstances when the uncertainty in one parameter is related to the uncertainty in the other, that is, they are not independent? By the end of this chapter you will be able to deal with real life problems where you have some data, a conception of a model that described that data and you want to estimate multiple parameters in that model and assign confidence levels to these parameters.

4.1 Signal detection: a 2-parameter case study

Picture this. You are up late at night in the telescope operating room taking the spectrum of an exoplanet. You are really hoping to see the presence of aurorae on that exoplanet. No one has seen such a signal. If you succeed, it will be a great discovery.

The plan is to look for the presence of a Balmer line of hydrogen around 6560 Angstrom wavelength. If you see excess emission at that wavelength, it will prove that auroral activity exists on the exoplanet. You see the spectrograph readout and it looks like the curve in Figure 4.1.

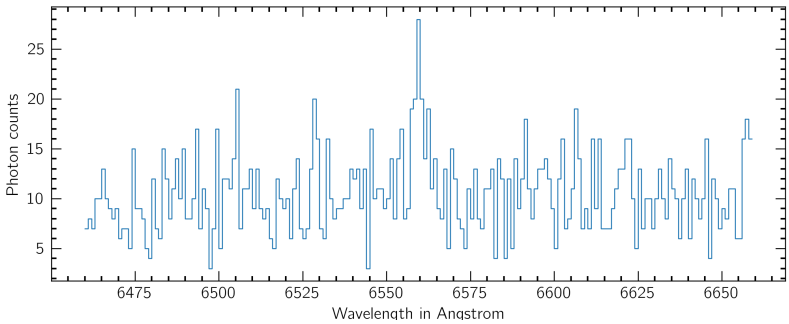


Figure 4.1: Spectrum in the vicinity of a Balmer series line at 6560 Angstrom.

There is something at 6560 Angstrom. Is that the auroral signature? Well its not entirely clear as there are also seemingly random fluctuations at other wavelengths. How to calculate our confidence in the veracity of the auroral line?

In problems like this, it helps to take a step back and construct a mental model of why the data looks they wait it does. In this situation, the model must of course be based on our prior and theoretical understanding of physics. What we are seeing here is the number of photons that arrive in different small wavelength bins. The number of photons must be a Poisson random variable. So if the Poisson rate parameter in the vicinity of 6560 Angstrom is significantly higher than the baseline value then we know that the signal is real. Notice also that the photon counts at all wavelengths are at a mean level of around 10. This is because there is some background light (in addition to any potential aurora) that is coming from either the exoplanet itself, or some diffuse sky background (the sky is not perfectly dark even at night), or maybe some stray light that is entering the telescope. Based on your knowledge of physics you assert that this

background should not have a strong wavelength dependence. This is to say that the Poisson rate parameter for the background does not depend on the extract value of the wavelength, and that the number of photons from the background is uncorrelated from one wavelength bin to another. OK that explains the background fluctuations. What about the potential aurora itself? What is it's shape? You know that the aurora should be centred on 6560 Angstrom. Can you just look for this signal in that one wavelength bin? That is an option. But what if the line was 10 Angstroms wide? The photons from the aurora will then be spread out over 10 wavelength bins. A better option would then be to use all 10 bins. That would give you less uncertain estimate of the aurora's brightness because more data means less uncertainty. But how you combine the data in the 10 bins also depends on what you expect the line profile to be. Is it like a top hat? Is it some smooth curve? You know from physics that the line must have a Gaussian profile if it originates in a thermal medium[1]. So with all of this information you can arrive at the following model

$$N_k = b_k + a_k$$

$$b_k \,\text{drawn from}\, \mathcal{P}(\lambda_b)$$

$$a_k = \,\text{drawn from}\, \mathcal{P}\left(\lambda_a e^{-(k-k_0)^2/(2\sigma^2)}\right)$$

where N_k is the number of photons in the k^{th} wavelength bin and is comprised of the background b_k and signal a_k. b_k are Poisson random variables with rate parameter λ_b. a_k are Poisson random variables with rate parameter that depends on k and this dependence is given by a Gaussian profile centred at $k_0 = 6560$ Angstrom and width σ.

So now that we are given the data, we have to estimate 3 parameters: λ_b, λ_a and σ. In other words, we wish to calculate the posterior

$$\text{Prob}(\lambda_b, \lambda_a, \sigma | N_1, N_2 ...).$$

Let us for the sake of simplicity of illustration, make this a 2-parameter estimation by fixing the value of $\sigma = 2$ Angstrom. You cannot do this in reality without some other knowledge of the width but we do it here to make our discussion simple.

Using Bayes theorem we can write

$$\text{Prob}(\lambda_a, \lambda_b | N_1, N_2 ...) \propto \frac{\text{Prob}(N_k | \lambda_a, \lambda_b)}{\text{Prob}(\lambda_a, \lambda_b)}$$

[1]Gaussian here is the shape of the line versus wavelength curve, and not a random variable!

We know that the exoplanet aurora cannot know anything about the sky background photons so we have $\text{Prob}(\lambda_a, \lambda_b) = \text{Prob}(\lambda_a)\text{Prob}(\lambda_b)$ due to their independence. Further more, let us say that we have no prior knowledge of λ_a and λ_b and choose flat prior. We are then just left with evaluating the likelihood. Because each wavelength bins has photons arriving independently of the other bins, we have

$$\text{Prob}(N_1, N_2...|\lambda_a, \lambda_b) \propto \prod_k \text{Prob}(N_k|\lambda_a, \lambda_b)$$

Each term in the product is the probability of a random variables, which itself is a sum of two Poisson random variables (background and signal). It can be shown (proof in Appendix) that the sum of two Poisson random variables with rate parameters λ_a and λ_b is given by another Poisson random variable with rate parameter $\lambda_a + \lambda_b$. So the likelihood is

$$\text{Prob}(N_1, N_2..|\lambda_a, \lambda_b) \propto \prod_k \frac{e^{-\lambda_k}\lambda^{N_k}}{N_k!}$$

where the rate parameter of the k^{th} element is

$$\lambda_k = \lambda_b + \lambda_a e^{-(k-k_0)^2/(2\sigma^2)}$$

We can now do the calculations of the posterior on a computer. Remember that we now have two variables to estimate so we have to calculate the posterior on a 2-dimensional grid of values along the λ_a and λ_b axes. If we had retained σ as a third parameter to estimate, we would have to do the calculations on a 3-dimensional grid. Before we look at the posterior in Figure 4.2, it is worth commenting on some numerical tricks that are necessary. In high dimensions and/or when there are a large number of data points, each evaluation of the Poisson probability yields a very small number which has to me multiplied with a lot of other small numbers which can cause numerical resolution issues. An easy way to avoid these problems is to calculate the logarithm of the probability. Another tricks is that one can carefully isolate only log-probability terms that have the parameters in them and the rest of the terms can be cast as a constant (including the evidence). This constant need not be calculated as one can always normalise the posterior later.

The log posterior is given by

$$\log \text{Prob}(\lambda_a, \lambda_b|N_1, N_2...) = \text{constant} + \sum_k -\lambda_k + N_k \log \lambda_k$$

where 'constant' contains all terms that do not depend on the parameters λ_a and λ_b.

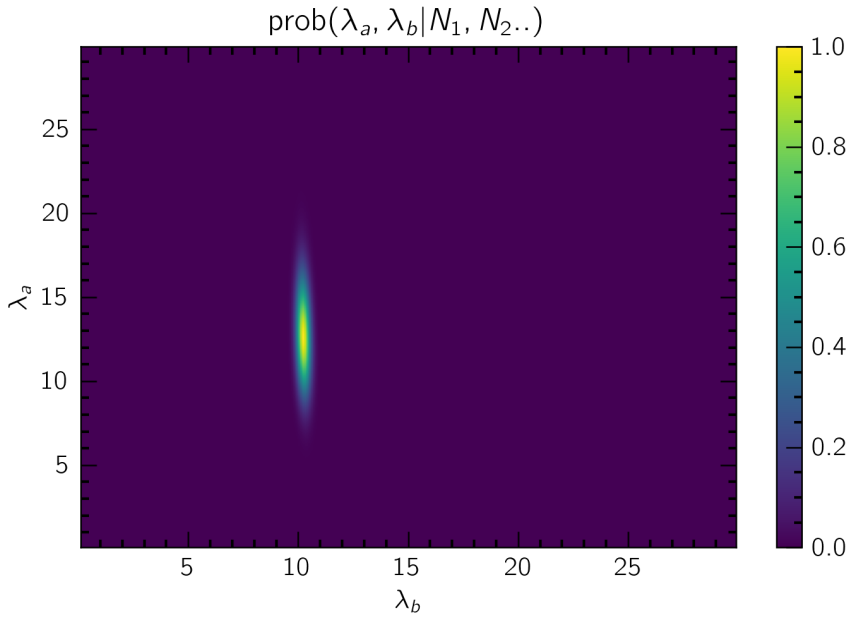


Figure 4.2: Posterior probability density of the parameter pair λ_a, λ_b.

OK with those said, let us look at the posterior in Fig. 4.2.

Ok we now see that the rate parameter for the background is around 10 as expected. But the crucial part is that the rate parameter for the aurora is around 10 as well! This looks promising but we need to get proper confidence intervals. In other words we want $\text{Prob}(\lambda_a|N_1, N_2...)$ and we only have $\text{Prob}(\lambda_a, \lambda_b|N_1, N_2...)$. We know how to do that; via marginalisation:

$$\text{Prob}(\lambda_a|N_1, N_2) = \int \mathrm{d}\lambda_b \, \text{Prob}(\lambda_a, \lambda_b|N_1, N_2...)$$

The marginalised PDF for λ_a is given in Figure 4.3.

From here on we know how to calculate confidence levels, central value estimates and uncertainty estimates. One thing which is worth remembering here is that such estimates can change due to marginalisation. For example, in our case, the peak of the 2-dimensional posterior, which is the mask likely value, occurs at $\lambda_a = 12.6$, $\lambda_b = 10.2$. However, the peak of the marginalised posterior occurs at $\lambda_a = 12.8$. In this case the most likely values are not too different but as the dimensionality of the problem increases the most likely values can be quite different.

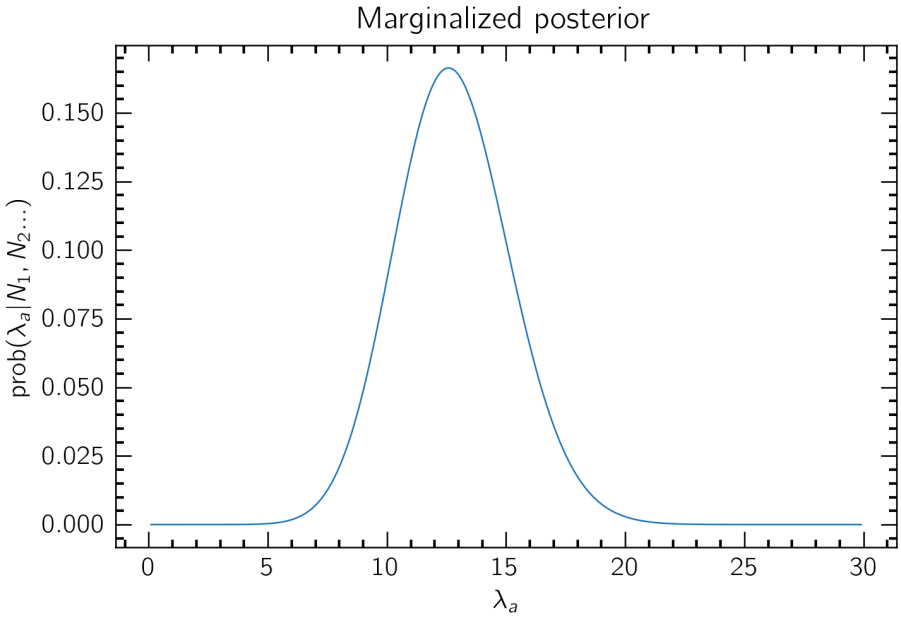


Figure 4.3: Marginalised PDF of λ_a

4.2 Least squares and shape of the posterior

That was fun! So let us consider another estimation problem. Imagine you were Edwin Hubble who had just measured the recession velocities and distances to a number of galaxies. You plot the data points up and get the following figure.

Let us assume that the distances are accurately measured and velocities are measured with substantial error (in reality it is actually harder to measure distances, but hey, this is just a hypothetical example). Let us assume zero uncertainty in distances and suppose you know from your measuring equipment that the uncertainty on the velocity is Gaussian distributed with zero mean and $\sigma = 200\,\mathrm{km/s}$. These errors are shown as bars spanning $\pm\sigma$ in the figure. The data points show that there is a linear relationship between the two quantities but the uncertainties really makes you wonder whether you can be statistically sure that there is a linear relationship.

Let us now write a model for the data. Of course we are trying to fit a straight line to the data with a slope, say H and intercept, say C. We are expecting C to be close to zero and H is the parameter that has the cosmological information we are after. Our model is

$$v_k = Hd_k + C + n_k$$

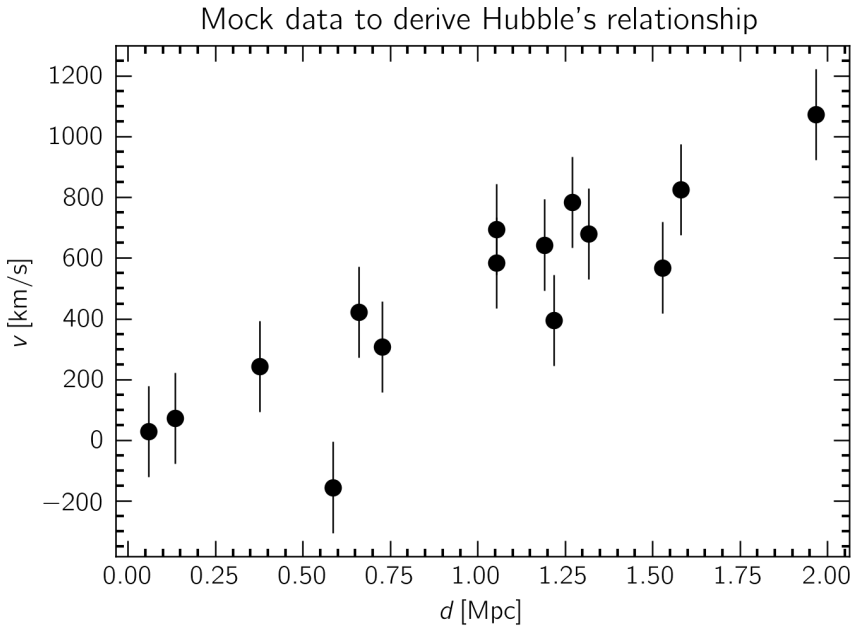


Figure 4.4: Mock data in our imaginary universe where you are Edwin Hubble

$$n_k \text{ drawn from } \mathcal{N}(0, 200)$$

where H and C are parameters to be estimated, v_k and d_k are the measurements and n_k represents noise realisations. Let us go through the routine we have so many times and write the posterior (we will use uniform priors on the parameters):

$$\text{Prob}(H, C|D) \propto \text{Prob}(D|H, C)\text{Prob}(H, C) \propto \text{Prob}(D|H, C)$$

where I have used the shorthand $D = v_1, v_2...., d_1, d_2...$ for the measured data.

Because the noise terms are Gaussian distributed and the noise in one measurement is independent of the others, we have

$$\text{Prob}(H, C|D) \propto \prod_k e^{-(v_k - Hd_k - C)^2/2\sigma^2}$$

As before, let us work with the log probabilities

$$\log \text{Prob}(H, C|D) \equiv \mathcal{L}(H, C|D) = \text{constant} - \sum_k (v_k - Hd_k - C)^2/(2\sigma^2)$$

We already know how to code this all up and get to the posteriors we want. But let us do something different for this problem and progress analytically. AS you will see, this will allow us to get some insights into the nature and shape of common posterior distributions.

Now, we want to find the pair (H, C) where the log-posterior, $\mathcal{L}$ is maximised. This will happen at the critical points when the derivative w.r.t to the parameters goes to zero. But there are two parameters, so we get two equations that can be solved simultaneously to get the values of the two parameters at the critical point.

$$\frac{\partial \mathcal{L}}{\partial H} = \frac{\partial \mathcal{L}}{\partial C} = 0$$

which gives

$$\sum_k (v_k - Hd_k - C)d_k = \sum_k (v_k - Hd_k - C) = 0$$

You may notice that this is a set of two linear simultaneous equations which must be solved to find the two parameters H and C. In fact the equations can be written in the matrix form

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

with solution

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

where

$$\mathbf{A} = \begin{bmatrix} \sum_k d_k^2 & \sum_k d_k \\ \sum_k d_k & \sum_k 1 \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} H \\ C \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} \sum_k v_k d_k \\ \sum_k v_k \end{bmatrix}$$

It is straightforward to calculate the matricities on a computer given the data d_k, v_k and there are any number of routines available to compute matrix inverses (e.g. `numpy.linalg.inv` in `python`) and matrix multiplications (e.g. `numpy.dot` in `python`). I have done this for our dataset and I get the values $H =$ and $C =$. Nice! If the posterior is a Gaussian distribution *and* the model is linear in the parameters, you can very quickly calculate the most likely value of the parameters (location where the posterior peaks) without actually calculating the whole posterior. This is very convenient.

The other thing we are usually interested in is the uncertainty in the parameter estimates. Can we also do this readily with some deft analytic work? Let us see. We already know that the uncertainty depends on how

quickly the posterior curve falls off as we go away from the central value (where it peaks). If it falls off rapidly then the uncertainty is small. We now have a way to get to the *location* of the peak. Can we do the same to get to the *shape* of the peak?

Let us try an old trick to approximate a curve: Taylor expand around a point. Let us Taylor expand the log likelihood, $\mathcal{L}$ around the peak value, say (H_0, C_0). We have

$$\mathcal{L}(H,C) \approx \mathcal{L}(H_0,C_0) + (H-H_0)\left[\frac{\partial \mathcal{L}}{\partial H}\right]_{H_0,C_0} + (C-C_0)\left[\frac{\partial \mathcal{L}}{\partial C}\right]_{H_0,C_0}$$

Well that is just not enough, because the partial derivatives are by definition zero at (H_0, C_0). This means that we need to expand to the next order to make any progress as the first order derivatives are both zero.

$$\mathcal{L}(H,C) \approx \mathcal{L}(H_0,C_0) + (H-H_0)^2\left[\frac{\partial^2 \mathcal{L}}{2\partial H^2}\right]_{H_0,C_0} + (C-C_0)^2\left[\frac{\partial^2 \mathcal{L}}{2\partial C^2}\right]_{H_0,C_0}$$
$$+(H-H_0)(C-C_0)\left[\frac{\partial^2 \mathcal{L}}{\partial H \partial C}\right]_{H_0,C_0}$$

Let us understand what this equation is telling us. The likelihood has a peak at H_0, C_0 and then will decline as we move away (the second derivatives are negative; else we are not even at the peak!). They decline quadratically due to the second order terms: $(H-H_0)^2$, $(C-C_0)^2$ and $(H-H_0)(C-C_0)$, and the rate of decline is decided by the factors (with second order derivatives) in front of the second order terms. The larger the factor the faster the decline. So the inverse of the factors should give us a measure of the width of the curve and therefore the uncertainty in the parameters.

But is this new measure we have the same as the standard deviation of the p[osterior distribution? Well that depends on the distribution. Let us digress to a simpler example where the posterior in a one-parameter estimation has a Gaussian shape with mean μ and deviation σ. The log-likelihood will then be quadratic:

$$\mathcal{L} = \text{constant} - (x-\mu)^2/(2\sigma^2)$$

The second derivative of the log likelihood is

$$\frac{\partial^2 \mathcal{L}}{\partial x^2} = 1/\sigma^2$$

which confirms our heuristic. The second derivative of the log-likelihood gives us the inverse of the variance. This result is exact for a Gaussian distribution; that is a distribution which is quadratic in log-likelihood. And

when we Taylor expand the log-likelihood of any other distribution to second order, what we are doing is approximating our posterior as a Gaussian distribution near the peak.

There is still one problem. If you have multiple parameters, there are multiple second derivatives. So how to use this to get the variance of each parameter. Let us now deal with this issue. Let me use a shorthand notation for the second derivatives and drop the arguments to make my life simple and write

$$\mathcal{L} \approx \mathcal{L}_0 + 1/2(H - H_0)\partial_H^2\mathcal{L} + 1/2(C - C_0)^2\partial_C^2\mathcal{L} + (H - H_0)(C - C_0)\partial_{HC}^2\mathcal{L}$$

This is just a second order equation in a 2-dimensional plane. It is of the form

$$f(x, y) = ax^2 + by^2 + cxy$$

where the origin has been shifted to the location of the peak and a, b, c are given by the second derivatives which have something to do with the inverse of the variances.

What is the shape of this function in the x, y plane? Well let us drop the third term involving xy then we get an equation that looks like the equation for an ellipse: $f(x, y) = ax^2 + by^2$. Nice! The semi-major and semi-minor axes have lengths of $\sqrt{f/a}$ and $\sqrt{f/b}$. Different values of f give you concentric ellipses with different axes lengths. Remember that we are interested in the variance which is the value of x at which our quadratic log-likelihood has the value $1/2$ less than the peak value. So we set $f = 1/2$ and call this our 'uncertainty ellipse'. This is the two-dimensional equivalent of saying same range is our $\pm\sigma$ uncertainty interval. Instead of an interval on a straight line, we now have an area on a 2-dimensional plane. The uncertainty ellipse has semi-major and semi-minor axes of lengths $(2a)^{-1/2}$ and $(2b)^{-1/2}$ which in our original notation is

$$(\partial_H^2\mathcal{L})^{-1/2} \text{ and } (\partial_H^2\mathcal{C})^{-1/2}$$

We could then quote these as the standard deviation of the posterior along the two parameters:

$$\sigma_H^2 = (\partial_H^2\mathcal{L})^{-1} \text{ and } \sigma_C^2 = (\partial_H^2\mathcal{C})^{-1}$$

The situation is complicated by the presence of the third term that involves xy factor. What does this term do? Well it turns out that this terms still gives an ellipse but this ellipse does not have its semi major and semi minor axes along the x and y axes. Instead it is a titled ellipse. What does this mean for our uncertainty? It means that we can no longer quote independent uncertainties along the H and C axes. We can only

quote two uncertainties along orthogonal axes that are at some angle in the H, C plane. We are then stuck with the situation where the uncertainty in H depends on the uncertainty in C and vice-versa. The two parameters are not independent and we use the wording: H and C have covariant uncertainties.

As you can imagine, doing the algebra with tilted ellipses will become very cumbersome. To gracefully deal with situations like this, we have over the decades, developed an efficient mathematical notation using the language of linear algebra and eigen-decomposition of matricies.

First let us write our quadratic posterior as a matrix equation.

$$\mathcal{L} = [H - H_0,\, C - C_0] \cdot \begin{bmatrix} \partial^2_H \mathcal{L} & \partial^2_{HC} \mathcal{L} \\ \partial^2_{HC} \mathcal{L} & \partial^2_C \mathcal{L} \end{bmatrix} \cdot \begin{bmatrix} H - H_0 \\ C - C_0 \end{bmatrix}$$

where we must always remember that the partial derivatives are evaluated at the peak H_0, L_0.

The tilting back of the ellipse is mathematical done via eigen-decomposition of the matrix

$$\mathbf{B} \equiv \begin{bmatrix} \partial^2_H \mathcal{L} & \partial^2_{HC} \mathcal{L} \\ \partial^2_{HC} \mathcal{L} & \partial^2_C \mathcal{L} \end{bmatrix}$$

which yields two eigen-values say λ_1 and λ_2 that are related to the semi-major and semi-minor axes of the tilted parabola. Unfortunately, we wont have time in this course to explore the general properties of the covariance matrix. We must at once return to our Edwin Hubble role-playing. For now, suffice it to say that the covariance matrix of the parameters is given by the matrix inverse $\mathbf{B}^{-1}$. If this inverse is a diagonal matrix then we return to our special case of separate uncertainties for the two parameters (non-tiled ellipse). If it is not then the uncertainties in the two parameter are coupled which is why this matrix is sometimes called the co-variance matrix, indicating that it is a multi-dimensional equivalent of the variance of the parameters.

OK back to Hubble. Let us now calculate the second derivatives of the log-likelihood.

$$\frac{\partial^2 \mathcal{L}}{\partial^2 H} = \frac{1}{\sigma^2} \sum_k d_k^2$$

$$\frac{\partial^2 \mathcal{L}}{\partial^2 C} = \frac{1}{\sigma^2} \sum_k 1 = N/\sigma^2$$

$$\frac{\partial^2 \mathcal{L}}{\partial H \partial C} = \frac{1}{\sigma^2} \sum_k d_k$$

Our matrix of second derivatives is then

$$\mathbf{B} = \begin{bmatrix} \frac{1}{\sigma^2}\sum_k d_k^2 & \frac{1}{\sigma^2}\sum_k d_k \\ \frac{1}{\sigma^2}\sum_k d_k & N/\sigma^2 \end{bmatrix}$$

and our parameter covariance matrix is just its inverse.

I have implemented the solution for the most likely value and the covariance matrix on a computer. I also calculated the posterior by brute-force (the old way). Figure 4.5 shows the result. As expected our analytical calculation of the most likely value, is bang on! I also did an eigen-decomposition of the covariance matrix and plotted the major and minor axes of the uncertainty ellipse which again agrees with our brute force calculation.

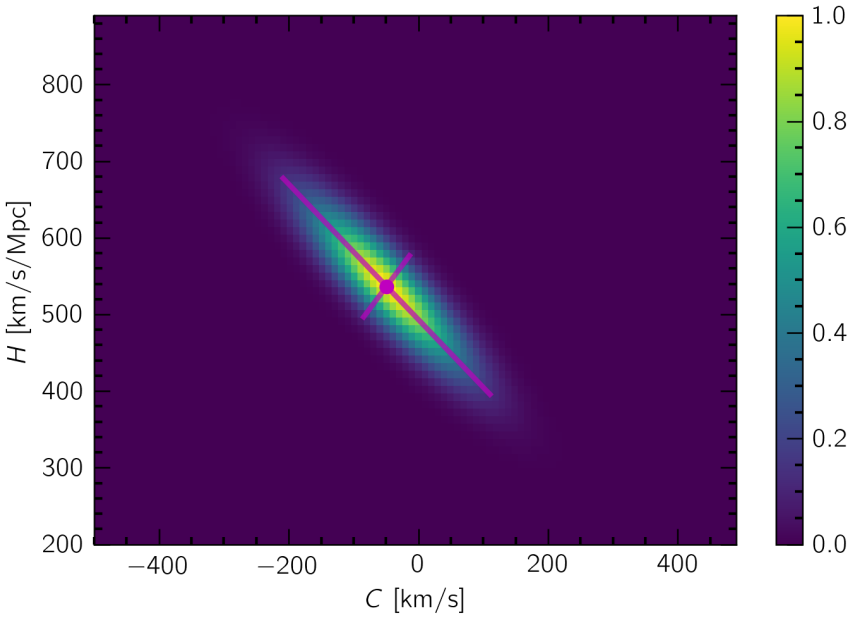


Figure 4.5: Posterior distribution of the Hubble parameter and offset for our mock data. The cross indicates the most likely value computed using the analytical expression derived in the text. The lines indicated the $\pm 2\sigma$ interval computed via eingen-decomposition of the covariance matrix.

Of course we can now go ahead and marginalise over the parameter C which we are not really interested in, to get the posterior distribution for the Hubble parameter H which we are interested in.

A final point worth making here is regarding the value of our analytical struggle to avoid brute force calculations. In this example we only had

two parameters to estimate. Image you had to estimate 5 parameters. If you went for a brute force evaluation then you would have to compute the posterior in 5-dimensional space. If you decided to place 100 grid points per dimension to accurately capture the posterior curve then you would end up with 10^{10} calculations of the posterior which will consume an insane amount of computational resources unnecessarily. So the short cut allowed by linear problems with Gaussian errors is very valuable! So let us end out chapter by writing down general equations for such estimations.

4.3 Linear estimation with Gaussian errors

We just saw how a straight line can be fit to data. This problem belongs to a general class of models that are linear in its parameters. For instance we could have a problem where a polynomial of the kind

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 ...$$

must be fit to a dataset that has the measurements $(x_1, y_1), (x_2, y_2)....$ This is also a linear model. Do not get confused by finding non-linear terms such as x^2. By linear we mean that the model is linear in its parameters that we are trying to estimate. The parameters here are $a_0, a_1 ...$ and the polynomial equation is clearly linear in the parameters.

We can go ahead of follow the same procedure as with the straight line fit and write down the posterior etc. But a lot of time and effort can be saved if we can derive some general properties of the posterior for all linear models with Gaussian errors. So let us do that now.

Let us collect all our parameters to be estimated in a 1-dimensional vector

$$\boldsymbol{\theta} = [\theta_0, \theta_1, \theta_2....]$$

Because our model is linear, it can always be written in the matrix form

$$\boldsymbol{b} = \boldsymbol{A} \cdot \boldsymbol{\theta}$$

In our polynomial example, we will have

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ .. \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & .. \\ 1 & x_2^2 & x_2^2 .. & \\ .. & .. & .. & \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ ... \end{bmatrix}$$

So our model will then be

$$\boldsymbol{b} = \boldsymbol{A} \cdot \boldsymbol{\theta} + \boldsymbol{n}$$

where $\boldsymbol{n}$ is a vector of noise realizations.

If now we assume that the noise realization in the different data points are independent and that the realizations are drawn from a Gaussian distribution with variance σ^2, then the log of the likelihood becomes

$$\mathcal{L} = \text{constant} - \frac{1}{2\sigma^2} \left| \boldsymbol{A} \cdot \boldsymbol{\theta} - \boldsymbol{b} \right|^2$$

where the norm of a vector, $|\boldsymbol{x}|$ should be understood as the length of the vector squared:

$$|\boldsymbol{x}| = \sum_i x_i^2,$$

and, as before, 'constant' refers to all terms that do not depend on $\boldsymbol{\theta}$. Notice that if the priors are uniform then maximizing the likelihood is the same as maximizing the posterior.

Maximizing the log-likelihood is equivalent to minimizing the sum of squares of the difference between the data and the model, hence this method is also called least-squares estimation or chi-squared minimization. The latter expression comes about because people use the Greek symbol χ^2 to denote the following sum over all data points

$$\chi^2 = \sum_k \frac{(\text{data}_k - \text{model}_k)^2}{\sigma_k^2}$$

where σ_k^2 is the variance of the noise in the k^{th} data point.

Anyways, let us return to our matrix formulation for linear problems with Gaussian errors. The log-likelihood is maximized at the extrenum point $\hat{\boldsymbol{\theta}}$ given by

$$\nabla_{\boldsymbol{\theta}} L = 0 \implies \boldsymbol{A} \cdot \hat{\boldsymbol{\theta}} - \boldsymbol{b} = \boldsymbol{0}$$

The maximum likelihood estimate is therefore

$$\hat{\boldsymbol{\theta}} = \boldsymbol{A}^{-1} \cdot \boldsymbol{b}$$

The covariance matrix of the parameters is given by

$$\sigma^2(\boldsymbol{\theta}) = \left[\nabla^2_{\boldsymbol{\theta}} \mathcal{L} \right]^{-1} = \sigma^2 \left[\boldsymbol{A}^T \cdot \boldsymbol{A} \right]^{-1}$$

Chapter 5

Non-linear parameter estimation

In the previous chapter we saw that if the errors in our measurement are Gaussian and if the model we have for our data is linear in its parameters, then a relatively straightforward expression can be derived for the most likely values for the parameters and their uncertainties. This is of great value in multi-parameter estimation problems where a brute force calculation of the posterior is prohibitively expensive. We are now prepared to tackle the next level of complexity: what is the model is non-linear in its parameters? Of course, new techniques have to be developed to deal with this— the subject of this chapter. We will learn the most common tricks: (a) linearizing the model, (b) iterative algorithms to find the peak of the likelihood, and (c) clever Monte-Carlo methods to sample the posterior instead of brute-force evaluations.

5.1 Complications of non-linearity

To appreciate the complexity of non-linear models, let us take a common example encountered in radio astronomy. Suppose you measure the flux density S of an astrophysical source at different radio frequencies, ν_i. Of course your detector will introduce some noise and therefore uncertainty in your measurement. Two types of emission are commonly encountered in radio sources: (a) synchrotron emission that gives power-law spectra of the form $S(\nu) \propto \nu^{\alpha}$ where α is typically negative and in the vicinity of -1 and (b) thermal emission which gives rise to a flat spectrum with $\alpha = 0$. Suppose you found the data show in Fig. 5.1 upon measuring the spectrum of a source at different radio frequencies. You now have a crucial question to answer before you know the true nature of the source. Is this thermal emission or synchrotron emission? The conclusions of the paper you are writing depend on it. What to do?

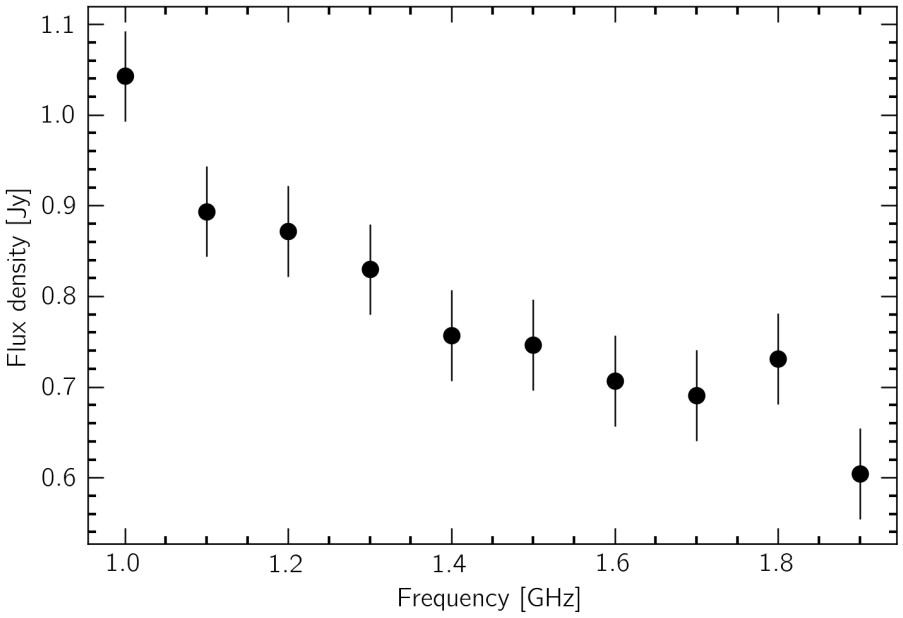


Figure 5.1: Mock measurements of the radio spectrum of an astrophysical source. Is this source showing thermal or synchrotron emission?

Well obviously thermal emission is a special case where $\alpha = 0$ so the problem is that of estimating α and seeing if very low values of α can be ruled out with some confidence. But you immediately notice the problem:

the model is non-linear in its parameters:

$$S_k = C\,\nu_k^\alpha + n_k$$

where C and α are parameters to be estimated, S_k, ν_k are the data and n_k are the noise realisations. Let us assume that the noise realisations are independent and Gaussian distributed with zero mean and variance σ^2. The log-likelihood is then

$$\mathcal{L} = \text{constant} - \frac{1}{2\sigma^2}\sum_k [S_k - C\nu_k^\alpha]^2$$

OK, we need to find the maxima of this function. So we will evaluate the partial derivatives and set them to zero.

$$\frac{\partial \mathcal{L}}{\partial C} = 0 = \sum_k \nu^\alpha (S_k - C\nu^\alpha)$$

and

$$\frac{\partial \mathcal{L}}{\partial \alpha} = 0 = \sum_k (S_k - C\nu_k^\alpha) C\nu^\alpha \log \nu$$

And now you are majorly stuck because these simultaneous equations but they are not *linear* simultaneous equations. So all the tricks with matrix notation and cute 'n' easy linear algebra solutions are out of the window. What to do now?

Well what did you do before in life when you had to find the roots of a non-linear equation? Say you had to find x such that $f(x) = (x-1)\sin x - 1 = 0$. Does Netwon's method ring a bell? This is a way to Taylor expand the equation and use that to set up an iterative algorithm. So if you have to solve $f(x) = 0$, you write

$$f(x) \approx f(x_0) + (x - x_0) f'(x_0)$$

and used that to iterate your way to the solution:

$$x_i = x_{i-1} - \frac{f(x_i)}{f'(x_i)}$$

In other words, you 'linearlized' the function at your best guess values of x and used the linear representation to find the root. Then you used this root as your new guess to repeat the whole procedure.

We can do the same with messy non-linear models. This procedure gives is a family of algorithms called gradient decent algorithms. Why the name? Because the gradient is just the name for the derivative taken in multiple dimensions. Just like Newton's method for the roots of a function used the first derivative to guide the iterations towards the true value, so does the multi-dimensional case use the gradient to guide the iterations along the path of steepest descent towards the minima of the function.

5.2 Gradient descent techniques

The linearised version of our model with our initial guess C_0, α_0 is

$$S_k = C(\nu_k^{\alpha_0}) + \alpha(C_0\nu_k^{\alpha_0}\log\nu_k) - \alpha_0 C_0\nu_k^{\alpha_0}\log\nu_k + n_k$$

Our log-likelihood is then

$$\mathcal{L} = \text{constant} - \frac{1}{2\sigma^2}\sum_k \left[S_k - C(\nu_k^{\alpha_0}) - \alpha(C_0\nu_k^{\alpha_0}\log\nu_k) + \alpha_0 C_0\nu_k^{\alpha_0}\log\nu_k\right]^2$$

As before to maximize the log-likelihood, we set the partial derivatives w.r.t to the parameters to zero. Setting the derivative w.r.t C to zero, one gets

$$\frac{\partial\mathcal{L}}{\partial C} = 0 = \sum_k \nu_k^{\alpha_0}\left[S_k - C(\nu_k^{\alpha_0}) - \alpha(C_0\nu_k^{\alpha_0}\log\nu_k) + \alpha_0 C_0\nu_k^{\alpha_0}\log\nu_k\right]$$

Setting the partial derivative w.r.t α to zero gives

$$\frac{\partial\mathcal{L}}{\partial C} = 0 = \sum_k C_0\nu_k^{\alpha_0}\log\nu_k\left[S_k - C(\nu_k^{\alpha_0}) - \alpha(C_0\nu_k^{\alpha_0}\log\nu_k) + \alpha_0 C_0\nu^{\alpha_0}\log\nu_k\right]$$

These equations are now linear in the parameters α and C, we can solve them with the usual matrix algebra tools. But this is not the end! Once we get our most likely estimates, we have to go back and linearlize the model at these new values substituted for α_0 and C_0 and repeat. If all goes to plan, then with each iteration, we get closer and closer to the true solution as Newton stated.

The linear simultaneous equations for C, α in matrix notation are

$$\begin{bmatrix} \sum_k \nu_k^{2\alpha_0} & \sum_k C_0\nu_k^{2\alpha_0}\log\nu_k \\ \sum_k C_0\nu_k^{2\alpha_0}\log\nu_k & \sum_k (C_0\nu_k^{\alpha_0}\log\nu_k)^2 \end{bmatrix} \cdot \begin{bmatrix} C \\ \alpha \end{bmatrix} = \begin{bmatrix} \sum_k \nu_k^{\alpha_0}(S_k + \alpha_0 C_0\nu_k^{\alpha_0}\log\nu_k) \\ \sum_k C_0\nu_k^{\alpha_0}\log\nu_k(S_k + \alpha_0 C_0\nu_k^{\alpha_0}\log\nu_k) \end{bmatrix}$$

Wow! That was cumbersome, but we got it done! Let us check if this works out as planned with the dataset in Fig. 5.1 To really test how well our algorithm does, let us start with some values of C and α that are way off. Say we start with $C_0 = 10$ and $\alpha_0 = 2.9$. Here are the outputs of the first 10 iterations

```
Iteration 1: C=0.6771, alpha=2.8044
Iteration 2: C=0.6939, alpha=1.3657
Iteration 3: C=0.9140, alpha=-0.1861
Iteration 4: C=1.0261, alpha=-0.7278
Iteration 5: C=1.0301, alpha=-0.7430
Iteration 6: C=1.0299, alpha=-0.7426
Iteration 7: C=1.0299, alpha=-0.7426
```

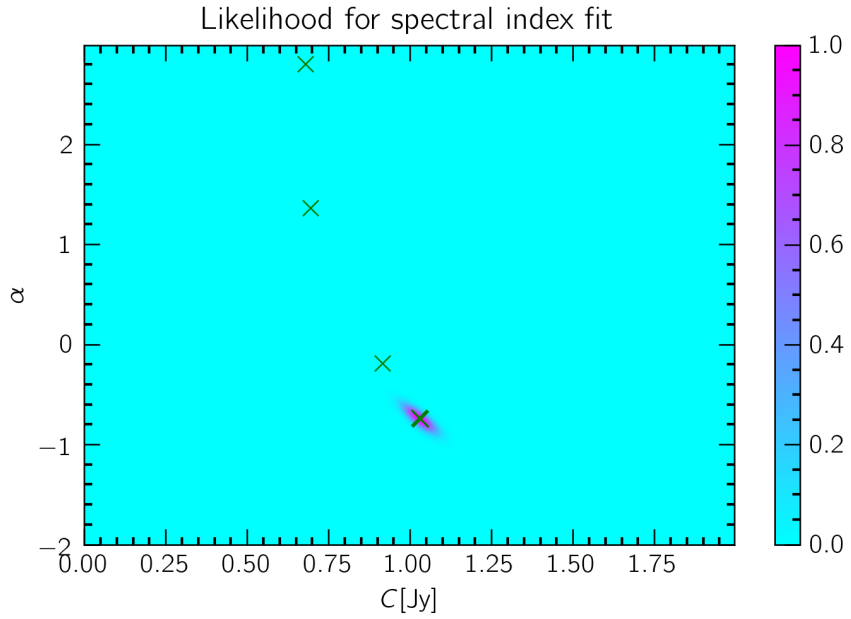


Figure 5.2: Likelihood of the flux density scale and spectral index. The crosses show the successive iterations of Newton's method converging to the peak of the likelihood.

Wow that converged quite quickly. Let us now see if they converged to the actual peak of the likelihood by brute-force computing the likelihood. The result is shown in Fig. 5.2.

An intuitive way to think about this is as follows. Imagine that the likelihood function is a mountain. You want to reach the peak. You start somewhere, even if it is way down at the base of the mountain. You then locate in which direction the peak is. You do this by computing the derivatives (gradient in this case). The gradient points you in the direction where the mountain is the steepest (as this gives the highest derivative/slope). You then proceed along this direction and keep doing this till you reach the peak. If the problem is cast in terms of least squares then you are trying to reach the minima and you just proceed in the direction of steepest descent. Note that the likelihood is the exponential of the negative of the sum of squares of model-data!

By the way, You can marginalize over C to calculate the posterior of the spectral index. I think in this problem we expect your conclusion to be that it is a synchrotron source. But to know the confidence with which you can say that you will have to numerically compute the confidence intervals.

Alternatively, you could calculate the second derivatives of $\mathcal{L}$ and find the find the co-variance matrix just as we did before.

OK that worked but it was a lot of painful algebra. Perhaps we can write things out in a generic but compact matrix notation to make our lives easier? And because we have gained some familiarity with matrix notation, let us use that and get even better at it. We have been given a model that is non-linear in the parameters $\boldsymbol{\theta}$. For any choice of parameters we can compute the log-likelihood $\mathcal{L}(\boldsymbol{\theta})$. The problem is to find $\boldsymbol{\theta}$ that maximizes $\mathcal{L}$. Let us Taylor expand $\mathcal{L}$ around our initial guess for the parameters $\boldsymbol{\theta}_0$:

$$\mathcal{L}(\boldsymbol{\theta}) \approx \mathcal{L}(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \nabla \mathcal{L}(\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \nabla^2 \mathcal{L}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + ...$$

Let us take the gradient of $\mathcal{L}$ and set it to zero. The first term on the R.H.S is not a function of $\boldsymbol{\theta}$ so its gradient w.r.t $\boldsymbol{\theta}$ is zero. The second term on the R.H.S is $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \nabla \mathcal{L}(\boldsymbol{\theta}_0)$. Its gradient can be computed from the product rule: $\nabla(\boldsymbol{a} \cdot \boldsymbol{b}) = \boldsymbol{a}\nabla \cdot \boldsymbol{b} + (\nabla \boldsymbol{a}) \cdot \boldsymbol{b}$. Notice also that the gradient of $\mathcal{L}(\boldsymbol{\theta}_0) = \mathbf{0}$, the null vector, as $\mathcal{L}(\boldsymbol{\theta}_0)$ it is not a function of $\boldsymbol{\theta}$ and $\nabla(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \mathbf{1}$, the unit vector.

$$\nabla \left[(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \nabla \mathcal{L}(\boldsymbol{\theta}_0) \right] = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \cdot \mathbf{0} + \mathbf{1} \cdot \nabla \mathcal{L}(\boldsymbol{\theta}_0) = \nabla \mathcal{L}(\boldsymbol{\theta}_0)$$

Using the product rule repeatedly, we can write down the gradient of the third term on the R.H.S. We get

$$\nabla \left[\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \nabla^2 \mathcal{L}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right] = \frac{1}{2} \left[(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \nabla^2 \mathcal{L}(\boldsymbol{\theta}_0) \right]$$

$$+ \frac{1}{2} \left[\nabla^2 \mathcal{L}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right]$$

Because $\mathcal{L}$ is a continuous function, we have $\partial^2 / \partial\theta_i \partial\theta_j = \partial^2 / \partial\theta_j \partial\theta_i$ This mean that $\nabla^2 \mathcal{L}$ is symmetric about its principal diagonal. In this case, the two terms on the R.H.S of the last equation are the same. So we have the gradient of $\mathcal{L}$ to be

$$\nabla \mathcal{L}(\boldsymbol{\theta}) \approx \nabla \mathcal{L}(\boldsymbol{\theta}_0) + \nabla^2 \mathcal{L}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

We can now set up an iterative scheme as follows:

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 - \nabla \mathcal{L}(\boldsymbol{\theta}_0) \left[\nabla^2 \mathcal{L}(\boldsymbol{\theta}_0) \right]^{-1}$$

Once the iterations have converged, the most likely values for the parameters, $\hat{\boldsymbol{\theta}}$ are found. The covariance matrix of the parameters that gives their uncertainties is then given by

$$\left\langle (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \right\rangle = \left[\nabla^2 \mathcal{L}(\hat{\boldsymbol{\theta}}) \right]^{-1}$$

The above equation can be applied to any model. Sometimes it is analytically very cumbersome to calculate the partial derivatives. We already saw that with the simple model for the radio spectrum. So the derivatives of the log-likelihood are calculated numerically. It is worth noting here that the two partial derivatives are so commonly used that they have their own names: $\nabla\mathcal{L}$ is called the Jacobian matrix and $\nabla^2\mathcal{L}$ is called the Hessian matrix.

5.3 Complex likelihood shapes

We saw in the previous section that the linearisation scheme can be used to solve any estimation problem, even if the model is non-linear. But just like Newton's method can have some sticky issues in root finding, so can our scheme; after all, they use the same iteration scheme. To illustrate this aspect of gradient methods, consider another common problem in astronomy: finding the presence of an exoplanet by detecting the reflex motion of its host star. If the orbit is circular then the radial velocities of the star will show a sinusoidal variation at the orbital frequency of the planet. Let us consider the mock data shown in Fig. 5.3[h]

The model we have for the data is a sinusoid. This means that we have to estimate 3 parameters: the amplitude of the sinusoid A, the frequency of the sinusoid P and the phase of the sinusoid ϕ at time $t = 0$. So our model is

$$v_k = A\cos(2\pi P t_k + \phi) + n_k$$

where n_k is the noise. We will assume that the noise is Gaussian distributed with known variance σ^2 and zero mean.

The parameter vector is

$$\boldsymbol{\theta} = \begin{bmatrix} A \\ P \\ \phi \end{bmatrix}$$

And with flat priors on the parameters our log-posterior is proportional to the log-liklelihood, which is

$$\mathcal{L} = \text{constant} - \frac{1}{2\sigma^2}\sum_k [v_k - A\cos(2\pi P t_k + \phi)]^2$$

It is clear that the likelihood is not linear in the parameters. So we must use our iterative scheme. There are three parameters which means that we must compute 3 partial first derivatives and 6 partial second derivatives. These can be done, no problem, but I chose this problem to illustrate a common problem with gradient algorithms. So I don't want us to get stuck

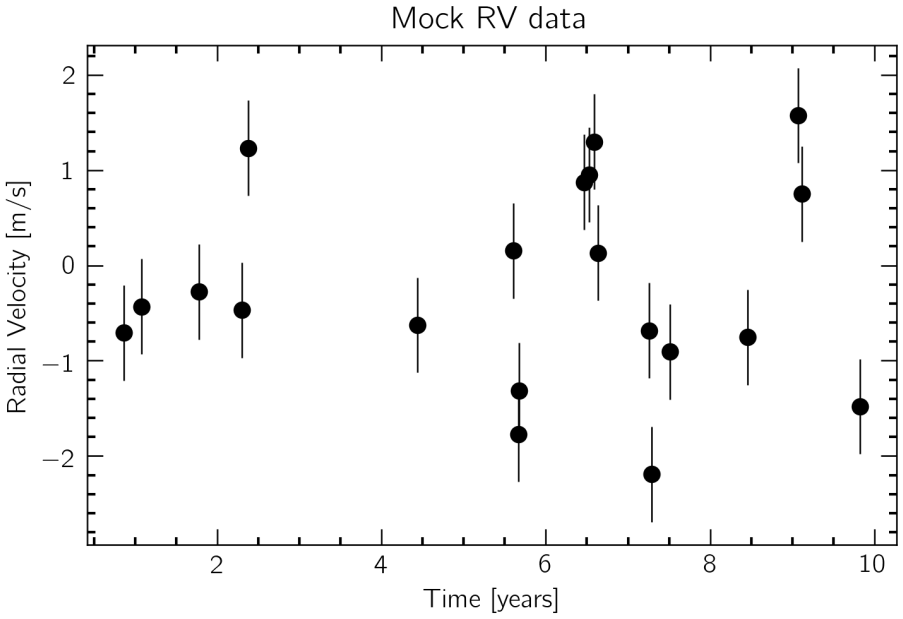


Figure 5.3: Mock RV data spanning 10 years. The RV shifts are caused by the presence of an exoplanet. The problem is to estimate the amplitude of the signal, the frequency and starting phase at time $t = 0$.

with the algebra and miss the main point. So for now, let me give you the true values of the amplitude and starting phase I used to generate this data: $A = 1$ and $\phi = 0$. So we can focus on estimating the orbital frequency which will turn out to be the tricky bit!

We now only have to evaluate the partial derivatives:

$$\frac{\partial \mathcal{L}}{\partial P} = -\sigma^{-2} \sum_k 2\pi t_k \sin(2\pi P t_k) \left[v_k - \cos(2\pi P t_k) \right]$$

and

$$\frac{\partial^2 \mathcal{L}}{\partial P^2} = -\sigma^{-2} \sum_k (2\pi t_k)^2 \sin^2(2\pi P t_k) + (2\pi t_k)^2 \cos(2\pi t_k P)(v_k - \cos(2\pi P t_k))$$

With these, we can set up our iteration scheme

$$P_i = P_{i-1} - \frac{\partial \mathcal{L}}{\partial P} \left[\frac{\partial^2 \mathcal{L}}{\partial P^2} \right]^{-1}$$

In Fig. 5.4 I have plotted the best estimates for the orbital frequency as a function of iteration number. The different curves are for different initial guesses. I have also shown the final χ^2 value defined as

$$\chi^2 = \sigma^{-2} \sum_k [v_k - \cos(2\pi P t_k)]^2.$$

It is clear that our scheme has failed; and miserably at that! How can

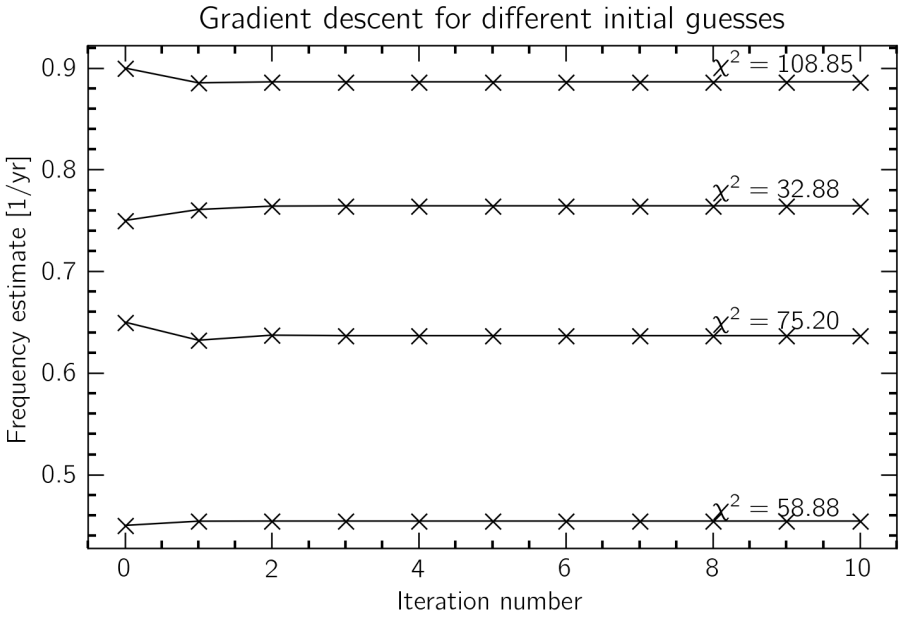


Figure 5.4: Iterative solution for the orbital frequency for different initial guess values.

the 'true' value found by the algorithm depend on the initial estimate? It should not. Also look at the χ^2 values. Clearly, it is lowest when the values converge to a period around $0.76\,\mathrm{year}^{-1}$. But why did the algorithm not always find this value? Why did it get stuck at solutions with such high χ^2 values?

To understand what is going on, we must brute-force compute the curve of $\mathcal{L}$ versus P. I have done this in Fig. 5.5 and the reason for our failure is now obvious.

The likelihood space has a lot of peaks and valleys. The maximum does occur around $P = 0.76\,\mathrm{year}^{-1}$, sure, but imagine a starting value that is not very close to this true value. Suppose we start at an initial guess of 0.9. The algorithm senses that the peak is to the left of 0.9 so it ends up

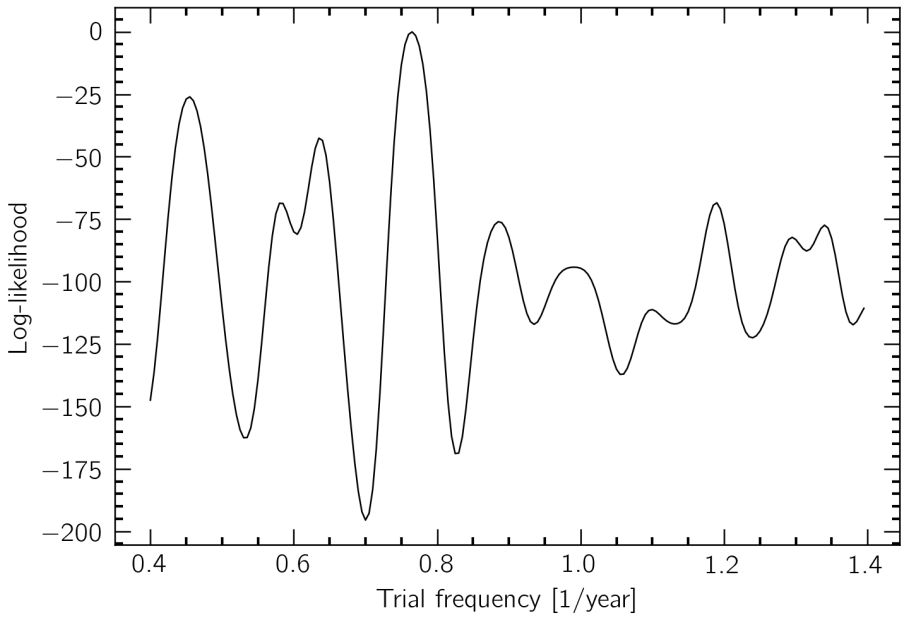


Figure 5.5: Brute force calculation of the log-likelihood for different orbital frequency trials showing the complex likelihood space.

climbing to the peak just to the left of 0.9. To the left of this peak is a valley and the highest peak is to the left of this valley. The algorithms, by design, decides that it would rather stay there than descent down towards areas of lower log-likelihood. Because the algorithm only uses the derivatives of the function computed where it is, i.e locally, it cannot know the existence of an even higher peak at 0.76. It gets stuck in the wrong place.

It is now clear that this gradient technique only works when we have a well-behaved likelihood shape. By well behaved we mean that the likelihood has a single peak and no valleys. Fortunately a lot of problems do fall in this category. I chose period/frequency finding which is quite a basket-case because the periodic nature of the sinusoids creates a 'wavey' likelihood with many smaller peaks and valleys. Nevertheless, we should always be careful to check if our algorithm arrives at the same value regardless of what the initial guess it.

How do I solve problems with funky likelihood shapes, you may be wondering. Let us learn some new techniques to deal with these problems and then return to our radial velocity exoplanet data to do a simultaneous fit of all three parameters in all its glory.

5.4 Monte-Carlo techniques

We will now study the basics of a whole bunch of new tricks to calculate the posterior distributions in estimation problems. As you will find out, these tricks are remarkably better than gradient techniques in handling complex posterior shapes, and they also handle higher dimensional optimisations quite well. But this will require us to look at the problem of posterior computation from a totally different angle. So we will have to go back to some basic ideas of probability.

What does a place known for gambling have to do with solving estimation problems with complex likelihood shape? The name is a reference to Stanisław Ulam's uncle whose compulsive gambling habit in Monte Carlo was his ruin. Stanisław fared somewhat better in life. He fled Poland mere days before its invasion in 1939, arrived in the USA and ended up working at Los Alamos on the hydrogen bomb. Anyways, Stanisław's essential insight can be highlighted by a common probability calculation involving birthdays (he used the example of Solitaire; but I have forgotten the game's rules, so I am using this instead).

Suppose there are 30 students in a classroom. What is the probability that at least two students share a birthday? The way you have been taught to handle this problem is as follows: Use algebra to find all the possible birthday combinations. Then do more algebra to find all the possible ways two or more people can have the same birthday. Divide one by the other and you have your answer. In fact, you are stuck with this method in the absence of computers. Stanisław's essential insight was that computers open a new door. Here is what he would have suggested. You write code to draw 30 random natural numbers less than 366 and check if there are any duplicates. You repeat this some large number of times (say 10^5) to see how often duplicates appear. Voila! You have a good estimate of the probability. No algebra needed! I wrote the following code to do this and got 70% probability which is very close to the correct value[1].

```
import numpy as np
nmc = 100000 # Number of Monte-Carlo runs
N = 30 # Number of students
yes=0 # counter
for i in range(nmc):
   bdays = np.ceil(np.random.rand(N)*365).astype(int)
   unique_bdays = list(set(bdays)) # set extracts unique elements
   if len(unique_bdays)<N:
```

[1] I will leave it to you to solve it with algebra. Hint: It is easier to compute the probability that no two people share a birthday.

```
      yes+=1
print (yes/nmc)
```

Notice what we actually did here? We used a random number generator to compute the value of an algebraic expression. This is what is referred to as a 'Monte-Carlo' method. OK what has this got to do with ill-behaved posterior distribution? Notice that the posterior distribution in a non-linear estimation problem is actually an algebraic expression, so we should try to find a clever algorithm that can use the random number generator on a computer to find the shape of the posterior curve.

Consider the following algorithm to take your hiker trying to climb the posterior peak to the next level. Suppose the hiker starts at some location $\boldsymbol{\theta}_0$ but can teleport to any location in the parameter space. Pick a random proposed location $\boldsymbol{\theta}_p$ drawn from the prior distribution. Compute the likelihoods at the current location and the proposed location, $L(\boldsymbol{\theta}_0)$ and $L(\boldsymbol{\theta}_p)$ (note: this is not the log-likelihood). Hop to the proposed location with a probability of $\min[1, L(\boldsymbol{\theta}_p)/L(\boldsymbol{\theta}_0)]$ that is, always hop to $\boldsymbol{\theta}_p$ if its likelihood is higher, but only sometimes hop if the likelihood is lower. Keep doing these hops for a large number of steps, N_{mc}. That's it! The number of times you visited some location in the parameter space is proportional to the posterior probability at that location. You can intuitively see why that is the case. The probability of hopping to some location in this scheme depends on two things: (a) how often this location is proposed which depends on how high the prior probability is that location and (b) how often this proposal is accepted which depends on how high the likelihood is at this location. The product of the prior and the posterior probabilities is proportional to the posterior and therefore the histogram of your visits is an approximation of the posterior distribution. What's more, the inherent randomness in your proposals insures against getting stuck in some local peak like the gradient algorithm would. So as long as you have a powerful enough computer to carry out this procedure for a sufficiently large N_{mc} you will find a good-enough approximation for the posterior distribution.

Intuition aside, why does this algorithm approximate the shape of the posterior? Consider two locations $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ and without loss of generality, let us say that $L(\boldsymbol{\theta}_B) > L(\boldsymbol{\theta}_A)$. Let the fraction of visits to $\boldsymbol{\theta}_A$ be f_A and the fraction of visits to $\boldsymbol{\theta}_B$ be f_B. Now the $A \rightarrow B$ hop is always accepted but the reverse hop is accepted with a probability of $L(\boldsymbol{\theta}_A)/L(\boldsymbol{\theta}_B)$. Because we are using the prior distribution $P(\boldsymbol{\theta})$ for hop-proposals, B is proposed at a rate proportional to $P(\boldsymbol{\theta}_B)$ and A is proposed at a rate proportional to $P(\boldsymbol{\theta}_B)$. The rate of hops from $A \rightarrow B$ is therefore $f_A \times P(\boldsymbol{\theta}_B)$ and the rate of hops from $B \rightarrow A$ is $f_B L(\boldsymbol{\theta}_A)/L(\boldsymbol{\theta}_B) \times P(\boldsymbol{\theta}_A)$. If we run our algorithm for a sufficiently long time, then we will reach an equilibrium where the

fractions f_A and f_B do not change. That is, the rate of hops from A to B are the same as the rate of hops from B to A. That is when

$$f_A \times P(\boldsymbol{\theta}_B) = f_B \times L(\boldsymbol{\theta}_A)/L(\boldsymbol{\theta}_B) \times P(\boldsymbol{\theta}_A)$$

or when

$$\frac{f_A}{f_B} = \frac{L(\boldsymbol{\theta}_A)P(\boldsymbol{\theta}_A)}{L(\boldsymbol{\theta}_B)P(\boldsymbol{\theta}_B)}.$$

In order to reach this equilibrium, transitions between any two points in the posterior parameter space must be permitted, which our algorithms does. Nice! This algorithm is commonly referred to as the Metropolis-Hasting algorithm.

Armed with this knowledge, let us boldly return to our orbital frequency estimation problem. Here is a simple algorithm that I implemented on a computer by assuming a uniform prior on the frequency between $0.4\,\mathrm{yr}^{-1}$ and $1.4\,\mathrm{yr}^{-1}$.

```
for each Monte-Carlo iteration:
   propose new location drawn from Uniform[0.4,1.4]
   compute likelihood L at proposed location
   draw a random number u from Uniform[0,1]
   if u <= (new likelihood/old likelihood):
      hop to propose location
   else:
      stay in current location
```

This gives an array of locations in orbital-frequency space shown in Fig. 5.6 and the histogram of these values is given in Fig. 5.7.

Let us understand these outputs. The iterations start at an initial value that I randomly gave to be $0.2\,\mathrm{yr}^{-1}$. The random proposals are usually accepted when they go towards higher likelihood values, so the hops gradually bring out teleporting hiker toward the highest peak around $0.76\,\mathrm{yr}^{-1}$. Once our hiker has reached the area around this peak, most of the random suggestions that bring our hiker far down the peak end up getting rejected. However suggestions that bring our hiker not too far below the peak have a decent chance of being accepted. The upshot is that our hiker spends her time hopping around the peak thereby sampling the shape of the posterior around the peak. The initial period when the hiker is randomly hopping until she arrives in the vicinity of the peak is called the "burn-in" or "warm-up" period. Because we are mainly interested in the location and shape of the peak, we ignore this burn-in period and make a histogram of the remaining values. Just as we anticipated this histogram very well reproduces the actual shape of the posterior with the advantage that it reached the 'correct' and highest peak.

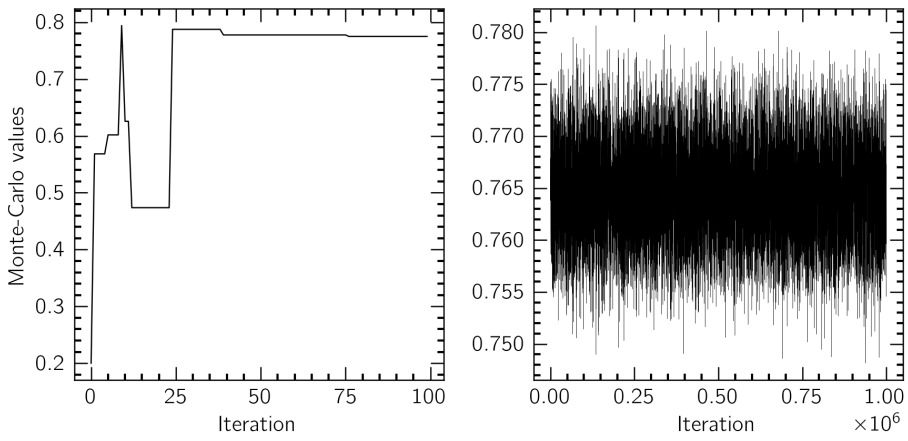


Figure 5.6: Monte Carlo values for the orbital frequency. Left panel shows the initial samples finding their way towards the peak. The right panel shows the bulk of the later samples which are samples drawn from the posterior distribution around the global peak.

5.4.1 Markov Chain Monte Carlo

Over the years, many additions and improvements have been made to the Metropolis-Hastings algorithm. An important improvement is the use of a sampling function. In our implementation, we randomly sampled the posterior to make a suggestion for a hop. This can become inefficient after the burn-in period as most suggestions will be far away from the peak, have very low likelihood, and therefore be rejected. This was no much of a limitation in the one-dimensional parameter example we solved above. But as the number of dimensions increases, the likelihood space one has to search increases so rapidly that the rejection-ratio begins the cripple the algorithm. The efficiency of sampling the posterior around the peak can be improved by using a sampling function that privileges shorter hops. This is done by drawing samples from a new distribution, called the proposal distribution, say $q(\boldsymbol{\theta}_p|\boldsymbol{\theta})$ which is the probability that a new proposal $\boldsymbol{\theta}_p$ will be made given the current location of the sampler $\boldsymbol{\theta}$. It is important to choose a density that is symmetric, that is, $q(\boldsymbol{\theta}_p|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\theta}_p)$ such that the proposal does not modify the equilibrium condition that is at the heart of the algorithm. It is common to use Gaussian proposal densities as a simple starting point. Deciding the width of the proposal sometimes has to be done with trial and error. Too small a width, and the sampler will tend to get stuck around the first peak it finds. Too broad a proposal and we return to

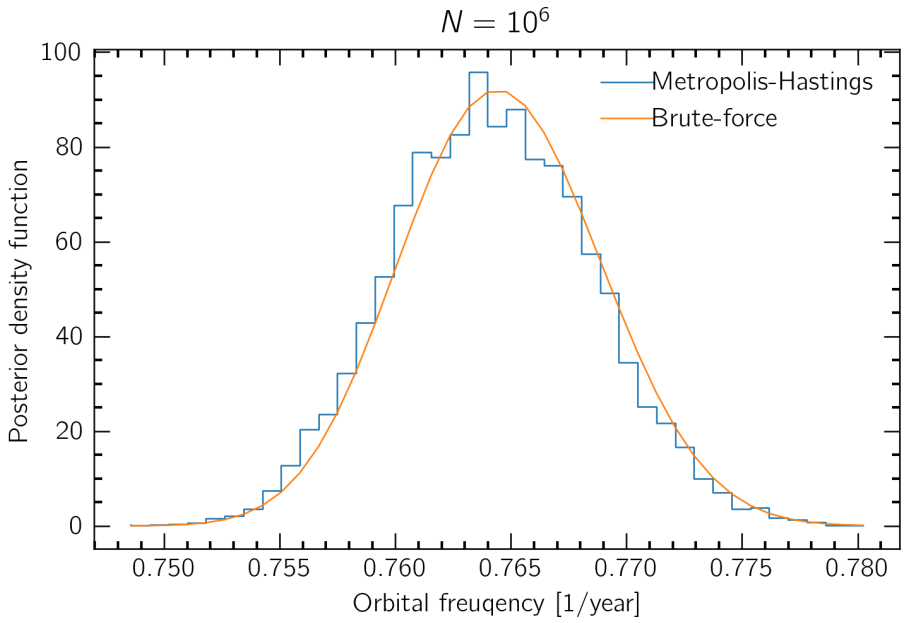


Figure 5.7: Histogram of the posterior samples after the burn-in period from Fig. 5.6. Also plotted is the brute-force computed posterior distribution which is the same as the curve in Fig. 5.5.

our initial problem of an unacceptably high rejection ratio. In any case, the introduction of a proposal distribution now makes any samples statistically dependent on the previous value of the sample. This dependency between the various 'states a system can be in' is usually referred to as a Markov chain and the therefore, this technique of posterior estimation is called the 'Markov Chain Monte Carlo' method or MCMC.

Let us use a Gaussian proposal density to solve our original 3-dimensional problem of estimating the amplitude, phase and orbital-frequency of our radial velocity problem. Here is what our new algorithm looks like with a proposal distribution that has a standard deviation of 0.1 along all three axes.

```
Start with some initial estimate of A0, P0 and phi0
Calculate the likelihood at this initial estimate
for each Monte-Carlo iteration:
   draw a sample from
   [Normal(A0,0.1), Normal(P0,0.1), Normal(phi0,0.1)]
   calculate the likelihood at the proposed location
```

```
draw a random number u from Uniform[0,1]
if u<= (new likelihood/old likelihood):
   hop to proposed location
else:
   stay in current location
```

With N such Monte Carlo iterations we will now get a matrix of size $3 \times N$ with the samples. The histogram of these samples will approximate the posterior distribution in 3-dimensions. Fig. 5.8 shows the samples after the burn-in period which I had to now extent to the first 1000 samples because of the increase in the dimensionality of the problem.

Unfortunately it is not easy to display a 3-dimensional distribution on a sheet of paper (or computed screen) in an easy-to-understand quantitative manner. So there exists some really handy tools for visualisation and marginalisation in cases with high-dimensionality. I have used the python tool `corner` for this and generated Fig. 5.9

The figure displays marginalised density functions. For example, the plots along the upper diagonal display the 1-dimensional distributions after marginalisation. The top plot for the parameter A is the result of the following marginalisation (implemented numerically of course)

$$\text{Prob}(A|\text{data}) = \int \text{d}\phi \int \text{d}P \ \text{Prob}(A, P, \phi|\text{data}),$$

and so on. The dashed vertical lines show the 16^{th}, 50^{th} and 84^{th} percentile values. If the posterior is approximated as a Gaussian, then these correspond to the mean value and $\pm 1\sigma$ deviation on either side of the mean. In other words, for a Gaussian distribution, 16% of he samples are lower than the mean minus one standard deviation and 84% of he samples are smaller than the mean plus one standard deviation. So the percentiles can be understood as the $\pm 1\sigma$ uncertainty range. As for the panels in the Figure with a Gray colormap, these show the marginalised two-dimension posterior. So for instance, the middle panel of the left-most column is evaluating

$$\text{Prob}(P, A|\text{data}) = \int \text{d}\phi \text{Prob}(A, P, \phi|\text{data}),$$

and so on. In the central regions of the panel where the density of sample points is high, the sample density (which is proportional to the probability density) is shown as a colormap with contours. away from the central region where the sample density is low, the samples are depicted as points.

It is always important to check a few things when one does MCMC calculation of the posterior:
(1) Check if you are merely sampling from the prior: This can happen

when the likelihood is far broader than the prior or in other words when the new data does not bring more information on the parameters than what the prior already did. In this case the posterior width with encompass the prior width. This just means that your data is not very useful; tough!
(2) Check if the posterior distribution depends on the starting values for the parameters (at the beginning of the burn-in period). If it does, then sampler is likely getting stuck at local peaks. Consider increasing the width of the proposal distribution and/or the number of Monte Carlo iterations.
(3) Check if the posterior distribution is sufficiently smooth. If not then it could mean that you are not running enough iterations given the dimensionality of your problem.
(4) Check if the posterior peak is too close to the limit allowed by the prior. This would mean that the data is telling you that the parameter's likely value is inconsistent with the prior. In this case, the sampler has done its job; you need to go back to the drawing board and examine why this is happening.
(5) Finally, always check how well the parameters fit the original data. This is your last line of defence against getting stuck on some unwanted peak in an unwanted corner of a high-dimensional likelihood space. Let us end this chapter by doing just that with the radial velocity data. In fig. 5.10 I have drawn 100 random samples from the MCMC sampler output and computed the model for these samples and over-plotted them on the original data.

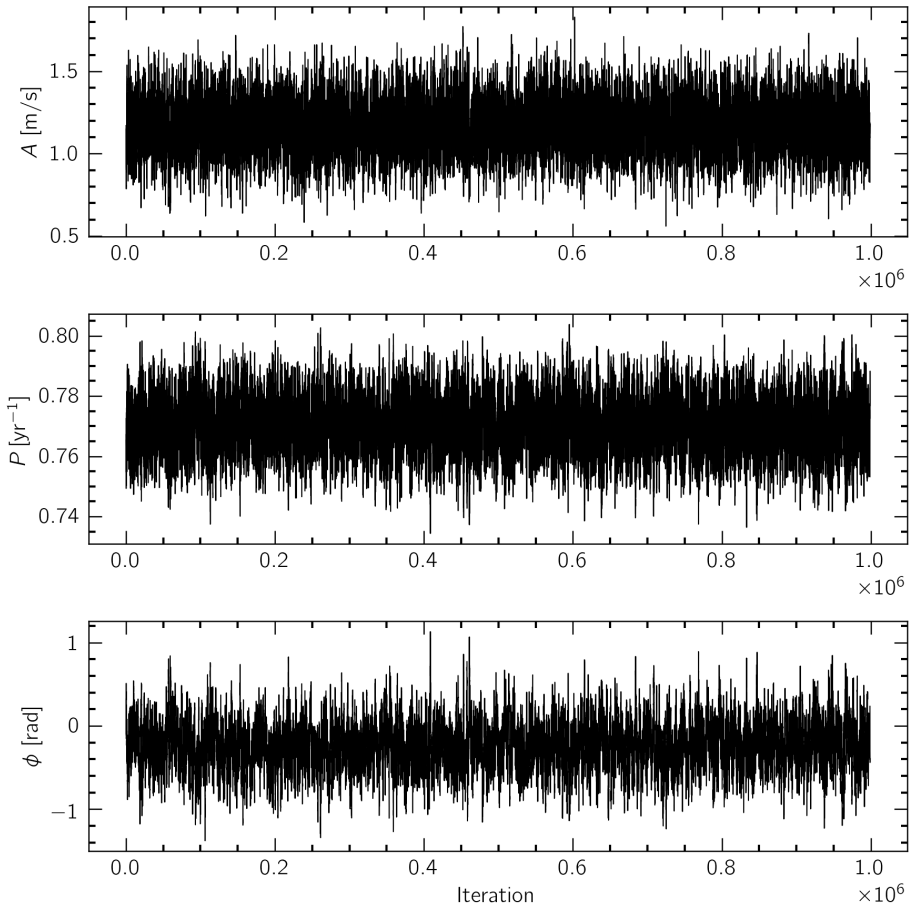


Figure 5.8: Markov Chain Monte Carlo samples of the posterior distribution for the radial velocity problem. The burn in period of 1000 samples has been excluded. Due to the absence of abrupt and large jumps, in the values of the parameters, the plot shows that the sampler has identified a peak and is sampling the shape of the 3-dimensional posterior in the vicinity of the peak.

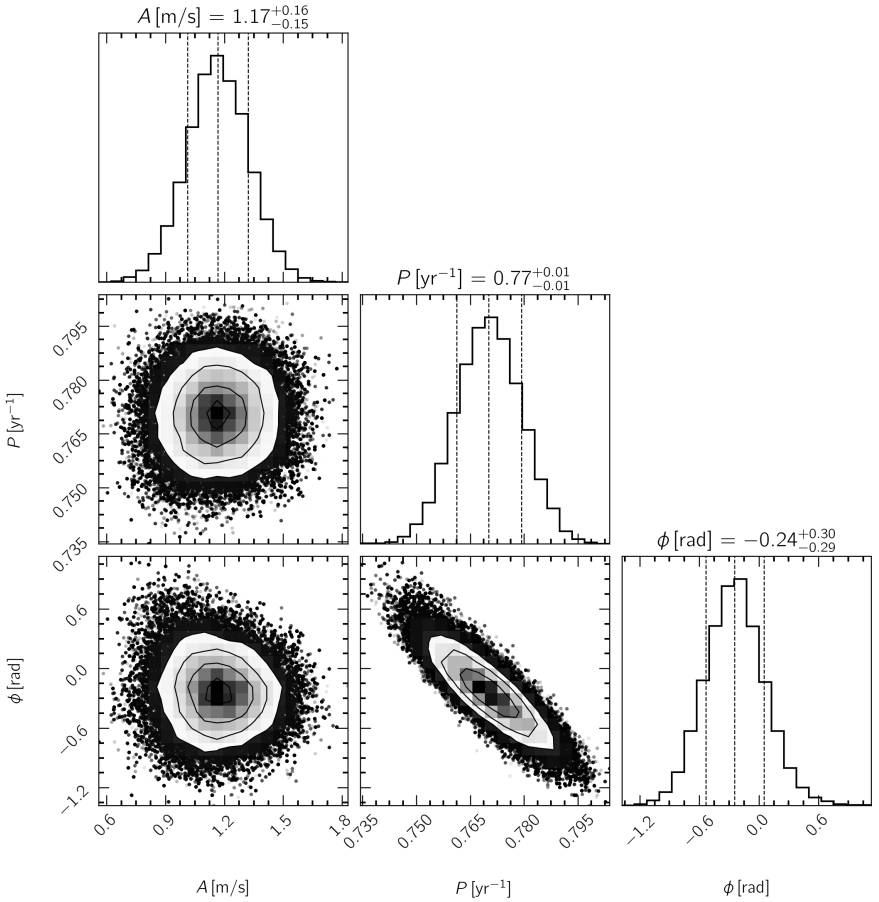


Figure 5.9: Posterior distribution of the parameters constructed from the samples in Fig. 5.8 using the package `corner`.

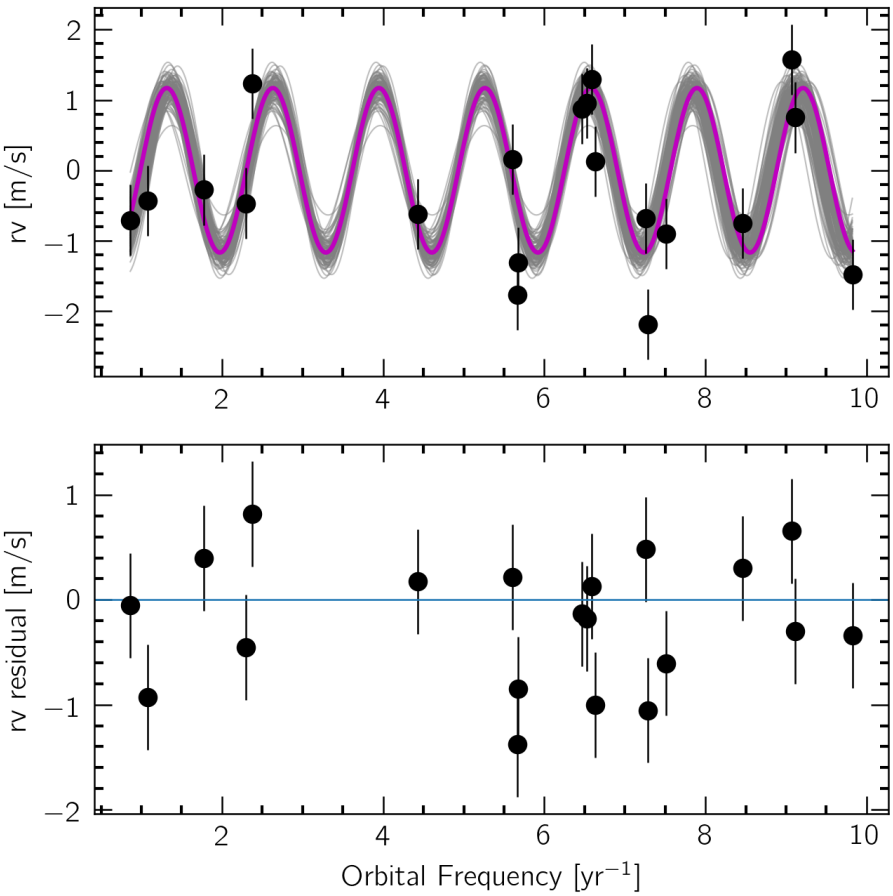


Figure 5.10: Top panel: Radial velocity mock data (black points with error bars) from Fig. 5.3, the best fit curve (magenta curve) and fits drawn from the posterior samples (Gray lines). Bottom panel: Residual between the data and the best fit. Because the error bars span $\pm 1\sigma$, and because 68% of Gaussian samples lie within 1σ is the mean, we expect around 68% of the error-bars to cross the zero-line if the fit is as good as it can possibly be.

Chapter 6

Model selection

We have so far learnt a number of techniques of parameter estimation. In all of these problems we had some data or information, and we have a model for how the data was generated and a model for the uncertainties in the data. For instance, in case of the radial velocity exoplanet detection, we knew that the data had the form of a sinusoid and that the noise were drawn from a Gaussian distribution. With this knowledge, we could estimate the model's parameters (amplitude, frequency and phase). However, in my cases you will encounter in life, it is unclear what the model is. Or maybe there are two competing models that could have led to the data. Worse yet, the two models could fit the data equally well. What to do then? Which model must be preferred? This chapter will give you the statistical tools to deal with such situations. We will pose the question of model selection in a Bayesian framework, learn about Occam's razor and come up with tools to deal with such model selection situations using simple examples.

6.1 Bayesian view of model complexity

Suppose you invited your friend to watch a football game at a bar. You reach the place, the match kicks off and your friend is nowhere to be seen. He shows up at half-time and says that he had a flat tire and it took a while to fix up his bike. Why did he not call? Oh that's simple, when he was rushing to fix the tire, his phone fell off and died. He lives in a student dorm so could have asked any of your common friends to drop a message. He says that he tried but could not find anyone around. Now, do you feel your sense of doubt rise in your mind? Do you feel that he just forgot about your date and is now lying to not feel embarrassed? Why do you have this feeling?

What you just did there was Bayesian model selection. You were presented with two models to explain the data: your friend showed up 45 minutes late. Model A is that your friend just forgot. Model B is the convoluted story that he told you. You don't know which one is correct for sure, but you prefer model A. Why? Because it is the *simplest* explanation for what happened. Model B needs a string of three unlikely things to happen in succession on the same day: (a) the flat tire, (b) the phone dying and (c) no common friends around in the dorm. It is *complex*. A complex model has too many knobs that have to be *fine tuned* to fit the data. We all have a natural sense of *parsimony* that tends to reject complex models with fine tuning. And if we abandon that sense, we will loose discretion.

The same is true in science. There are often competing ideas or theories and we use data to test them. We can always come up with a complex theory that has so many knobs that can be turned to perfectly explain the every little wiggle in the data (fine tuning). Does that mean that theory is to be preferred over a much simpler theory that has few knobs to turn but can more-or-less explain the data? Remember the 'date with a friend' example. We must always want our models (or theories) to be "as simple as possible." But how can we be quantitative and objective about this notion of complexity or simplicity of models? Let us explore this with an example.

Suppose you have some data D similar to the one we used to derive Hubble's law in Chapter 4 (Fig. 4.4). Suppose, you have two competing models to explain the data. Model A says that the data is a straight line though the origin. So it has one free parameter— the intercept, say a. Model B says that the data can be any straight line, so it has two free parameters— the slope and intercepts, say b and c. We wish to know which model must be preferred given the data; the one with one knob or the one with two knobs. Or in other words we want to find out the values of

$$\text{Prob}(A|D) \text{ and } \text{Prob}(B|D)$$

which are the probabilities that model A is true given the data D and the probability that model B is true given the data D.

Let us use our old friend, Bayes' theorem to write

$$\text{Prob}(A|D) = \frac{\text{Prob}(D|A)\text{Prob}(A)}{\text{Prob}(D)}$$

and similarly

$$\text{Prob}(B|D) = \frac{\text{Prob}(D|B)\text{Prob}(B)}{\text{Prob}(D)}$$

You can see that the $\text{Prob}(D)$ is common to both denominators, so it is quite convenient to work with the ratio

$$\frac{\text{Prob}(A|D)}{\text{Prob}(B|D)} = \frac{\text{Prob}(D|A)}{\text{Prob}(D|B)} \frac{\text{Prob}(A)}{\text{Prob}(B)}$$

which is called the 'Bayes' factor'. It is a product of two ratios. One of the ratio of prior probabilities of the models being true (before you even saw the data) and the other is the ratio of the evidence for the two models. As we will soon see it is this ration of evidences that codify our innate sense that prefers simpler explanations that fit the data over mode complex ones.

Let us say that we have no reason to privilege one model over another prior to the data. So we have $\text{Prob}(A)/\text{Prob}(B) = 1$. You already know how to write down the evidence for a model. It was the denominator when we applied Bayes' theorem for parameter estimation that we always ignored as it was 'just a normalising constant'. Well, you will soon see that when it comes to model selection it is not 'just' an normalisation inconvenience and you will see why it is called the 'evidence' in favour of the model.

The evidence for the two models is

$$\text{Prob}(D|A) = \int \text{d}a\, \text{Prob}(D|A, a)\text{Prob}(a)$$

and

$$\text{Prob}(D|B) = \int \text{db} \int \text{d}c\, \text{Prob}(D|B, b, c)\text{Prob}(b, c)$$

Basically we have written down an integral over the likelihood times the prior for the two models. These can be evaluated on a computer, but we will gain further insight if we soldier on analytically even if we made some grossly simplifying assumptions.

Let us assume uniform priors on the parameters over the ranges $[a_{\text{min}}, a_{\text{max}}]$, $[b_{\text{min}}, b_{\text{max}}]$ and $[c_{\text{min}}, c_{\text{max}}]$. Let us also assume that the maximum likelihood values of the parameters are a_0, b_0 and c_0 and that the likelihoods

curve have some characteristic width. In case of Model A, let the width be Δa and in case of Model B we have a characteristic area $\Delta b \Delta c$.[1].

The integrals have approximate values of

$$\begin{aligned}\text{Prob}(D|A) &= \frac{1}{a_{\max} - a_{\min}} \int_{a_{\min}}^{a_{\max}} \mathrm{d}a \, \text{Prob}(D|A,a)\text{Prob}(a) \\ &\sim \frac{\text{Prob}(D|A,a_0)\Delta a}{a_{\max} - a_{\min}}\end{aligned}$$

and

$$\begin{aligned}\text{Prob}(D|B) &= \frac{1}{(b_{\max} - b_{\min})((c_{\max} - c_{\min})} \int_{b_{\min}}^{b_{\max}} \mathrm{d}b \int_{c_{\min}}^{c_{\max}} \mathrm{d}c \\ &\quad \text{Prob}(D|B,b,c)\text{Prob}(b,c) \\ &\sim \frac{\text{Prob}(D|B,b_0,c_0)\Delta b \Delta c}{(b_{\max} - b_{\min})((c_{\max} - c_{\min})}\end{aligned}$$

Let us understand what these terms are telling us. The maximum likelihood value, $\text{Prob}(D|A,a_0)$ telling us how likely it is that this data D can be generated with this model with parameter a_0. a_0 is the best the model can do to explain the data. So this factor is telling us how well the model can describe the data. A model that fits all the fluctuations in the data well is rewarded by this factor *regardless* of how complex the model is. Next, look at the factor $\Delta a/(a_{\max} - a_{\min})$. This factor is a ratio of the range of acceptable values of a that can fit the data to the total range of possible values of a. In other words it is telling us what fraction of all possible values of a can fit the data, or how finely-tuned does a have to be to fit the data. It is a measure of the 'fine-tuniness' of the model (I made that word up; but you know what I mean!) and is often referred to as 'Occam's factor'.

So the evidence for any Model is a product of 'how well the model fits the data' and 'how fine tuned the model must be to fit the data'. The Bayes factor is just a ratio of evidences. It rewards a Model for fitting the data well but also punishes it for having too many knobs that must be tuned to fit the data. And in simultaneously doing both it finds the optimal balance and gives us an objective method for model selection.

You may have noticed that the Occam's factor necessarily depends on the prior range you allow for the parameters. In the examples we have done so far, we have not been forced to carefully consider our prior range. Bayesian model selection forces us to select this range carefully without looking at the data. That is we are not allowed to find the most likely value of the parameter and *then* choose the prior range. That is cheating. we have to choose the prior range without any knowledge of the data.

[1]Strictly speaking the characteristic area is an ellipse as we saw in Chapter 4 with mock Hubble's law data.

6.2 Example: polynomial order selection

Suppose you are investigating a car crash. You wish to know if the accused was decelerating in the moments before the crash or was moving at a constant speed (e.g. on cruise control). Here is all the information you have: (1) based on the lack of 'unusual speed events' on the radars on the highway the car was moving between 15 m/s and 33 m/s during and immediate before the crash. (2) Based on the condition of brakes on the car, the deceleration could not have exceeded $5\,\mathrm{m/s^2}$. (3) A traffic camera recorded the crash and based on the footage the exact location of the car is known at some time, say $t = 0$ seconds and thereafter the distance travelled by the car has been estimated with independent Gaussian errors as given in Fig. 6.1. Did the accused apply any brakes at all[2]?

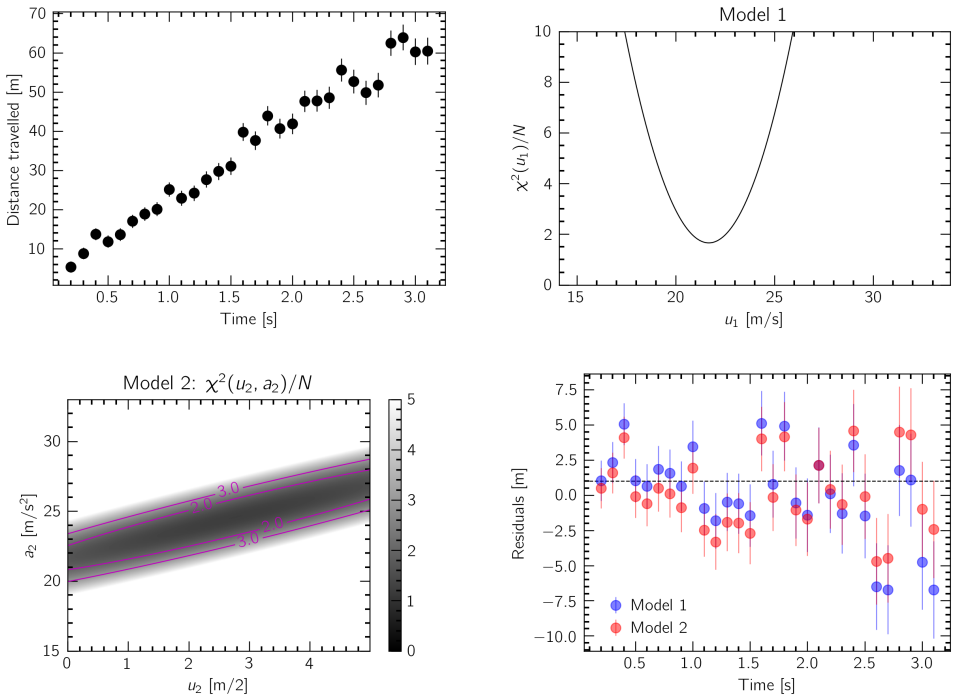


Figure 6.1: Top left: Data on the distance travelled by the car as a function of time (with error bars). Top right: chi-squared per data point for Mode 1. Bottom left: chi-squared per data point for Model 2. Bottom right: Data minus best fit for the two models.

[2]Let us assume that the punishment depends on this and you have been asked to testify (because you took this course)!

Based on what we have learnt before this chapter we would model the distance travelled, s versus time, t at the k^{th} datum as

$$s_k = ut_k + \frac{1}{2}at_k^2 + n_k$$

where u is the speed at time $t = 0$, the distance travelled is measured from the time $t = 0$ and n are the noise realisations. And then you would try to fit for the parameters u and a and then check of the fit can reject $a = 0$ at some confidence. But now you have learnt to do something better. You can cast this as a model selection problem. You now have two models: Model $M1$ with no deceleration and Model $M2$ with deceleration. We have

$$\text{Model}\, M1 :\ s_k = u_a t_k + n_k$$

and

$$\text{Model}\, M2 :\ s_k = u_b t_k + \frac{1}{2}a_b t_k^2 + n_k$$

We are asking the question: which model is preferred by the data. To be fair, we do not give any prior preference to the models. The expression for the Bayes' factor is then just the ratio of evidences for the two models.

The evidence for Model $M1$ is

$$\text{Prob}(D|M1) = \int \mathrm{d}u_1 \text{Prob}(D|u_1)\text{Prob}(u_1)$$

The prior on the parameter is

$$\text{Prob}(u_1) = \left\{ \begin{array}{ccc} \frac{1}{33-15} & ; & 33 > u_1 > 15 \\ 0 & ; & \text{otherwise} \end{array} \right\}$$

The likelihood in the integrand is

$$\text{Prob}(D|u_1) = \prod_k e^{-(s_k - u_1 t_k)^2/(2\sigma_k^2)} / \sqrt{2\pi\sigma_k^2}$$

Similarly, for Model $M2$ we have the evidence

$$\text{Prob}(D|M2) = \int \mathrm{d}u_2 \int \mathrm{d}a_2 \text{Prob}(D|u_2, a_2)\text{Prob}(u_2, a_2)$$

The prior is

$$\text{Prob}(u_2, a_2) = \left\{ \begin{array}{ccc} \frac{1}{(33-15)(5-0)} & ; & 33 > u_2 > 15, 5 > a_2 > 0 \\ 0 & ; & \text{otherwise} \end{array} \right\}$$

and the likelihood in the integrand is

$$\text{Prob}(D|u_2) = \prod_k e^{-(s_k - u_2 t_k - 0.5 a_2 t_k^2)^2/(2\sigma_k^2)} / \sqrt{2\pi\sigma_k^2}$$

We therefore have the Bayes' factor of

$$\mathcal{B}(M1/M2) = \frac{\int \mathrm{d}u_1 \text{Prob}(D|u_1)\text{Prob}(u_1)}{\int \mathrm{d}u_2 \int \mathrm{d}a_2 \text{Prob}(D|u_2, a_2)\text{Prob}(u_2, a_2)}$$

$$= \frac{5}{1} \frac{\int_{15}^{33} \mathrm{d}u_1 \prod_k e^{-(s_k - u_1 t_k)^2/(2\sigma_k^2)}}{\int_{15}^{33} \mathrm{d}u_2 \int_0^5 \mathrm{d}a_2 \prod_k e^{-(s_k - u_2 t_k - 0.5 a_2 t_k^2)^2/(2\sigma_k^2)}}$$

I have implemented this integrals numerically on a computer and the fits are shown in Fig. 6.1. The Bayes factor ratio I get is

$$\mathcal{B}(M2/M1) = 310$$

How to interpret the Bayes factor? A common prescription is as follows: Bayes factor of 20–150 is akin to a confidence of 95–90% that one Model is better than the other
Bayes factor greater than 150 is akin to a confidence exceeding 99%. Hence our conclusion here is that even though Model 2 has one extra knob to turn, the data requires this knob to be present and hence we have strong evidence in favour of the brakes being applied.

For future applications, here is a table that gives guidance on how to interpret the Bayes' factor.

Bayes factor	**Strength of evidence**
1–3	Weak
3–10	Moderate
10–30	Substantial
30–100	Strong
100–300	very strong
> 300	Decisive

6.3 Example: spectral line fitting

Now that we have been introduced to model selection let us take up a good practical astrophysical example. It concerns fitting of absorption components to a spectrum which is a commonly encountered problem. Consider the spectrum in Fig. 6.2. What is happening here is that a background source's light is being absorbed by intervening clouds of gas. So if the background source has a flux density at frequency ν of $S_0(\nu)$, then the absorption causes the observed flux to have the form

$$S(\nu) = S_0(\nu) \exp\left[-\tau(\nu)\right]$$

where $\tau(\nu)$ is called the optical depth. In the limit where the optical depth is much smaller than unity, we have the approximation

$$S(\nu) \approx S_0(\nu) \left[1 - \tau(\nu)\right].$$

The cloud's preferentially absorb radiation near certain wavelengths which means that the optical depth $\tau(\nu)$ is a sharply peaked function of frequency ν. Based on theoretical arguments, we can usually anticipate the shape of the optical depth. Let us here assume that the optical depth is given by a Lorentzian profile with width parameter set to unity:

$$\tau(\nu) = \frac{\tau_0}{1 + (\nu - \mu)^2}$$

where τ_0 is the overall strength of the line and μ is the frequency at the line centre. We will also make the simplifying assumption that the background flux of the source is unity, that is $S_0(\nu) = 1$.

So far so good. We have learnt of non-linear multivariate parameter estimation techniques to estimate ν_0 and τ_0 given a spectrum of the kind in Fig. 6.2.

But what if there are two or more along the sight-line to the background source? These clouds may have their own line-of-sight velocities which will Doppler shift their apparent line centres. So in this case, the data should be seen as a sum of different absorption components of each cloud:

$$S(\nu) \approx S_0 \exp\left[-\sum_i \tau_i(\nu)\right] \approx S_0 \left[1 - \tau_1(\nu)\right] \left[1 - \tau_2(\nu)\right] \ldots.$$

But we don't know apriori how many clouds there are; much in the same way we did not know apriori which order polynomial we should use to fit the data. Hence this is also a model selection problem where a large number of models may be considered: Model $M1$ has one intervening cloud, Model $M2$ has two intervening clouds and so on.

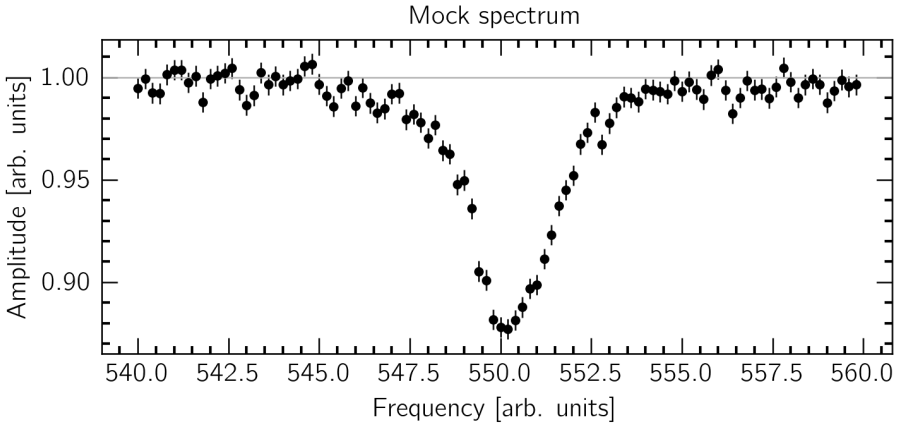


Figure 6.2: Mock data showing absorption in the spectrum of a background source. Is the absorption feature caused by a single cloud or multiple clouds.

So the question before us is: how many cloud components should we use to model the data in the Fig. 6.2. That is, we need to find a model from among the set [$M1$, $M2$, $M3$ ...] that is complex enough to fit the data but no more. Of course, we will use a Bayes' factor comparison to find the right amount of model complexity.

I will show you the calculations for Model $M1$ and Model $M2$, but the other models can be treated in an analogous way. What we expect to see as we increase model complexity is this.
(a) Initially the model is too simple and cannot fit the data adequately. These models will have high Occam's factors but very low peak likelihood values.
(b) As the model complexity increases, the Occam's factor will decrease but the likelihood will increase so rapidly that the overall evidence will increase.
(c) At some level of complexity the model can adequately describe all the variation in the data (minus noise, of course) and this is the highest level of likelihood that can be achieved. The highest possible value of the peak likelihood times the Occam's factor has been achieved.
(d) Increasing the complexity any further cannot increase the peak likelihood. It was only decrease the Occam's factor and therefore the evidence starts to drop.

Ok let us do the estimation for the two models, Model $M1$ and Model

$M2$ gievn by

$$\text{Model } M1: \; S_k = 1 - \frac{\tau_0}{1 + (\nu_k - \mu)^2} + n_k$$

and

$$\text{Model } M2: \; S_k = 1 - \frac{\tau_1}{1 + (\nu_k - \mu_1)^2} - \frac{\tau_2}{1 + (\nu_k - \mu_2)^2} + n_k$$

where n_k are noise samples drawn from a Normal distribution with zero mean and variance of σ^2.

Clearly the model is non-linear in its parameters, so we will have to either use the gradient technique or a Monte Carlo approach. Let us use the Monte Carlo technique. We have learnt in the last chapter how to do that so we don't need to repeat it here. However, the priors have to be specified sincerely for the Bayes' factor to be taken seriously. It is clear that the location of the peaks cannot be outside the extent of the spectrum so we will use a uniform distribution between frequencies of 540 and 560. As for the optical depth, we need to specify that based on the physical expectations of the clouds. Suppose that range is between 0 and 1. With these the log evidence for each model can be calculated using the posterior distribution of the Monte-Carlo runs.

and Fig. 6.3 and Fig. 6.4 show the corner plots for the parameters [h]

The Figures show that all parameters have posteriors that have a single peak and a roll-off on either sides which means that the parameters' posterior has been well sampled. As we said at the end of the last chapter, it is also good practice to check how well the estimation did. The residuals are given in Fig. 6.5. We can see that Model $M1$ has residuals that deviate around the location of the line whereas Model $M2$ fits the data much better around the line centre. So clearly $M2$ will have a higher peak likelihood than Model $M1$. It remains to be seen if the difference in Occam's factors can offset this. So let us compute the evidence values for the Models. We need to integrate the un-normalized posterior distribution (which is the prior times the likelihood) to find the evidence. We can do this numerically by constructing the integrand as the histogram of the Monte Carlo samples. Another quicker approximation which I want to demonstrate is by approximating the integrand as a multi-dimensional Gaussian distribution which can be integrated in closed form. For a single variable, the density distribution has the form

$$\text{Prob}(x) = \frac{1}{2\pi\sigma^2} \exp\left[-(x - m)/(2\sigma^2)\right]$$

where x is the variable and m and σ^2 are the mean and variance respectively. Similarly for a multi-dimensional Gaussian distribution of a set of variables

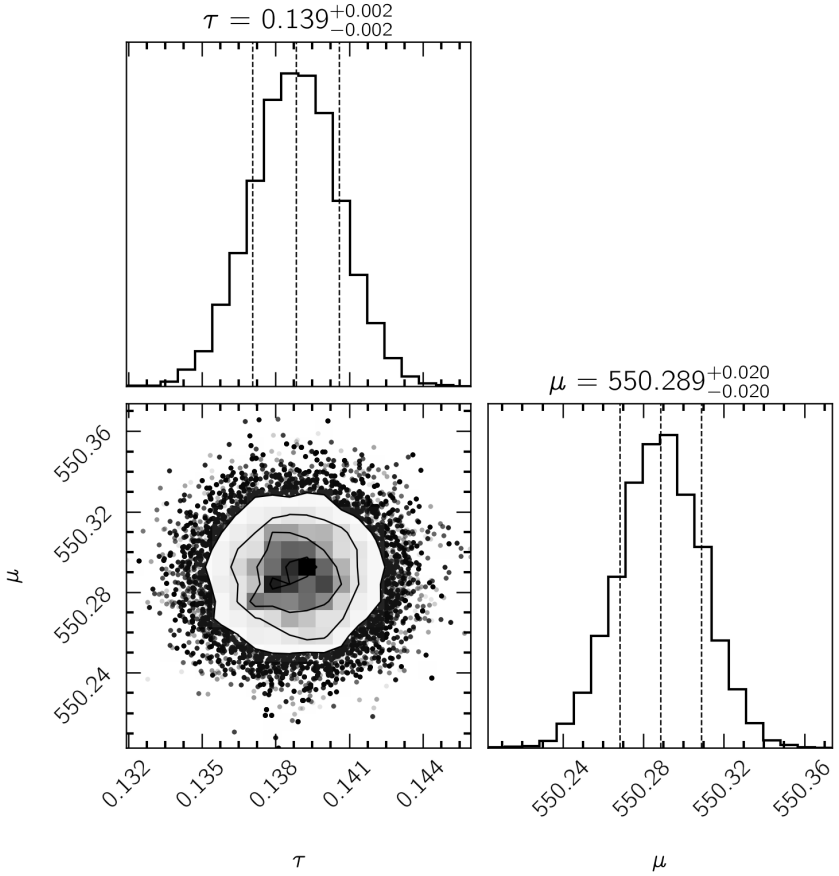


Figure 6.3: Parameter posterior distribution for Model $M1$.

collected in the vector $\boldsymbol{x}$ we can collect the means in the vector $\boldsymbol{m}$ and co-variances in the matrix $\boldsymbol{\Sigma}$. Notice that if there are n variables then $\boldsymbol{x}$ and $\boldsymbol{m}$ have the shape $n \times 1$ and $\boldsymbol{\Sigma}$ has the shape $n \times n$. This is because the matrix $\boldsymbol{\Sigma}$ has all the pairwise covariance values. The i, j^{th} element of that matrix is the co-variance

$$\Sigma_{i,j} = \langle (x_i - m_i)(x_j - m_j) \rangle$$

The multi-variate Gaussian distribution can then be written as

$$\text{Prob}(\boldsymbol{x}) \equiv \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma}) = [(2\pi)^n \det(\boldsymbol{\Sigma})]^{-1/2} \exp\left[-(\boldsymbol{x} - \boldsymbol{m})^T \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{x} - \boldsymbol{m})\right]$$

where 'det' refers to the determinant. You can verify that this reduces to the single variable expression for $n = 1$. Now this distribution is properly

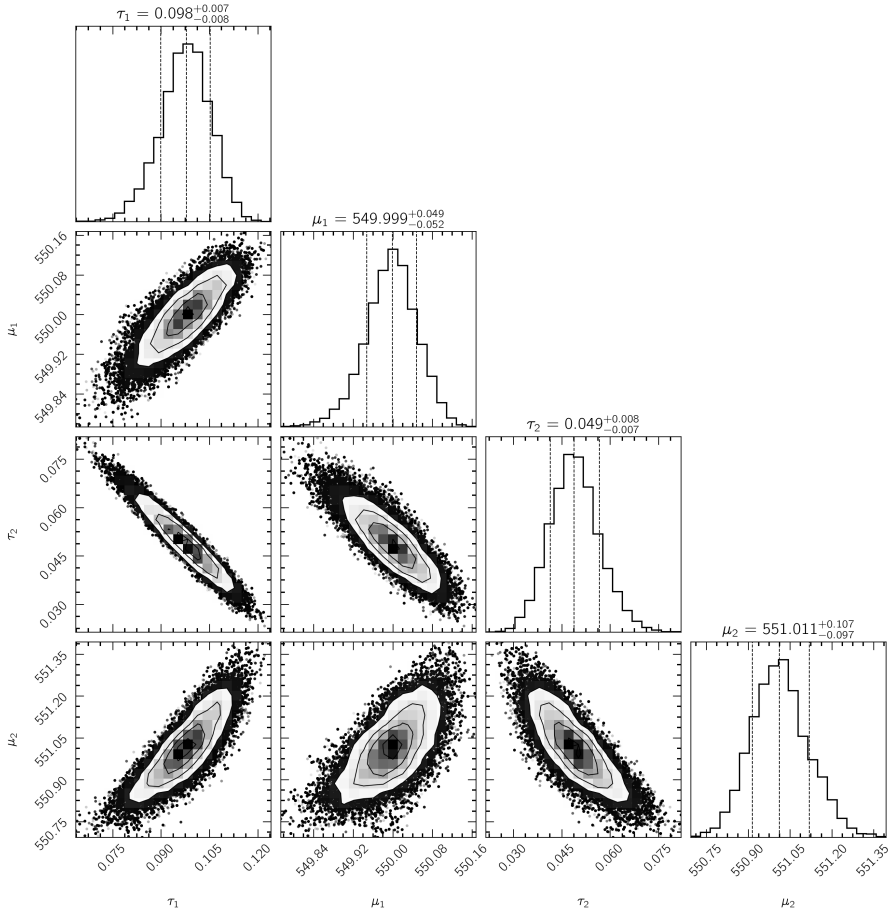


Figure 6.4: Parameter posterior distribution for Model $M2$.

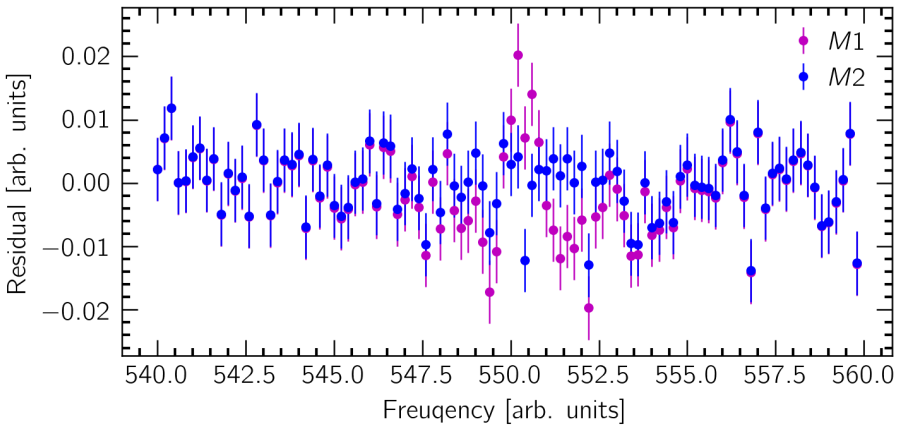


Figure 6.5: Residual spectral fluctuations after the best fit models have been subtracted from the data

normalised, which is to say that

$$\int \mathrm{d}x_1 \int \mathrm{d}x_2 \ldots . \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma}) = 1$$

The evidence for Model $M1$ is the integral of the likelihood times the prior:

$$E_1 = \int \mathrm{d}\tau \int \mathrm{d}\mu (\text{likelihood} \times \text{prior})$$

$$= \frac{1}{\tau_{\max} - \tau_{\min}} \frac{1}{\mu_{\max} - \mu_{\min}} \int \mathrm{d}\tau \int \mathrm{d}\mu \text{ likelihood}$$

We are going to approximate the likelihood to be a multivariate Gaussian. Let the peak value of the likelihood be reached at some location $[\tau_p, \mu_p]$ and its value be $L(\tau_p, \mu_p)$. The integral then reduced to the area under an improperly normalised multivariate Gaussian distribution:

$$E_1 = \frac{1}{\tau_{\max} - \tau_{\min}} \frac{1}{\mu_{\max} - \mu_{\min}} L(\tau_p, \mu_p)[(2\pi)^n \det(\boldsymbol{\Sigma})]^{1/2}$$

We can obtain good approximations for $L(\tau_p, \mu_p)$ and $\boldsymbol{\Sigma}$ using the samples from our Monte-Carlo iterations. For examples, we can record the peak likelihood value reached by the walker as $L(\tau_p, \mu_p)$ and we can compute the sample co-variances (using `numpy.cov`) as $\boldsymbol{\Sigma}$. With these we have all the information we need to compute the evidence for any model. In fact just like in the case of likelihoods, we will find that it is better to calculate the log of the evidence: $\mathcal{E} \equiv \log E$.

Here are the outputs of the code that show some relevant values for the two models.

```
Model 1
ML value = [1.4239e-01 5.5031e+02]
Covariance matrix =
[[3.2346e-06 2.6089e-07]
 [2.6089e-07 4.0224e-04]]
Chi2 = 180.13
log evidence = -101.45

Model 2
ML value = [9.0763e-02 5.4995e+02 5.9517e-02 5.5092e+02]
Covariance matrix =
[[ 0.0002  0.0009 -0.0002  0.0016]
 [ 0.0009  0.0068 -0.001   0.0092]
 [-0.0002 -0.001   0.0002 -0.0015]
 [ 0.0016  0.0092 -0.0015  0.0186]]
Chi2 = 104.90
log evidence = -71.87
```

Notice the χ^2 values for the two models. There are 100 points in the data so the chi-squared per data point is 1.8 for model $M1$ and around 1 for model $M2$. This means that Model 2 is definitely providing a much better fit to the data. But is this 'better' justified given the added complexity of model $M2$ (4 knobs to turn instead of 2 knobs). The log evidence says that it is. So we must prefer Model $M2$. Although I have shown the example here of just two models to keep the discussion and the code simple, we can also fit the data for model $M3$ with three absorption components. Because Model $M2$ has already reached as good a χ^2 value is feasible, it is very likely that Model $M3$ will be punished for it unnecessary complexity.

Chapter 7

Non-parametric methods & entropy

We are rapidly reaching the end of the course, which means we may not have the time necessary to delve deeper into many more interesting topics. So this week, we will study a few assorted topics in statistics. We will start with non-parametric methods. These are statistical methods to deal with situations where there is no parametric/algebraic model that can be specified to explain the data. But we still have certain statistical questions to ask of the data. The most common example of this is a 'test for correlation' between two variables based on data collected about the two variables. Another common example we will explore is the question of whether two sets of samples are drawn from the same probability distribution function. Then we will 'get meta' and talk about information theory, which is a methodical way to measure the amount of information contained in data. This will lead us to to some interesting prescriptions for the assignment of prior distributions, and the comparison between frequentist and Bayesian techniques.

7.1 Tests of correlation

Sometimes you just want to know if two variables are related. You don't even know what the exact relationship should or could be. Take for example a hypothetical situation from the early days of modern astronomy. Consider the data shown in Fig. 7.1. Here I am plotting mock measurements of a set of white dwarfs' radius and mass. These measurements are counter-intuitive as they seem to show that more massive white dwarfs are smaller, opposite of what is expected of normal stars[1]. Suppose that you don't yet have a detailed enough physical theory that allows you to formulate a parametric (or algebraic) relationship between the mass and radius. For the moment you just want to know if this inverse relationship is really what the data is telling you or if it is just a statistical anomaly.

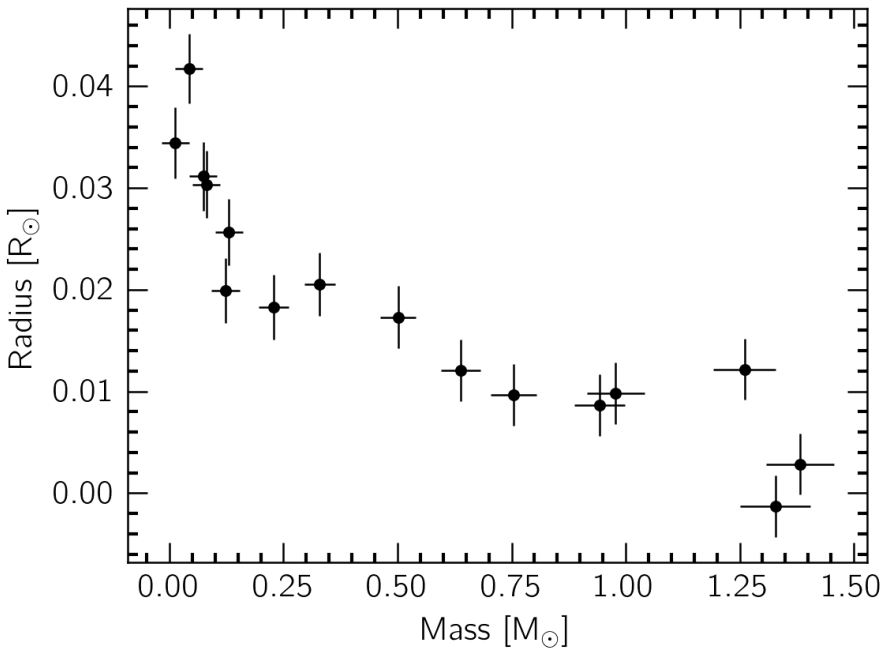


Figure 7.1: Mock measurements of a set of white dwarfs' mass and radius.

OK, one thing you can always do is just check 'by eye' if the points show any trend at all. As in do the points seem to lie randomly all over the

[1]In reality, the puzzle posed by white dwarfs was their apparent faintness despite their high temperature, implying that they must have very small radii and astonishing densities. See Holberg (2009) for a nice historic overview.

place or do they look like they are scattered around some (imaginary) line or curve. It seems from the plot that there is an inverse relationship but at the end of the day, you need a quantitative measure of the existence of a relationship, or lack thereof. You have to be sure that you are not seeing patterns that don't exist as this is a common human fallacy.

The next best thing you can do is try to fit a straight line through the data points. If the slope of the straight line has different from zero in a statistically significant way, then you can say that the two variables must be related physically. Let us do that. Suppose we fit the line $y = ax + b$ to the data points (x_k, y_k). We know from previous chapters that if the x_k are assumed to be measured perfectly and all uncertainties are somehow attributed to y_k, then the least-squares estimate of the slope a is

$$\hat{a} = \frac{N \sum_k x_k y_k - \sum x_k \sum y_k}{N \sum y_k^2 - \left(\sum y_k\right)^2}$$

Because the model is linear, we have also learnt that we can compute the uncertainty of the slope parameter and then check the consistency (or lack thereof) of the slope with a value of 0 (no correlation).

Here is the next best thing to do. It would be handy if we can define some new measure that is 0 for no-correlation and 1 for perfect correlation (of -1 for perfect anti-correlation)— sort of a 'normalized slope'. Why do we need that? Suppose you measure y in some different units, say $z = 10 \times y$. Then the slope and its uncertainty also are now scaled by a factor of 10. Life would be easier if we can come up with a measure of 'relatedness' of x and y that did not depend on the units we use to measure x and y. Life would also be amazing of this measure was symmetric with respect to x and y. That is if the measure was not specific to the formulation $y = f(x)$ and could also accept the formulation $x = g(y)$ to be equally valid.

Let us consider the other formulation of the same model: $x = g(y)$. In this formulation, we can assume that y_k are measured perfectly and all measurement uncertainties are somehow assigned to x_k. Then, we would fit the model $y_k = a'x_k + b'$. The least squares estimate of the new slope a' is

$$\hat{a'} = \frac{N \sum_k x_k y_k - \sum x_k \sum y_k}{N \sum x_k^2 - \left(\sum x_k\right)^2}$$

Now in reality there are uncertainties in both the measurement of x_k and y_k and we have no reason to prefer one formulation of the model over the other. Remember again that this is *not* model comparison; they are the *same* model, just formulated differently. In fact, we have $aa' = 1$. Let us

see what we obtain when we multiply the two slope estimates

$$\sqrt{\hat{a}\hat{a}'} \equiv R = \frac{N\sum_k x_k y_k - \sum x_k \sum y_k}{\sqrt{N\sum x_k^2 - (\sum x_k)}\sqrt{N\sum y_k^2 - (\sum y_k)}}$$

Now we have a quantity, R, that treats x_k and y_k in a symmetric fashion. And as a consequence of this symmetry, R is so normalised that it is independent of units used to express x and y. In fact, it is also invariant under any translations of the location of the origin in the x, y plot. You can prove this with the substitution $x \to c + dx$, $y \to e + fy$ which will still give you the same value of R. R is also easy to interpret: (a) if $R = -1$ then the two variables are perfectly anti-correlated; by which we mean that y must decrease in proportion to x's increase, (b) if $R = 0$ then x and y are uncorrelated that is there is no linear relationship between the two variables, and (c) of $R = 1$ then the two variables are perfectly correlated; that is x increases in proportion to an increase in y.

R here is called the 'Pearson's correlation coefficient'. It is a very quick and easy way to see if two variables are correlated and if it is worth any effort at all to understand *how* they are correlated. R is also sometimes called the 'linear correlation coefficient'. This is because R works out to be ± 1 only when the two variables are *linearly* related (and there is negligible uncertainty). If suppose y were equal to ax^2 then R will not be unity, even though x and y are related. How to deal with situations like this?

Suppose we only want to know the answer to 'Does y increase/decrease when I increase x?'. We don't care in what proportion it increases/decreases; that is we don't care at what *rate* it increases/decreases: It could be linear, quadratic, exponential, logarithmic, whatever else; we don't care. We just want to know if there is *any* monotonic relationship between x and y. Then what we can do is assign ranks to the value of x and y. That is, we say that the rank is 1 for the lowest measured value of x_k, rank is 2 for the next highest value among x_k and so on until rank n. We do the same for the values of y_k. We then get two vectors of ranks: $r_k^{(x)}$ and $r_k^{(y)}$. We now take the 'Pearson's correlation coefficient' of these two vectors of ranks. This new coefficient is called the 'Spearman's rank correlation coefficient'. What is this new co-efficient telling us?

Notice what we did when we took the ranks instead of the actual values. We made the correlation coefficient only dependent on the relative largeness or smallness of the variables and not their actual values. For instance, numbers drawn from the arithmetic progression 3, 7, 11, 15 ... will have the same ranks as the corresponding numbers drawn from the geometric progression 3, 12, 48, 192, ... In other words, the ranks remove the effects of the rate at which some variable increases with respect to some other

variables and only distils the direction in which the variation happens. This is precisely what we want when we want to test for a generic monotonic relationship. This is why the Spearman's rank correlation coefficient can be used to detect any monotonic relationship.

I have implemented both the Pearson's and Spearman's correlation coefficients for the data shown in Fig. 7.1 on a computer and I get the following result Pearsons coeff = -0.85 Spearman coeff = -0.92 They are both negative which makes sense as there is an inverse relationship between the mass and radius. Furthermore, Spearman's coefficient is closer to -1 which *suggests* that perhaps the inverse relationship is non-linear. We better work on a good theory of how these stars can have this inverse relationship and use the data to test the precise details of the theory.

In any case, we calculated the correlation coefficients but we still have to be very careful in claiming an anti-correlation because the coefficients are themselves statistical estimates which means that they also have an associated uncertainty. Of course the uncertainty must depend on the size of the uncertainties in the measurements. They should also depend on the number of measurements (you can always fit a straight line between any two points!). How to assign uncertainties to these numbers? If we precisely knew the statistics of the measurements errors then we could maybe work things out analytically, which is what we have done in the course so far when we computed likelihoods and posteriors and all that. What if we don't have this information? There is a neat trick called 'Bootstrapping' to deal with such situations. And the bootstrapping methods can be used beyond correlation analysis, even in parameter estimation problems. It comes to the rescue in situations where the uncertainty on the data cannot be estimated properly.

7.2 Bootstrapping

Computationally speaking, bootstrapping is a clever trick to use the power of modern computers to assign uncertainties in a 'Monte-Carlo sort of way'. Let us learn about this technique using an example. Suppose you want to know the average age of stars in some cluster. You collect age measurements of the cluster members from some old literature reference. Suppose you get the following values in units of gigayears

```
1.09, 0.70, 1.38, 1.11, 0.79, 1.11, 1.18, 0.62, 0.87,
0.81, 0.80, 1.01, 1.07, 0.83, 0.75, 0.95, 1.05, 1.22,
1.31, 0.94
```

You add the numbers up and divide by the number of stars and you get

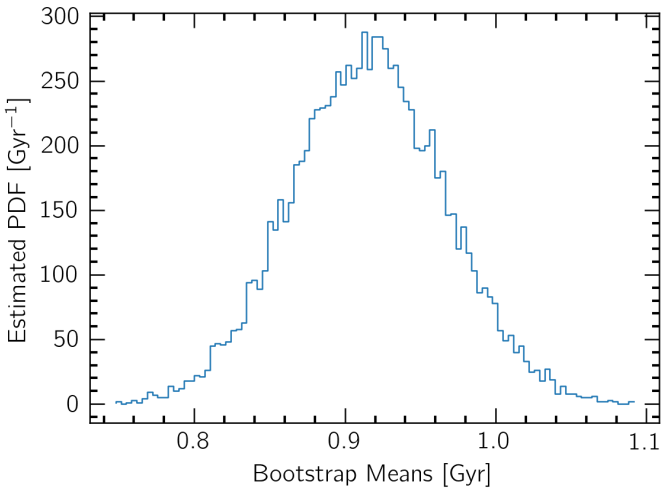


Figure 7.2: Histogram of the Bootstrapped values

```
Sample mean = 0.93
```

Now the problem is: what is the uncertainty on this mean value? Same questions as we asked before about the uncertainty on the correlation coefficient. Suppose that the paper gave no reliable error bars on those measurements. What to do now?

Enter Bootstrapping; here is what you do. You run a large number of simulations, say 10000. In each iteration, you sample 20 numbers from the list with replacement and calculate the mean of these numbers. In every iteration, the mean will be slightly different and therefore have a different *realisation* of the underlying noise in the measurement. The histogram of the means is then an approximation of the inherent uncertainty on the sample mean. Fig. 7.2 shows the histogram of Bootstrapped means and the standard deviation of the values in the histogram is

```
Bootstrap std = 0.05
```

Therefore we would decide that the average age of the stars in the cluster is $0.93 \pm 0.05\,\mathrm{Gyr}$.

Why does Bootstrapping work? There is a deeper reason for this related to the central limit theorem which we will not have time in this course to go into. But it is what saves us in many situations where a good understanding of uncertainties, or more generally, an understanding of the underlying probability density function is unknown. Take for example, our example of straight line fitting with mock data on Hubble's law from Chapter 4.

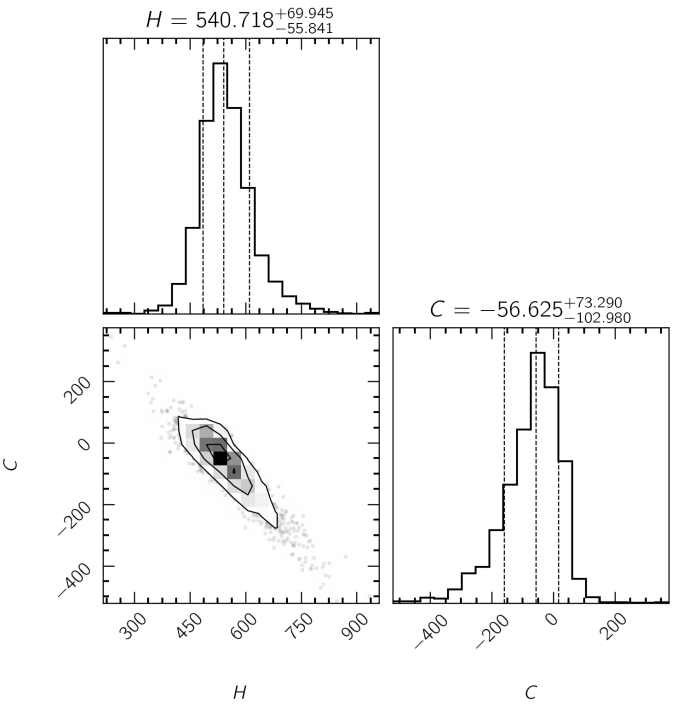


Figure 7.3: Posterior distribution of the slope, H and intercept C obtained via 2000 Bootstrapping iterations.

Suppose we did not have any idea of the uncertainties in the measurement of the galaxies recession velocity. We could then use Bootstrapping by drawing the distance–velocity measurement pairs at random from the data with replacement and fit straight lines to the data. We would then get a large number of slope and intercepts whose histogram would approximate the posterior distribution of the slope and intercept parameter. I have implemented this for the same data as used in Chapter 4 (Fig. 4.4) with 2000 Bootstrapping iterations and obtained the posterior distribution shown in Fig. 7.3. If we compare this to the posterior in Chapter 4, then we see that the Bootstrapping did quite well despite not being provided *any* information on the uncertainties in the measurements.

OK good. Let us now return to our correlation coefficients of white dwarfs' mass and radius. How can we use the idea of bootstrapping here? We could again draw with replacement, data pairs and compute the correlations. I have done that and Fig. 7.4 shows the histogram of the Pearson and Spearman correlation coefficients.

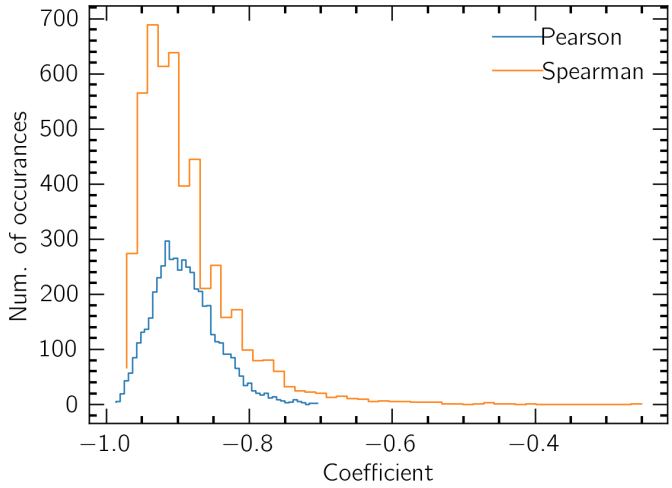


Figure 7.4: Histogram of bootstrapped Pearson and Spearman correlation coefficients.

With this histogram now we have everything we need. We can normalise the histogram and find the 99% confidence interval for our correlation values. We also see that the Spearman coefficient has a peak at a different location than the Pearson coefficient. This also shows that the relationship is very likely not linear, because if it were then the two coefficients would have very similar histograms.

7.3 Hypothesis testing

We just learnt how we can use Bootstrapping to do our uncertainty analysis in a non-parametric way— useful when we have poor knowledge of the uncertainty in the data. Can we also make statistically sound statements on the veracity of a theory or hypothesis in a non-parametric way? This will again become necessary in cases where we have poor knowledge of the uncertainty in the data. In such cases we cannot write down expressions for likelihoods and that means we have no way of writing down expressions for the posterior or the Baye's factor.

In such situations we can use what is commonly called a 'Hypothesis test'. I wish to teach you this as it can come in very handy but I must forewarn you that if you are not super careful, this kind of hypothesis testing can lead you down the wrong path.

Now that you have been warned, here is what we do in hypothesis test-

ing.
(a) We state a 'null-hypothesis': This is usually the opposite of what we are setting out to prove. It is a statement of 'no-effect'. So in our white dwarf example, the null hypothesis will be 'There is no relationship between the mass and radius of a white dwarf.' In case of drug-testing it will have the form: 'There is no effect of <the drug> in mitigation of <the disease>.
(b) You then calculate the probability of obtaining the data given the null hypothesis. If you do not have a good way to estimate the uncertainties in the data, then you will have to calculate this probability using a non-parametric way.
(c) If this probability is below some pre-defined threshold then you have reached a statistical contradiction. You then conclude that the 'null-hypothesis' must be rejected.
(d) Now comes the controversial last step: 'If the null hypothesis is rejected, then the alternate hypothesis must be true'. This is what gets most people down the wrong path. You can only make this conclusion if the null and alternative hypothesis are mutually exclusive and exhaustive (their probabilities add to 1). In other words, this is only correct when there is no other alternate third possibility.

Let us understand the above algorithm better with our white dwarf example. Under the null-hypothesis, there is no relationship between the mass and radius. This means that the measured mass–radius pairing is incidental: that is, if we scramble the numbers in the mass and radius vectors it will make no statistical difference to the correlation coefficient. Ok let us try that. We can run a large number of iterations on a computer where in each iteration, we randomly scramble the mass vector and compute the Spearman's coefficient. We can collect these coefficients in a histogram (see Fig. 7.5).

The Figure shows that even after 1000 random realizations under the null-hypothesis, no coefficient value reaches that measured on the data. This means that there is a less than one in 1000 chance that the inverse relationship seen in the data was caused by pure chance. Hence we can reject the null-hypothesis (no relationship between mass and radius) with a significance of more than 99.9%.

I will end by re-iterating my caution. Just because we rejected the null-hypothesis does not mean that the alternate hypothesis (which we set out to prove) must be necessarily true. It could be that the data is inconsistent even with the alternate hypothesis! It could be that both the null and alternate hypothesis are false and some third hypothesis (which we did not even consider) is actually true. In our simple example, we cast the world into a binary: on in which there is a correlation and one in which there

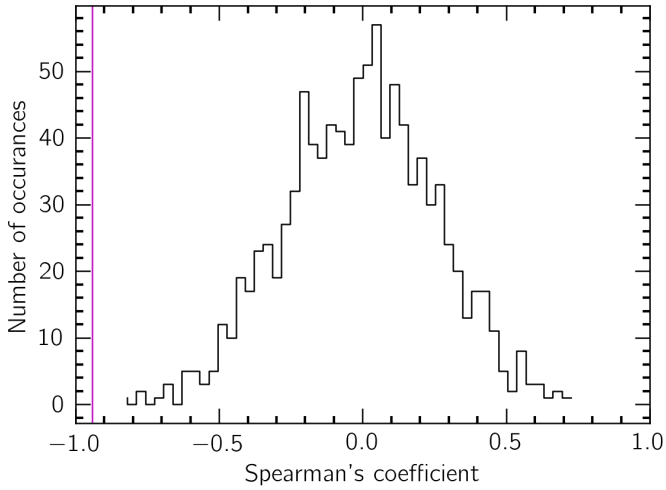


Figure 7.5: Hypothesis test for the white dwarf mass–radius correlation. The histogram ($N = 1000$) shows the outcome of random evaluations under the null-hypothesis (no correlation) and the magenta line shows the actual coefficient (no randomization).

isn't. So we can feel safe in asserting that a rejection of the null hypothesis means that the alternate is true. But this is often not the case in complex real-world situations. My recommendation is to use this kind of hypothesis test as an informal first-pass to see if the data is telling you something interesting, and then cast the problem in a Bayesian framework to get to your final conclusion.

7.4 Information and entropy

This is the part of the course where we need to get a little 'meta' about probabilities. Information theory is a whole new branch which we wont have the pleasure of learning due to limited time, but it is worth knowing something about it as it teaches us how to properly assign priors to variables. So let us get started.

Information is a measure of how much you know about something. Entropy is a measure of randomness of something. Say the probability of an event is 1. That means the entropy is 0; there is no randomness. When that event occurs, you have learnt precisely *nothing new*. Suppose the probability of an event is 0.5. Now there is some inherent randomness to it: you don't know if it will happen or not. That means that when it does (or

does not) happen, you actually have learnt something. So there is a deep relationship between information and entropy.

What has this to do with the assignment of priors? Well if you truly have no information on the prior value of a parameter you are trying to estimate, you must assign a distribution that is 'non-informative' or a distribution with the most entropy. We naively assumed that this was a uniform distribution. But now we will ask if that really so? To know the answer we have to first define information and entropy more precisely and mathematically.

Entropy of a random variable x is given by

$$H(x) = -\sum \text{Prob}(x) \log \text{Prob}(x)$$

in case the random variable takes discrete values and

$$H(x) = -\int \mathrm{d}x \, \text{Prob}(x) \log \text{Prob}(x)$$

in case the random variable tables continuous values. One can think of the entropy as the expected value of the logarithm of the PDF of the random variable. So $H(s) = \langle \log(\text{Prob}(x)) \rangle$

Because the theory of information entropy was developed by Shannon as applied to communication, the logarithm is sometimes taken to base 2 instead of the natural logarithm and the entropy is expressed in 'bits' of information. It is sometimes called Shannon entropy. Say for example we take the toss of a fair coin. We have two outcomes each with probability of 0.5. So the entropy is

$$H = -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1$$

So a coin toss has a Shannon entropy of 1 bit.

Another way to intuitively understand entropy that is often suggested is to see it as a 'surprise value'. If someone said to you that it rained, you should not be surprised as you live in northwestern Europe. If someone said that in Death valley, California, that would be very surprising and therefore carry a much higher information entropy. Similarly a totally biased coin has no 'surprise value' when it is tosses. A fair coin on the other hand has the maximum 'surprise value' and this is why you take an interest in the toss before a game.

The negative logarithm of the PDF is chosen as the right function to measure the level of surprise or information because it has the following properties.

(1) If the probability is 1, then the entropy is zero

(2) The entropy is a monotonically decreasing function of probability. Less probable events carry more information/surprise.

(3) Entropy of different independent events is additive. If two events happen with probability p_1 and p_2 then the information you have learnt after both have happened is the sum of their individual information.

Ok so with that, now we understand the motivation to define entropy as minus the expected value (mean value) of the logarithm of the probability. Now back to the coin. Suppose the coin was biased and showed tails with a probability of 0.2 and heads with a probability of 0.8. The Shannon entropy of that distribution is

$$H = -(0.2 \log_2 0.2 + 0.8 \log_2 0.8) = 0.72 \,\text{bits}$$

which is less than the 1-bit we got for a fair coin. Why? Because the highest level of randomness is obtained when the coin is fair. The bias reduces the randomness and therefore the degree of surprise.

Good. Now let us get to the topic of assigning probabilities and what it has to do with entropy. Remember when we said so many times that we have no prior information on some parameter so we assign the uniform distribution? What we should have said was what is the distribution that has most entropy that I can assign to this variable about which I don't know anything. How to do that? Let us take a silly example coutersy of Gull & Skilling.

Suppose two-thirds of Kangaroos are right handed and one-thirds of kangaroos prefer beer over whiskey. What is the probability that a randomly selected kangaroo is a left-handed beer lover? You think this is silly? Actually, it is a very deep question! Let us understand why. The information given to us is this.

	BEER	WHISKEY	
Left-handed	$p11$	$p12$	$1/3$
Right-handed	$p21$	$p22$	$2/3$
	$1/3$	$2/3$	

where $p11$ is the probability that a kangaroo is left-handed *and* and beer-lover and so on. Be careful: this is not the marginal probability that a a kangaroo is left-handed *given* that it is a beer-lover! So we have $p11 + p12 = 1/3$, $p21 + p22 = 2/3$, $p11 + p21 = 1/3$ and $p12 + p22 = 2/3$ and we are asked to find $p11$. Ok we have 4 equations for 4 variables. It all seems alright.... Until you try to solve it that is, because the system us under-determined. Here are the possibilities

	BEER	WHISKEY	
Left-handed	$p11$	$1/3 - p11$	$1/3$
Right-handed	$1/3 - p11$	$1/3 + p11$	$2/3$
	$1/3$	$2/3$	

So $p11$ can be anything between 0 and 1/3 and be consistent with all the information. What to do? Which value to prefer? Let us calculate the entropy of this distribution. All we have to do is calculate the (minus) sum of probabilities times their logarithms. We can do this for different possible value of $p11$ and see what we get (Fig. 7.6).

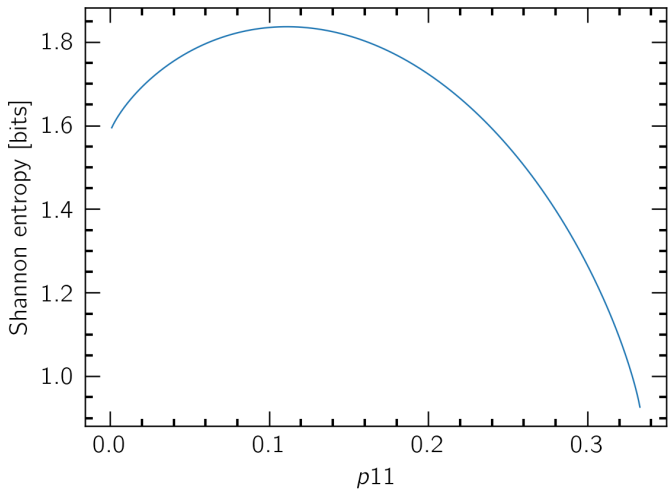


Figure 7.6: Entropy of the Kangaroos' probability distribution

Whoa! That is interesting. The entropy peaks at 0.11. So this is the value of $p11$ that gives the most randomness to the table of probabilities. Most randomness means that this is least amount of order *that you put in* by assuming $p11 = 0.11$. But what is so special about this $p11 = 0.11$. Why does the entropy curve peak here? Notice that the data tells you the probability of left- and right-handedness and it tells you about affinity to beer versus whiskey. It tells you nothing about whether left-handed kangaroors prefer beer more often than right handed kangaroos. So if you don't have any idea of the drinking preferences of left handed kangaroos, then to be fair you must choose a value of $p11$ that does not inadvertently impose a conditional probability Prob(left|beer). In other words the handedness of kangaroos and their drinking preferences should be kept independent to maximize entropy. If you don't keep them independent then you reduce the 'surprise value' because now you can say something about the handedness of kangaroos based on their drinking preferences and vice-versa. In other words, your choice of $p11$ should be such that

$$\text{Prob}(\text{left}, \text{beer}) = \text{Prob}(\text{left}) \times \text{Prob}(\text{beer}) = 1/9 = 0.11$$

That is it! The entropy peaked at $p11 = 1/9$ because picking any other value would mean that you have unknowingly introduced information that you did not have. So the fair thing to say is that there is a 1 in 9 chance that a randomly picked kangaroo is left handed and likes beer.

What we did there was choose a 'non-informative' prior. Can we extend this to more complicated situations that we encountered in our parameter estimation problems?

7.5 Jeffreys' prior

In the previous example, we had to choose a single probability value. Now we have to choose a whole distribution from among the host of candidate distributions. And this choice has to be made to suit a parametric model whose parameters are being estimated. First let us understand what is meant by a host of candidate distributions using an example that we already solved before in class— estimating the bias of a coin. In that example, we had to choose a prior on the bias b which we define as the probability of tails. We then define a uniform prior on b and said that was fair as any value of b should be made equally likely. And therein lies the problem. What if we had chosen a different formulation? What if we have chosen the following parameter:

$$\eta = \log\left(\frac{b}{1-b}\right),$$

where $\eta = -\infty$ corresponds to coin that always gives heads and $\eta = +\infty$ corresponds to a coin that always gives tails. There is nothing wrong in this choice. But now if we say, we assume no prior knowledge of η and choose a uniform prior on η, is that the same as choosing a uniform prior on b?

Let us see. Because the probability in this case is a *density* function, we have

$$\mathrm{d}\eta\,\mathrm{Prob}(\eta) = \mathrm{d}b\,\mathrm{Prob}(b)$$

Let us take our new choice of a uniform prior on η. We have $b/\mathrm{d}\eta = b(1-b)$ and hence

$$\mathrm{Prob}(b) \propto \frac{1}{b(1-b)}$$

which is obviously not a uniform distribution. In fact this new prior peaks around $b = 0$ and $b = 1$. This exercise has alerted us to the fact that just choosing a uniform prior *does* introduce some information into our calculations. We were just unaware of it! Just like choosing some random value for $p11$ in $(0, 1/3]$ did introduce information that we had no right to. Just like we had to choose $p11 = 1/9$ in the Kangaroo problem to

maximize entropy, there are prescriptions based on information theory to properly choose a prior distribution.

There is a choice for a prior distribution that is invariant under different formulations. It was suggested by Harold Jeffreys, and is unsurprisingly called Jeffreys' prior:

$$\text{Prob}(\theta) \propto \sqrt{|\mathcal{I}(\theta)|}$$

where θ is the parameter and $\mathcal{I}(\theta)$ is called the Fisher information. It is computed on the log-likelihood function, $\mathcal{L}(x|\theta) = \log L(x|\theta)$ as

$$\mathcal{I}(\theta) = \left\langle \left[\frac{\partial}{\partial\theta} \mathcal{L}(x|\theta) \right]^2 \right\rangle$$

where the angular brackets denote expected value. That is

$$\mathcal{I}(\theta) = \int \mathrm{d}x L(x|\theta) \left[\frac{\partial}{\partial\theta} \mathcal{L}(x|\theta) \right]^2 .$$

We will also state without proof here that if the log-likelihood is twice-differentiable then we the Fisher information can also be computed as

$$\mathcal{I}(\theta) = -\int \mathrm{d}x L(x|\theta) \frac{\partial^2}{\partial\theta^2} \mathcal{L}(x|\theta).$$

Jefferys' initially suggested this prior as it was invariant under monotonic transformation like the one we suggested for the coin-bias case. However it has later been recognised that Jeffreys' choice has deeper information theory roots. Unfortunately ,that is an advanced topic for another course but here is my best attempt at giving you some intuitive understanding of Fisher information. The expression for Fisher information does have the form of the definition for entropy: instead of just the expected value of the logarithm of a PDF, it is the expected value of the second derivative of the log-likelihood. Remember that the second derivative of the log-likelihood is a measure of the width of the likelihood distribution (remember it was related to the Hessian and the parameter covariance matrix?). So this second derivative is really telling us how sensitive the likelihood is to changes in the parameter. If this second derivative is very large then it means that a small change in the parameter will give a large change in the likelihood. To maximise entropy we must choose a prior distribution that maximises the 'surprise-factor', that is a distribution that has a preference for areas where the likelihood is very sensitive to the parameter. This is precisely what Jeffreys' prior does. If that explanation is not satisfactory then don't worry. At some level you will have to take Jefferys' suggestion as a recipe for now and wait for a more advanced statistics course to get into the weeds.

Anyways, let us follow this recipe for our coin bias estimation problem.

$$L(x|b) = {}^{N}C_r b^r (1-b)^{n-r}$$

where r is the observed data and b is the parameter.

$$\mathcal{L}f(r|b) = \log L(r|b) = \log({}^{N}C_r) + r\log(b) + (n-r)\log(1-b)$$

The first partial derivative is

$$\frac{\partial}{\partial b}\mathcal{L}(r|b) = 0 + \frac{r}{b} - \frac{(n-r)}{1-b}$$

And the second partial derivative is

$$\frac{\partial^2}{\partial b^2}\mathcal{L}(r|b) = -\frac{r}{b^2} - \frac{(n-r)}{(1-b)^2}$$

The expected value of the second partial derivative is (minus) the Fisher information:

$$\mathcal{I}(b) = \langle r \rangle\, b^{-2} + (n - \langle r \rangle)(1-b)^{-2}$$

The expected value of r is just nb, so

$$\mathcal{I}(b) = \frac{n}{b} + \frac{n}{(1-b)} = n\frac{1}{b(1-b)}$$

So Jeffreys' prior for the bias parameter will be

$$\text{Prob}(b) \propto \frac{1}{\sqrt{b(1-b)}}$$

This is interesting, as it asks use to choose a prior that does not treat all biases equally. But why? To understand this, we must compute the variance of the Binomial distribution, ${}^{n}C_r b^r (1-b)^{n-r}$. I will skip the algebra and state the result: the variance is given by $nb(1-b)$. Notice that the variance depends on the parameter b and its value tends to zero as the bias b tends to either 0 or 1. Low variance means tighter constraints from the new data. So the prior prioritizes these regions where the new data has the most impact in providing constraints. More generally, Jeffreys' prior is design such that the posterior is invariant under monotonic transformation of the parameter to be estimated. So using this prior is 'non-informative' in that your answer will not depend on how you formulate the problem.

7.6 Thumb rules for prior assignment

Jeffreys' prior is not the be-all-end-all of prior assignment: it has its advantages and disadvantages. Going deeper into the philosophy behind prior assignment and the information theoretic reasons is beyond the scope of an introductory Bachelor's source. Yet you will have to make choices, so I will leave you with some rules of thumb:

- If you only know the expected value of a parameter or random variable (say x) then the maximum-entropy prior should be used, which is the exponential distribution: $\text{Prob}(x) = \lambda \exp(-\lambda x)$.

- If you know the mean and variance of a random variable then the maximum-entropy prior should be used which is a Gaussian distribution: $\text{Prob}(x) = (2\pi\sigma^2) \exp(-(x-\mu)^2/(2\sigma^2))$. This is the most common case you will encounter as you will have some information on the parameter you are trying to estimate from previous experiment/research/theory.

- If you know absolutely zilch about the parameter you are trying to estimate, then you will have to first identify the kind of variable you are dealing with and go from there:
  - If the parameter is denoting a proportion (e.g. number of heads in n tosses, or the fraction of male lambs born in a farm) then you know that the likelihood will have a binomial form and you should use Jeffrey's prior: $\text{Prob}(b) \propto [b(1-b)]^{-1/2}$.
  - If the parameter is denoting a rate at which discrete events occur, i.e. if you are dealing with a Poisson process, then you should use Jeffreys' prior: $\text{Prob}(\lambda) \propto \lambda^{-1/2}$.
  - If the parameter is the mean of a Gaussian process then you should use Jeffreys' prior on the mean: a uniform distribution.
  - Similarly if the parameter is the standard deviation, then you should use a flat prior on its logarithm: $\text{Prob}(\log(\sigma)) \propto 1$ or $\text{Prob}(\sigma) \propto \sigma^{-1}$.
  - For more complex model parameters, use a uniform prior on the parameter if it is an offset parameter (details below) and use a uniform prior on the logarithm of the parameter if it is a scale parameter (details below).

7.6.1 Scale and offset parameters

An offset parameter is one that shifts the origin of some co-ordinate system inherent to the problem and a scale parameter is one that stretches and squeezes the co-ordinate system inherent to the problem. Take the example from Chapter 5 of fitting a sinusoid to the radial velocity measurements to detect an exoplanet. The model we fit was of the form

$$A\cos\left(2\pi Pt+\phi\right),$$

where A, P and ϕ were parameters to estimate. Now imagine the graph of radial velocity measurements on the y axis and time on the x axis. What will happen as we vary the value of ϕ? The entire graph will shift to the left and right along the x-axis. So ϕ is an offset or translational parameter. We should assign a uniform prior on ϕ:

$$\text{Prob}(\phi)\propto 1$$

What about A. If we vary A then the the graph stretches and squeezed along the y axis. It does not move up and down. So A is a scale parameter and we should assign the prior

$$\text{Prob}(A)\propto A^{-1};\ \text{Prob}(\log(A))\propto 1$$

What about the frequency P? If we change P then we will stretch and squeeze the graph along the x axis (not move it left and right). So we again have

$$\text{Prob}(P)\propto P^{-1};\ \text{Prob}(\log(P))\propto 1.$$

What about the spectral-line fitting example from Chapter 6? There the centre of the line μ is an offset parameter so we should use a uniform prior on it. But the depth of the line τ is a scale parameter so we should use a uniform prior on its logarithm.

What about the straight line fit we did in Chapter 4 (Hubble's law)? This one needs a bit more thought. Our first instinct is to say that in the model $v=Hd+C$, C is an offset parameter and H is a scale parameter. This is not correct; changing H does not stretch of squeeze the graph, it rotates the line. So the correct way to think about this model is to say that the orientation of the line, θ, where θ is the angle made with the x axis should be uniformly distributed. H is related to θ according to $H=\tan\theta$. What about the intercept with the y axis C? Here we have to make a choice that does not depend on formulating the model in terms of the y axis intercept (we could have chosen an x axis intercept with equal validity). An impartial choice is the offset of the line from the origin which is given by

$C\cos\theta$. So our prior should be a uniform distribution for $\tan\theta$ and $C\cos\theta$. If one goes through the algebra of PDF transformation we get the following joint prior:

$$\text{Prob}(H, C) \propto (1 + H^2)^{-3/2}$$

That ends our discussion on the assignment of priors. The last topic I wanted to cover in this chapter is the Kolmogorov–Smirnov test but we have run out of time! So we must end the notes here. I will leave open the possibility to type up some notes based on questions you will ask me in the last section, given below.

7.7 Assorted topics

7.7.1 Propagation of uncertainties

We already learnt how to transform the PDF of a variable x into the PDF of a new but related variable $y = f(x)$. If $f(x)$ is a monotonic function then we showed that

$$\text{Prob}(y) = \text{Prob}(x)\left|\frac{\partial x}{\partial y}\right|$$

Suppose x was a variable that you estimated but what you are really interested in is the variable y. You now know how to use the posterior distribution of x to get the posterior distribution of y. Once you have the distribution of y, you have all the information you need. But sometimes you just need a point estimate for y (e.g. expected value) and a simple measure of its uncertainty (e.g. $y = 18.9 \pm 0.2$). For these cases, it makes sense to work out a simple expression for the transformation of not PDFs but just variances.

OK let us work this out. We will start with the definition of variance[2]

$$\sigma^2(y) = \left\langle [y - \langle y \rangle]^2 \right\rangle$$

Let us suppose that our function f is linear, so we have $y = f(x) = ax+b$ where a and b are constants. Then we have

$$\begin{aligned}\sigma^2(y) &= \left\langle [ax + b - \langle ax + b \rangle]^2 \right\rangle \\ &= \left\langle [ax + b - a < x > -b]^2 \right\rangle \\ &= a^2 \left\langle [x - < x >]^2 \right\rangle = a^2\sigma^2(x)\end{aligned}$$

[2]Note that in literature, σ^2 is often used for variance, regardless of whether the distribution is Gaussian or not.

So we see that the offset b has no effect on the uncertainties but the scale factor a does have a significant impact. This makes intuitive sense because the offset cancels out when we take *central* moments whereas the scale factor being multiplicative also multiplies the uncertainties.

Let us increase the complexity by one level. Suppose we use a matrix relationship between the parameter vector $\boldsymbol{x}$ and the variable vector we are interested in which is $\boldsymbol{y}$:

$$\boldsymbol{y} = \boldsymbol{A} \cdot \boldsymbol{x}$$

Now we must talk about the *covariance matrix* of the parameters instead of a single variance:

$$\boldsymbol{\Sigma}(\boldsymbol{x}) = \begin{bmatrix} \sigma^2(x_1) & \sigma^2(x_1, x_2) & \sigma^2(x_1, x_3) & ... \\ \sigma^2(x_2, x_1) & \sigma^2(x_2, x_2) & \sigma^2(x_2, x_3) & ... \\ ... & & & \\ .. & .. & .. & \sigma^2(x_n, x_n) \end{bmatrix}$$

We are interested in the covariance matrix of the new parameter vector $\boldsymbol{y}$. This can be shown to be

$$\boldsymbol{\Sigma}(\boldsymbol{y}) = \boldsymbol{A} \cdot \boldsymbol{\Sigma}(\boldsymbol{x}) \cdot \boldsymbol{A}^T$$

Finally, what if the transformation $y = f(x)$ is non-linear in x? Well then unfortunately we do not have a simple closed form expression for the variance of y. We will have to linearise the model around the central values to propagate uncertainties in an approximate way. We will then write

$$y = f(x) \approx f(x_0) + (x - x_0)f'(x_0) = xf'(x_0) + f(x_0) - x_0 f'(x_0)$$

By analogy with our linear case, we see that the scale factor if $f'(x_0)$ and therefore

$$\sigma^2(y) = \left[f'(x_0)\right]^2 \sigma^2(x)$$

Here is a commonly encountered error propagation situation. I will leave it to you to prove these result to get some practice and insight. The notation used is that x_i are parameters and a_i are constants

$$y = a_1 x_1 + a_2 x_2;\ \sigma^2(y) = a_1^2 \sigma^2(x_1) + a_2^2 \sigma^2(x_2) + 2a_1 a_2 \sigma^2(x_1, x_2)$$