Statistics of Measurement (Lectures 1-4)
Lecture 1 - Introduction and properties of probabilities

Heather Graven, 12 May 2023

1 Introduction

Statistics is a study of two somewhat separate issues, which we shall call **probability** and **estimation** (although people will often use 'statistics' to mean just the latter). For physicists, the main aim of studying statistics is to be able to interpret the results of experiments so as to derive information on the values and uncertainties of physical quantities. This is essentially the 'estimation' part; we estimate values, e.g. for the mass of the electron, using data we acquire during an experiment.

However, no experiment can measure a quantity *exactly*, so there will inevitably be some level of random error associated with each value obtained. Hence, estimation requires a knowledge of 'probability', and so in essence, we must study this aspect first. We can think of the study of probability as being the way of predicting how often we would get the various possible outcomes of our experiment, e.g. how often we would measure the electron mass to have some particular value. Hence, you can think of the two as going in opposite directions; i.e. studying probability allows one to take physical constants and predict the likely outcomes of an experiment, whilst estimation takes the actual outcome of a particular experiment and finds the values of physical constants which are compatible with the data. You will sometimes hear estimation referred to as the 'inverse problem' to probability.

One thing to note however is that 'probability' and 'estimation' are two very different topics: Probability is a mathematically rigorous field; whilst, estimation is less straightforward, often with no absolutely 'right' answer and is to some extent arbitrary. As we will see later, there is even a basic disagreement at the fundamental level as to what we should be estimating. It is common for students to be fairly comfortable with probability but can be confused with estimation; I hope the latter will not be true for you. We will cover probability in the first part of the lecture course and estimation in the second part. We will not deal with systematic errors greatly, but you should be aware of these, partly through your experience in labs this year.

2 Frequentist and Bayesian probabilities

Everyone will have an intuitive feel for what we mean by a probability; e.g. "The probability of throwing a die and getting a six is 1/6". However, it turns out that actually defining probability can be quite tricky.

The most obvious definition would be that the probability of something is the fraction of times it occurs when the number of 'experiments', sometimes called 'trials', becomes very large. In fact the actual definition would be the limit of this fraction as the number of experiments tends toward infinity. This is called the **frequentist** (or 'classical') definition, as it defines probability through the frequency with which something occurs in a large number of experiments. This definition means that the probability will be a fixed value for a given experimental setup and everyone will use the same value. Note, the critical concept here is *repeatability*.

In contrast, the **Bayesian** idea of probability (named after Thomas Bayes) is that it is related to the *degree of belief* in the outcome, which will depend on the available knowledge of the experimenter. Depending on the extent of their knowledge, different people might therefore assign different probabilities to the outcome for the same experiment. The positive side is that this gets around an issue with the frequentist definition, which is that it is often difficult or impossible to repeat the experiment. The most extreme example would be cosmology studies

of the development of the Universe, which can predict e.g. the probability distribution for sizes of galaxies. We cannot redo the 'experiment' of the creation of the Universe to check our probabilities and it is not clear if a frequentist probability definition is even conceptually sensible. The Bayesian approach can be seen as generalising the frequentist definition. However, the downside is that it does mean that probabilities can be harder to evaluate uniquely in the Bayesian approach. For most of the estimation part of the course we will take the frequentist view, while in the last lecture and the second seminar we will discuss the Bayesian approach explicitly.

3 General properties

Although the definition of probability is tricky, the mathematical properties of probabilities are uncontentious and well-defined. If the possible outcomes of an 'experiment' (in the general sense, e.g. throwing a die) are x_i with probabilities $P(x_i) = P_i$, then the P_i have to satisfy the **axioms of probability** introduced by Andrey Kolmogorov in 1933. The probability of an event is a non-negative real number, the probability of x_i or x_j is the sum of the probabilites of i and j, and the sum of probabilities of all possible outcomes is 1:

$$P_i \geq 0, \qquad P_{i \text{ or } j} = P_i + P_j, \qquad \sum_i P_i = 1$$

Many other properties can be derived from these; e.g. clearly every $P_i \leq 1$, so probabilities are bounded between 0 and 1. Also, the probability of something not happening must be

$$P_{\text{not } i} = \overline{P_i} = 1 - P_i = \sum_{j \neq i} P_j.$$

4 Random variables and probability distributions

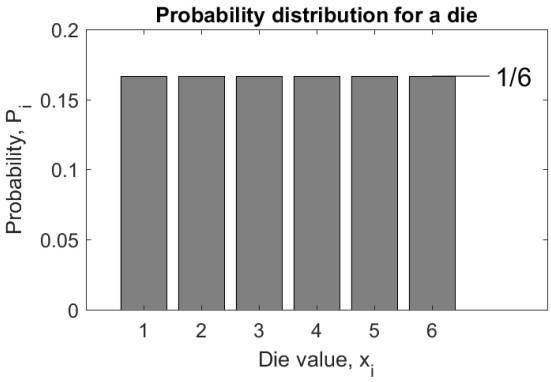
Formally, for any given 'experiment', the possible outcomes are called the 'sample space' and each outcome will have a probability assigned to it. e.g. if I throw a fair die, the sample space is the set of values {1, 2, 3, 4, 5, 6} and by symmetry, each has a probability of 1/6 so that they sum to one as required. If I throw a fair coin, the sample space is {Heads, Tails} and again by symmetry, each must have a probability of 1/2 so as to sum to one. The quantities which we measure, and which can take any one of the values in the sample space, are called 'random variables'; in these examples they are the top face of the die or coin. Each time we do an experiment, the outcome is randomly picked from one of the outcomes in the sample space, with an appropriate probability assigned to it.

In both these cases, the probabilities P_i for each of the outcomes within the sample space x_i are equal, by assumption. The probability distribution is **uniform**. For a uniform distribution with N possible outcomes, let $P_i = k$, a constant. We can use the normalisation of all the probabilities to fix the constant; since

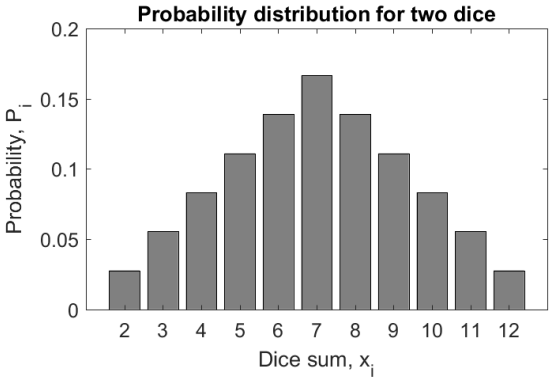
$$\sum_{i=1}^{N} P_i = Nk = 1 \qquad \text{then} \qquad P_i = k = \frac{1}{N}$$

which agrees with the probabilities we assigned above, with $N = 6$ for the die and $N = 2$ for the coin. We can draw such a probability distribution for all elements in the sample space as a graph; e.g. the case of the die is shown below.

In general, probabilities are NOT typically all equal; for example, the sum of two dice is not uniform, even though the outcome of each die independently is. This is straightforward to



understand; any given outcome of the first die and the second die has a probability of $(1/6)^2 = 1/36$. There is only one way to make the sum be 2 (or 12), namely two ones (or two sixes), and so these indeed each have a probability of 1/36. However, there are two ways to make a sum of 3, namely a one on the first die and a two on the second, or vice versa. Therefore, a sum of 3 has a probability equal to the sum of the separate probabilities (using Kolmogorov's axioms) and hence is $1/36 + 1/36 = 1/18$. The same holds for 11. The probabilities increase for larger sums than 3, as there are more and more combinations which give a particular sum, and peak at the sum of 7, which has six combinations and so a probability of $6/36 = 1/6$. The probability distribution for this is shown below. Note here, as in all probability distributions, the total area under the distribution must sum to one to satisfy Kolmogorov's axioms.



5 Expectation value and variance

The **expectation value**, or 'expected' value, is the long-run average value of the distribution.

You can think of it as the average that would be obtained from a very large number of experiments, given the distribution. Since the probability of getting a value x_i is P_i, then for $N \to \infty$ experiments, the number of outcomes equal to x_i would be NP_i (from the frequentist definition of probability). The expectation value is then given by

$$E(x) = \langle x \rangle = \frac{1}{N} \sum_i x_i (NP_i) = \sum_i x_i P_i$$

For the case of two dice, this gives

$$E(x) = \langle x \rangle = \frac{1}{36}(2+12) + \frac{2}{36}(3+11) + \frac{3}{36}(4+10) + \frac{4}{36}(5+9) + \frac{5}{36}(6+8) + \frac{6}{36}(7) = 7$$

as is obvious from the symmetry of the probability distribution. For any uniform distribution, the expectation value is

$$E(x) = \langle x \rangle = \sum_i x_i \frac{1}{N} = \frac{1}{N} \sum_i x_i$$

i.e. what we think of as the usual average. For a single die, then this would be

$$E(x) = \langle x \rangle = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = \frac{7}{2} = 3.5$$

This shows that the expectation value does *not* have to be one of the possible outcomes from the probability distribution in general; it happens to be so for two dice but not for one die.

It is often useful to have a measure of the spread of values in a distribution. The **variance** is the expectation value of the square of the difference of the x_i from the average value

$$V(x) = \langle\ (x - \langle x \rangle)^2\ \rangle$$

This can be written in a more useful form

$$V(x) = \langle x^2 - 2x\langle x \rangle + \langle x \rangle^2 \rangle = \langle x^2 \rangle - 2\langle x \rangle \langle x \rangle + \langle x \rangle^2 \langle 1 \rangle = \langle x^2 \rangle - \langle x \rangle^2$$

which means

$$V(x) = \left(\sum_i x_i^2 P_i \right) - \left(\sum_i x_i P_i \right)^2$$

Variance is commonly used rather than just the average deviation from the mean because positive and negative deviations will cancel to get an average deviation of zero. Also, as you have used in labs, the variance of the sum of two independent distributions is the sum of their variances, but the same is not true for the deviation from the mean. To get a quantity related to the spread which has dimensions of $[x]$, we take the square root of the variance to calculate the **standard deviation**.

For a die, the variance is

$$V(x) = \frac{1}{6}(1+4+9+16+25+36) - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12} = 2.92$$

and so the standard deviation is $\sqrt{2.92} = 1.71$. This gives some idea of the spread of the distribution.

6 Two dimensional probability distributions

So far we have discussed probability distributions of x_i in one dimension. It is possible to generalise the idea into more than one dimension, e.g. for two dimensions, where we would have x and y random variables related to two different sets of outcomes.

Let's take a simple example of rolling a die and also tossing a coin, so x is the die outcome and y the coin outcome. The x_i all have probability $1/6$ and the y_i have $1/2$. We define the **joint probability** for the outcomes x_i and y_j as $P_{ij} = P(x_i, y_j)$. If I roll the die and toss the coin, and ask for the probability of getting a six on the die as well as getting heads on the coin, then the probabilities are multiplied

$$P(\text{Six},\ \text{Heads}) = P(\text{Six}) \times P(\text{Heads}) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

However, as we will see, this is true only because they are **independent**. The probability of the die outcome is unrelated to the coin outcome so these two are indeed independent. It is often convenient to display these values as a table, showing the probabilities of the possible outcomes we are interested in, as shown below, where for simplicity we only care if we get a six or not.

		Heads	Tails
		1/2	1/2
Six	1/6	1/12	1/12
Not-six	5/6	5/12	5/12

The table above is a representation of the two-dimensional probability distribution. Each entry in the main part of the table corresponds to one of the joint probabilities; for example $P(\text{Six, Heads})$ corresponds to the upper left entry. In this case, the probability in each entry in the table is the product of the two overall probabilities, again only because they are independent. The columns sum to the probabilities written along the top, which give the probabilities for the coin outcome by itself. Similarly, the rows sum to those on the left side, which are the die outcome probabilities. [Nb. Strictly speaking, the case of two dice discussed above is another independent 2D case, but by only being interested in the sum of the dice values, we reduced it to a 1D distribution.]

In general, the probabilities in a 2D distribution do not have to be independent; if not, they are said to be dependent or **correlated**. Here's an example that should illustrate the point: within the UK population, 50% of people are female and 50% are male. Also, approximately 1% of people are pregnant and 99% are not pregnant. However, the probability of being male and pregnant is not

$$P(\text{Male, Pregnant}) = P(\text{Male}) \times P(\text{Pregnant}) = 0.5\%$$

but is effectively zero. The equivalent table in this case is shown below.

		Male	Female
		50%	50%
Pregnant	1%	0%	1%
Not pregnant	99%	50%	49%

Note the entries, i.e. the joint probabilities, are *not* the products of the overall probabilities in this case. Hence, in general the joint probabilities only factorise to $P(x_i, y_j) = P(x_i) \times P(y_j)$ in the special case of independent variables x and y. In the general case, having dependent variables means that if we know x_i, then the probability for y_j is different from what it would be if we knew nothing about x, i.e. $P(y_j)$ by itself. We write this as $P(y_j|x_i)$, the **conditional probability**, which should be read as the probability of getting y_j, given that x_i is known. Similarly, if we know y_j, then the probability of getting x_i is changed from $P(x_i)$ to $P(x_i|y_j)$. For the special case of independent variables, then $P(x_i|y_j) = P(x_i)$ (as it is not dependent on y_j) and similarly $P(y_j|x_i) = P(y_j)$.

Since the joint probability of x_i and y_j happening together must be given by the probability of x_i and then the conditional probability of y_j given x_i, or the probability of y_j and then the conditional probability of x_i given y_j, then we have

$$P(x_i, y_j) = P(x_i)P(y_j|x_i) = P(y_j)P(x_i|y_j) \tag{1}$$

We can find the conditional probabilities from the joint probabilities using Equation 1 above; for example

$$P(y_j|x_i) = \frac{P(x_i, y_j)}{P(x_i)}$$

The conditional probability is a probability distribution in the first variable only, not the second (as that is known). Hence, the normalisation of the probability distribution requires for any x_i

$$\sum_j P(y_j|x_i) = 1$$

Be careful when using conditional probabilities as they are *not* equal to each other. For the example above, having selected a person who is female, the probability of the person also being pregnant is

$$P(\text{Pregnant}|\text{Female}) = \frac{P(\text{Female}, \text{Pregnant})}{P(\text{Female})} = \frac{0.01}{0.50} = 2\%$$

while having selected a person who is pregnant, the probability of them also being female is

$$P(\text{Female}|\text{Pregnant}) = \frac{P(\text{Female}, \text{Pregnant})}{P(\text{Pregnant})} = \frac{0.01}{0.01} = 100\%$$

as you would expect.

7 Bayes' theorem

The equality of the last two terms in equation 1 is called **Bayes' Theorem** (after the same Thomas Bayes for which the probability definition above is named). This allows us to convert between the two conditional probabilities, i.e.

$$P(y_j|x_i) = \frac{P(x_i|y_j)P(y_j)}{P(x_i)}$$

Essentially, we can use Bayes' theorem to improve an estimate of the probability of y_j given new information about the probability of x_i and the conditional probability of x_i given y_j. In the example above, if we want to know the probability of a person being pregnant without any other information, our first guess would be the probability of any person being pregnant, 1%. Then, if we received new information that the person was also female, we could use Bayes' Theorem to update the probability of that person being pregnant, given that the person was female

$$P(\text{Pregnant}|\text{Female}) = \frac{P(\text{Female}|\text{Pregnant})P(\text{Pregnant})}{P(\text{Female})} = \frac{1 * 0.01}{0.50} = 2\%$$

We thus scale the initial, or prior, probability with the new information to produce an improved posterior estimate.

We do need to be careful with the denominator here. We can combine the rules above to give the 'marginalisation rule'. The total probability for any x_i must be the sum of the joint probabilities of x_i with every possible outcome of y, i.e.

$$P(x_i) = \sum_j P(x_i, y_j) = \sum_j P(x_i|y_j)P(y_j)$$

This is often the best way to calculate the denominator when using Bayes' theorem. Bayes' theorem itself then reads

$$P(y_j|x_i) = \frac{P(x_i|y_j)P(y_j)}{\sum_{j'} P(x_i|y_{j'})P(y_{j'})}$$

Hence the denominator effectively acts as a normalisation for the terms in the numerator so that

$$\sum_j P(y_j|x_i) = 1$$

for any x_i, as required.

Bayes' theorem is incredibly useful and it has applications today in the modelling of climate change, astrophysics and stock market analysis to name but a few examples. Bayes' Theorem can also be used for probability distributions and in cases where the prior or new information is not very precise but still useful in constraining a parameter of interest. We will return to Bayesian estimation towards the end of this course.

Statistics of Measurement
Lecture 2 - Discrete probability distributions

Heather Graven, 15 May 2023

8 Introduction

We have seen a few probability distributions in the last lecture, namely those for tossing a coin, and throwing one or two dice. These are examples of discrete (as opposed to continuous) distributions, meaning the outcomes can only have specific values, e.g. 'Heads' or Tails' for the coin, or integers for the dice. We will look at two important discrete distributions for physics in some detail in this lecture and will discuss continuous distributions in the next lecture. The two distributions considered today are the **binomial** and the **Poisson** distributions.

9 Binomial distribution

In the last lecture, we considered tossing a coin to get heads or tails. What if we toss it 10 times; what is the probability of getting e.g. 3 heads? For a process with just **two possible outcomes**, that is independently repeated N times (or 'trials'), the binomial distribution tells us the probability of getting n of one of the two outcomes. This clearly applies to tossing a coin, where there are obviously only two outcomes. However, it seems not to apply to a die, where there are six outcomes. But this is a matter of what we are interested in; if we restrict ourselves to asking if the die gives a six or not-six, then we have reduced the situation to two outcomes and so can use the binomial for this too.

Let's take the six or not-six example and see how it works. Take a simple case; how many times will we get two sixes with three throws of a die? Clearly, with one throw of the die, we will get six with probability 1/6 and not-six with probability 5/6. Let $p = 1/6$ so $5/6 = 1 - p$. (Note, $1-p$ is often written as q in some textbooks.) We could get a six on the first throw (with probability p) and again on the second throw (again with probability p) and then a not-six on the last throw (with probability $1 - p$), giving a total probability of $p \times p \times (1 - p) = p^2(1 - p)$ as they are independent. However, there are clearly other ways to get two sixes. One is to get a six, a not-six and then a six, with probability $p \times (1 - p) \times p = p^2(1 - p)$ and the other is to get not-six, six, six with probability $(1 - p) \times p \times p = p^2(1 - p)$. Hence, each of these combinations with two sixes and one not-six has the same probability. If we are not interested in the order in which the sixes appear, but only in the total number of sixes, then (by Kolmogorov) we add the probabilities for all combinations of getting two sixes, which here gives $3p^2(1 - p)$.

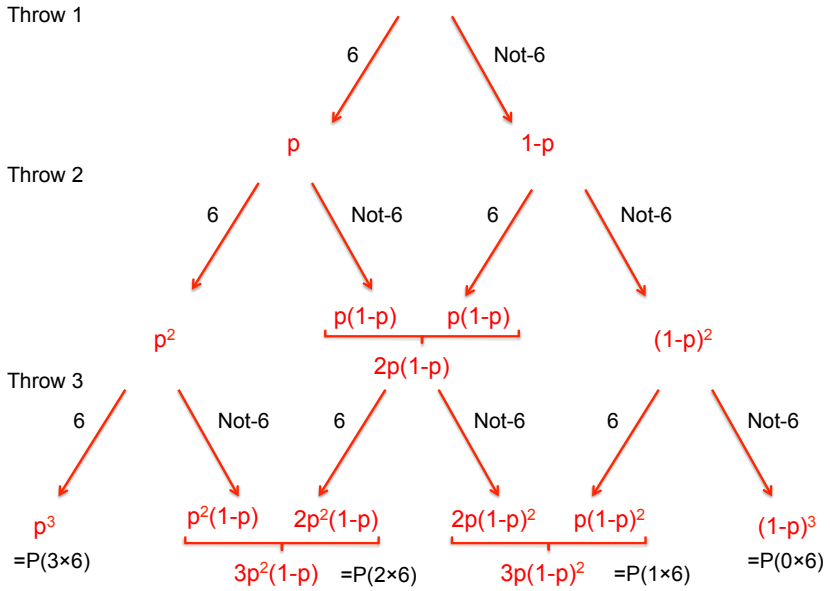
In general, for N throws, then to get n sixes, and hence $N - n$ not-sixes, the probability of any particular combination is $p^n(1 - p)^{N-n}$. The numbers of combinations is given by the 'binomial coefficient'

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

so-called as they are the values in **Pascal's triangle**, which are the coefficients which appear in a binomial expansion

$$(a+b)^N = \sum_{n=0}^{N} \binom{N}{n} a^n b^{N-n} = \sum_{n=0}^{N} \frac{N!}{n!(N-n)!} a^n b^{N-n}$$

The relationship between these two is made clearer in the figure, showing the various possible outcomes of three dice throws.



Hence, the probability of getting n sixes out of N throws is given by

$$B(n;p,N) = \frac{N!}{n!(N-n)!}\, p^n (1-p)^{N-n}$$

where 0! is defined to be 1. For $n = 2$ and $N = 3$, we get

$$B(2;p,3) = \frac{3!}{2!1!}\, p^2(1-p) = 3p^2(1-p)$$

as before. For $p = 1/6$, this is numerically $= 0.069$.

Note the notation here; the random variable appears first and the parameters of the distribution appear after a semi-colon. The above expression is called the **binomial distribution**, named because of its close relation to the binomial expansion. It holds for any case of two (hence 'bi' in 'binomial') possible outcomes per independent trial. For example, a given nucleus will decay in some time period with a probability p; there are two outcomes at the end of the time period; it has decayed or not. If we have N such nuclei, then each is an independent 'trial' and the total number that have decayed will be given by the binomial distribution.

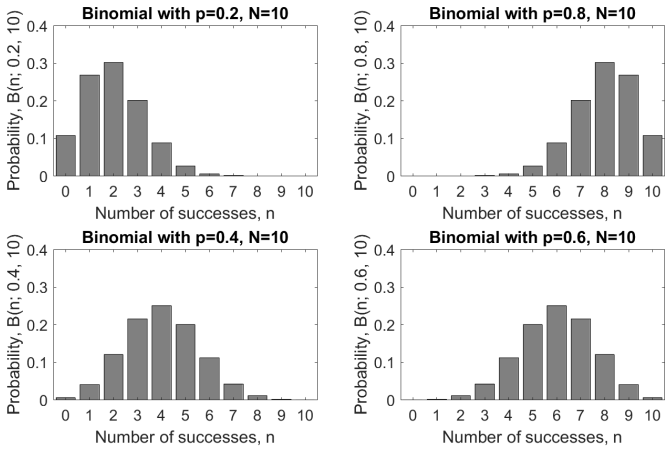
Another example: the probability of getting heads 3 times when tossing a fair coin 10 times is

$$B(3;0.5,10) = \frac{10!}{3!7!}\, 0.5^3\, 0.5^7 = \frac{10\times 9\times 8}{3\times 2\times 1}(0.5)^{10} = 0.117$$

Some binomial distributions for varying p and keeping N fixed at 10 are shown below. Clearly n can only have values between 0 and N. Basically, the outcome of each trial is one of two categories and the number of outcomes of one of these categories has the probability given by the binomial distribution. However, it should be clear that the other category will also follow a binomial as they are both possible outcomes. Specifically, the probability of the other is clearly $p' = 1 - p$ so $p = 1 - p'$, and the number of times the other occurs will be $n' = N - n$ so $n = N - n'$, such that

$$B(n;p,N) = \frac{N!}{n!(N-n)!}\, p^n (1-p)^{N-n} = \frac{N!}{(N-n')!n'!}\, (1-p')^{N-n'} p'^{n'} = B(n';p',N)$$

Hence, **the binomial is completely symmetric** with regard to which of the two outcomes is considered. This symmetry can be seen in the plots; the plots for $p = 0.2$ and 0.8 are mirror images of each other, as are $p = 0.4$ and 0.6. Note the particular outcome being considered is often described as a *success* but the other outcome, the *failure*, also follows the binomial distribution.



The relationship to the binomial expansion makes it easy to see that the probabilities sum to one. The binomial expansion states

$$(a+b)^N = \sum_{n=0}^{N} \frac{N!}{n!(N-n)!}\, a^n b^{N-n}$$

With $a = p$ and $b = 1 - p$, this reads

$$[p + (1-p)]^N = 1^N = 1 = \sum_{n=0}^{N} \frac{N!}{n!(N-n)!}\, p^n (1-p)^{N-n} = \sum_{n=0}^{N} B(n; p, N)$$

Since the right-hand side is the sum over all the binomial distribution probabilities, the distribution is indeed normalised as required.

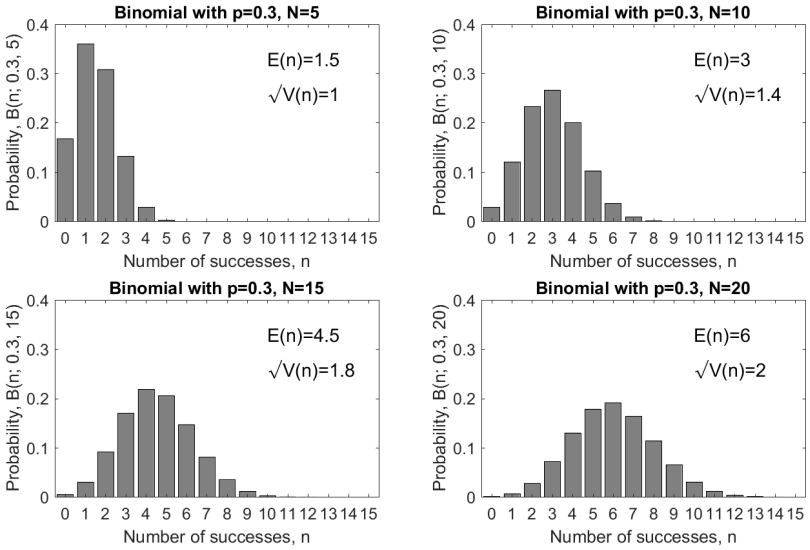
We defined the expectation value and variance of the outcomes for a probability distribution in the last lecture. These can be evaluated for a binomial (see the appendix) and are

$$E(n) = pN, \qquad V(n) = Np(1-p)$$

and hence the standard deviation is $\sqrt{Np(1-p)}$. The expectation value is the intuitive result; if there is a probability per trial of p, then the average number we would expect in N trials is pN. Plots of binomial distributions with the expectation values and standard deviations for p fixed at 0.3 but varying N are shown below. It can be seen how the expectation values and standard deviations increase with N. As you would expect, the number of successes increases with the number of trials, and the probability distribution spreads out as there are more potential combinations of outcomes.

10 Non-examinable: multinomial distributions

If we want to know the probability of throwing a die and getting n_6 sixes and n_5 fives out of N throws, then we need to go beyond the binomial distribution. We can actually find the



probability for this using the binomial twice. Firstly, it gives us the probability for n_6 sixes and therefore $N - n_6$ not-sixes, as before. However, we can then consider the not-sixes and use another binomial to give the probability of n_5 fives out of the $N - n_6$ not-sixes. The total probability is the product of the two terms. Since there are three possible outcomes (six, five and not-six-or-five) then the resulting distribution is called a trinomial, rather than a binomial. You will probably not be surprised to know that the coefficients in the trinomial distribution are the same as those that arise in the expansion of $(a+b+c)^N$. Also, this is now a 2D probability distribution and the variables are correlated; this is clear as e.g. a high number for n_6 leaves less possibilities for n_5 given a fixed total of N throws. This generalises to a 'multinomial' distribution for more possible outcomes

$$M(n_i; p_i, N) = N! \prod_i \frac{p_i^{n_i}}{n_i!}$$

where i runs over the outcomes, p_i is the probability of each outcome (with the constraint that $\sum_i p_i = 1$) and n_i are the observed numbers for each outcome, with $N = \sum_i n_i$ being the total number of trials.

11 Poisson distribution

The Poisson distribution describes the probability of obtaining a number of occurrences n given an average number μ, where the probability of each occurrence is independent. The distribution is

$$P(n; \mu) = \frac{\mu^n e^{-\mu}}{n!}$$

Here, n is unbounded above and so can take any discrete value from 0 to ∞. Some examples: the number of photons from a patch of sky in a particular photograph, the number of lightening strikes per hour during a storm, the number of buses passing through a road junction, or the number of car crashes on the M1 per year. It differs from the binomial because the occurances happen in a continuum, i.e. there is no specific number of trials and we can't say how often we didn't get an occurrence; e.g. it is meaningless to ask how many times a bus *didn't* go through the junction.

It is again easy to check the probabilities sum to one by using the standard expansion of an exponential

$$e^{\mu} = 1 + \mu + \frac{1}{2}\mu^2 + \ldots = \sum_{n=0}^{\infty} \frac{\mu^n}{n!}$$

Hence multiplying by $e^{-\mu}$ gives

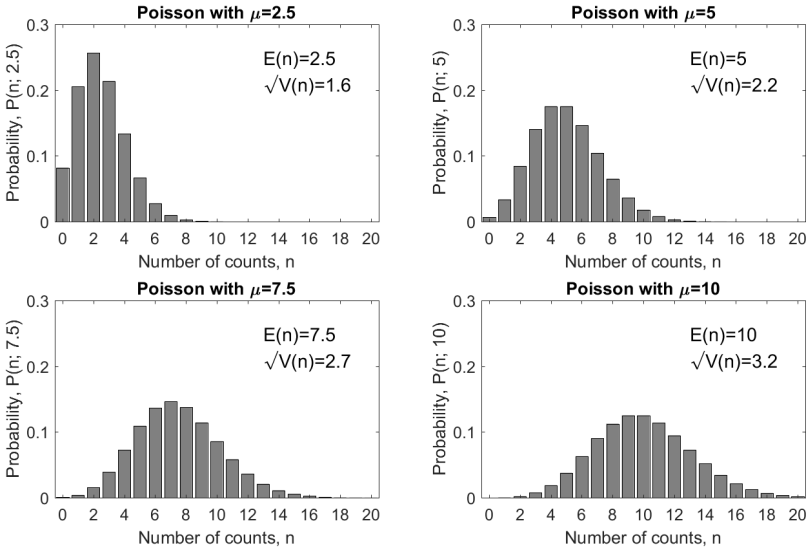
$$e^{\mu}e^{-\mu} = 1 = \sum_{n=0}^{\infty} \frac{\mu^n e^{-\mu}}{n!} = \sum_{n=0}^{\infty} P(n;\mu)$$

as required.

The expectation value and variance of a Poisson distribution can be found (see the appendix) to be

$$E(n) = \mu, \qquad V(n) = \mu$$

which means the standard deviation is $\sqrt{\mu}$. The expression for the expectation value shows that μ is indeed the average of the number of occurrences seen, as originally stated. Note, the parameter μ, being the average number, is not necessarily an integer, while the outcome n is the number seen and must be an integer. Some examples of Poisson distributions with the expectation value and standard deviation marked on them are shown below. It can be seen how the expectation value increases and the distribution spreads out as μ increases, similar to what we saw for increasing N in the binomial distribution.



In many cases, the Poisson mean μ is basically calculated from the rate of some process being measured for a fixed time or area. Take the example of photons from a patch of sky; if the weather is stable, there should be a constant rate of photons per unit time, λ. Hence, the average number of photons in time t would be $\mu = \lambda t$. Therefore, knowing the rate allows us to find the probabilities for n occurrences for whatever length of time t that we run our experiment for, i.e.

$$P(n;\mu = \lambda t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

Similarly, if we know the average density of buses per km^2 in London, ρ, then we can calculate the probability of n buses in an area A using $\mu = \rho A$. It is common to see the Poisson distribution written in forms like these.

12 Poisson as a limit of binomial

One critical difference between a Poisson and a binomial is that in the Poisson, the number of occurrences n can go to infinity, while in a binomial, it has a maximum of N. However, if p is small in the binomial, then the probability of getting n nearly as large as N is very small and then the upper limit becomes irrelevant. Let's consider the binomial as N gets very large and p gets very small, but with the constraint that pN remains finite. The binomial is

$$B(n;p,N) = \frac{N!}{n!(N-n)!}\, p^n(1-p)^{N-n}$$

Consider the numerator of the binomial coefficient $N!$. This is

$$N! = N(N-1)\dots(N-n+1)(N-n)(N-n-1)\dots 1 = N(N-1)\dots(N-n+1)\times(N-n)!$$

Hence, the numerator includes the $(N-n)!$ in the denominator as a factor and so these can be cancelled. (In fact, this is often the best way to calculate the binomial coefficients.) This leaves the product of n terms above, so

$$B(n;p,N) = \frac{N(N-1)\dots(N-n+1)}{n!}\, p^n(1-p)^{N-n}$$

In the limit of large N and small p, then the only values of n with non-negligible probabilities are small, i.e. $n \ll N$. Hence, all terms like $N-1$ or $N-n$ can be approximated to N, so the binomial is approximately

$$B(n;p,N) \approx \frac{N^n}{n!}\, p^n(1-p)^N$$

It is a standard result that in the limit of $N \to \infty$ and $p \to 0$, such that pN remains finite, then

$$(1-p)^N \to e^{-pN}$$

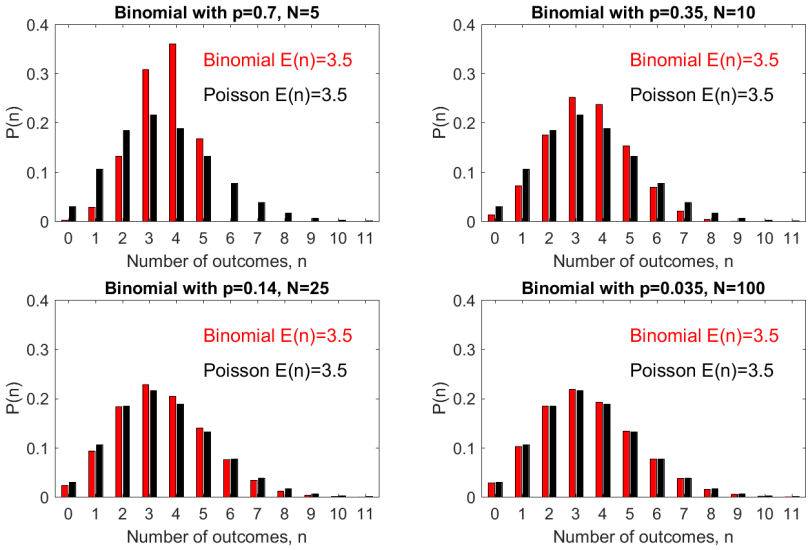
so

$$B(n;p,N) \approx \frac{(pN)^n}{n!}\, e^{-pN} = P(n;\mu = pN)$$

Hence, in this limit, the binomial becomes equal to the Poisson distribution, where the average value is $\mu = pN$, i.e. the same as the average of the binomial distribution. This is shown graphically in the figure below where the distributions become similar as N in the binomial gets larger. See also how you can have different binomial distributions with the same expectation value but different parameters, and different variances and standard deviations.

As stated above, the number of nuclear decays from a source will follow a binomial in principle, as we cannot see more decays than the number of nuclei we have. However, any macroscopic piece of radioactive material has a finite but huge number of nuclei, e.g. of order Avogadro's number. Hence, unless the decay is very rapid, the number of decays observed will be very small compared with the total number, so in this case, the number observed is a very good approximation to a Poisson distribution.

Conversely, in fact, it is very rare to ever get a 'real' Poisson distribution in principle. In the cases mentioned above, they are all in fact fundamentally binomial distributions but are extremely good approximations to Poisson distributions. For the number of photons from a patch of sky, the total possible number would be huge but cannot be infinite as there is only a finite amount of energy in the Universe. Lightning strikes must be limited by the amount of static generated on the clouds. Finally, the number of buses or car crashes must be limited to the total number of vehicles in the UK. However, in all cases, the number observed is minute compared with the maximum possible, so they are all very accurately described by a Poisson distribution.



13 Non-examinable: Appendix

13.1 Expectation value and variance of a binomial distribution

The expectation value of a binomial distribution is

$$\begin{aligned} E(n) \quad = \langle n \rangle \quad &= \sum_{n=0}^{N} n\, B(n;p,N) = \sum_{n=1}^{N} n \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \\ &= Np \sum_{n=1}^{N} \frac{(N-1)!}{(n-1)!(N-n)!} p^{n-1} (1-p)^{N-n} \\ &= Np \sum_{n=1}^{N} \frac{(N-1)!}{(n-1)![(N-1)-(n-1)]!} p^{n-1} (1-p)^{(N-1)-(n-1)} \end{aligned}$$

where in the first line we have made use of the fact that the $n = 0$ term in the sum is 0, hence we can sum from $n = 1$ onwards. Using the substitution $m = n - 1$ and $M = N - 1$ we obtain

$$E(n) = \langle n \rangle = Np \sum_{m=0}^{M} \frac{M!}{m!(M-m)!} p^m (1-p)^{M-m} = Np \sum_{m=0}^{M} B(m;p,M) = Np$$

as the sum above equals 1, being the sum of the terms of a binomial distribution which is normalized to a total probability of 1.

To compute the variance, we start by noticing that

$$V(n) = \langle n^2 \rangle - \langle n \rangle^2 = \langle n(n-1) + n \rangle - \langle n \rangle^2 = \langle n(n-1) \rangle + \langle n \rangle - \langle n \rangle^2$$

Hence, as we know $E(n) = \langle n \rangle$ already, we only need to compute $\langle n(n-1) \rangle$. Using a similar method to previously

$$\begin{aligned} \langle n(n-1) \rangle \quad &= \quad \sum_{n=0}^{N} n(n-1) \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} = \sum_{n=2}^{N} \frac{N!}{(n-2)!(N-n)!} p^n (1-p)^{N-n} \\ &= \quad N(N-1)p^2 \sum_{n=2}^{N} \frac{(N-2)!}{(n-2)!(N-n)!} p^{n-2} (1-p)^{N-n} \end{aligned}$$

$$= \quad N(N-1)p^2 \sum_{n=2}^{N} \frac{(N-2)!}{(n-2)![(N-2)-(n-2)]!} p^{n-2}(1-p)^{(N-2)-(n-2)}$$

Substituting $m = n-2$ and $M = N-2$ we obtain

$$\langle n(n-1)\rangle = N(N-1)p^2 \sum_{m=0}^{M} \frac{M!}{m!(M-m)!} p^m (1-p)^{M-m} = N(N-1)p^2 \sum_{m=0}^{M} B(m;p,M) = N(N-1)p^2$$

Thus

$$V(n) = \langle n(n-1)\rangle + \langle n\rangle - \langle n\rangle^2 = N(N-1)p^2 + Np - (Np)^2 = N^2p^2 - Np^2 + Np - N^2p^2 = Np(1-p)$$

and hence the standard deviation is $\sqrt{Np(1-p)}$.

13.2 Expectation value and variance of a Poisson distribution

Using the well-known properties of an exponential

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \qquad \text{and} \qquad \frac{d}{dx}(e^x) = e^x$$

this means

$$\frac{d}{dx}(e^x) = \sum_{n=0}^{\infty} n \frac{x^{n-1}}{n!} = e^x$$

Multiplying both sides by x gives

$$xe^x = \sum_{n=0}^{\infty} n \frac{x^n}{n!}$$

The expectation value of a Poisson distribution is therefore given by

$$E(n) = \langle n\rangle = \sum_{n=0}^{\infty} n\, P(n;\mu) = \sum_{n=0}^{\infty} n \frac{\mu^n}{n!} e^{-\mu} = e^{-\mu} \left(\sum_{n=0}^{\infty} n \frac{\mu^n}{n!} \right) = e^{-\mu}\,(\mu e^{\mu}) = \mu$$

To compute the variance, we need $V(n) = \langle n^2\rangle - \langle n\rangle^2$. Using a similar trick

$$\frac{d^2}{d^2x}(e^x) = \sum_{n=0}^{\infty} n(n-1) \frac{x^{n-2}}{n!} = e^x$$

and hence

$$x^2e^x = \sum_{n=0}^{\infty} n(n-1) \frac{x^n}{n!} = \sum_{n=0}^{\infty} (n^2 - n) \frac{x^n}{n!} = \sum_{n=0}^{\infty} n^2 \frac{x^n}{n!} - xe^x$$

Therefore

$$\sum_{n=0}^{\infty} n^2 \frac{x^n}{n!} = x^2e^x + xe^x$$

We can now compute

$$\langle n^2\rangle = \sum_{n=0}^{\infty} n^2\, P(n;\mu) = \sum_{n=0}^{\infty} n^2 \frac{\mu^n}{n!} e^{-\mu} = e^{-\mu} \left(\sum_{n=0}^{\infty} n^2 \frac{\mu^n}{n!} \right) = e^{-\mu}(\mu^2 e^{\mu} + \mu e^{\mu}) = \mu^2 + \mu$$

so that

$$V(n) = \mu^2 + \mu - \mu^2 = \mu$$

Hence, the standard deviation is $\sqrt{\mu}$.

Statistics of Measurement
Lecture 3 - Continuous probability densities

Heather Graven, 16 May 2023

14 General properties

In the previous lecture, we saw probability distributions where the outcomes were discrete, labelled as x_i, and had particular probabilities associated with them. However, it is often the case that we do not measure discrete quantities (like the number of nuclear decays in a second) but instead measure continuous quantities (like a voltage). The outcome of such measurements can take any value in a continuum, i.e. the sample space is continuous. We will again call the random variable x, although it is not necessarily a position in space.

The probability of any particular exact value such as $x = 4.3534679265873\ldots$ is clearly negligible. Hence, in the continuous case, the probabilities are given by a **probability density function** (PDF), here denoted by $\rho(x)$. This is defined such that the (infinitesimal) probability of finding the outcome value in the range x to $x + dx$ is $\rho(x)\,dx$ and so a specific value (like the one above) has an infinitesimally small probability. To get a finite probability, then the probability density function has to be integrated over some range. Specifically, the probability of getting x within the range $x_1 \leq x \leq x_2$ is

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} \rho(x)\,dx$$

Note, since the probability is dimensionless, $\rho(x)$ has dimensions $[x^{-1}]$.

The integral of probability density over the whole possible range of x must be 1, akin to the normalisation requirement on the discrete sum of probabilities:

$$\sum_i P_i = 1 \rightarrow \int_{-\infty}^{\infty} \rho(x)\,dx = 1$$

The integral may have narrower limits if the range of x is restricted. An important physics example of this is in Quantum Mechanics, where the modulus squared of the wavefunction gives the probability density for finding the particle at position x, i.e. $\rho(x) = |\psi(x)|^2$ and the normalisation condition can be written as

$$\int_{-\infty}^{\infty} |\psi(x)|^2\,dx = 1$$

which ensures that the particle is bound to be found somewhere. We will discuss several other probability density functions which commonly arise in physics below.

We also need to be able to calculate the expectation value and variance of a continuous probability density function. In the case of a continuous PDF, the probability is given by the continuous $\rho(x)\,dx$ rather than the discrete P_i, so the expression for the expectation value goes to

$$E(x) = \sum_i x_i P_i \rightarrow \int_{-\infty}^{\infty} x\ \rho(x)\,dx$$

Similarly, if we remember the expression for the variance is the expectation value for the squared difference from the mean

$$V(x) = E((x - \langle x \rangle)^2)\ = \langle x^2 \rangle - \langle x \rangle^2$$

then this goes from discrete to continuous as

$$V(x) = \sum_i x_i^2 P_i - \left(\sum_i x_i P_i \right)^2 \rightarrow \int_{-\infty}^{\infty} x^2 \, \rho(x) \, dx - \left(\int_{-\infty}^{\infty} x \, \rho(x) \, dx \right)^2$$

The standard deviation is again given by the square root of the variance.

15 Changing variables

For discrete distributions, then for (at least monotonic) functions $y = y(x)$, the probability of getting $y_i = y(x_i)$ is the same as getting x_i. However, when we deal with probability densities, then we have an integral for the probability and so the density function changes if we change the variable. If the probability density in x is $\rho_x(x)$, then as stated above, the probability of x being between x_1 and x_2 is given by

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} \rho_x(x) \, dx$$

If we want to change variables, then we need to express this as an integral over y. Using $dx = (dx/dy) \, dy$, and writing the limits as $y_1 = y(x_1)$, $y_2 = y(x_2)$, then this is

$$P(y_1 \leq y \leq y_2) = \int_{y_1}^{y_2} \frac{dx}{dy} \, \rho_x(x) \, dy$$

and hence, since changing variables cannot change the probability, we have

$$\rho_y(y) = \left| \frac{dx}{dy} \right| \rho_x(x) = \left| \frac{dy}{dx} \right|^{-1} \rho_x(x)$$

where the right-hand side terms should be expressed in terms of y. Note the modulus of the derivative, as a probability density has to be positive. Effectively, we are changing from dx to dy and hence for a PDF of more than one variable, this generalises to multiplying by the Jacobian.

For example, if $\rho_x(x) = x^3$ and $y = x^2$, then $dy/dx = 2x$ and so

$$\rho_y(y) = \frac{1}{2x} \, x^3 = \frac{x^2}{2} = \frac{y}{2}$$

Note it is *not* correct to state that since $x = y^{1/2}$ and $\rho_x(x) = x^3$ then $\rho_y(y) = (y^{1/2})^3 = y^{3/2}$. Therefore, changing variables is not quite as simple as plugging the expression for new variable into the original probability density.

Finally, one common source of error is non-monotonic functions, where different values of x give the same value of y. These have to handled by dividing up the PDF in x into monotonic pieces, applying the above, and then adding the resulting PDFs to get the total PDF for y.

16 Continuous uniform density function

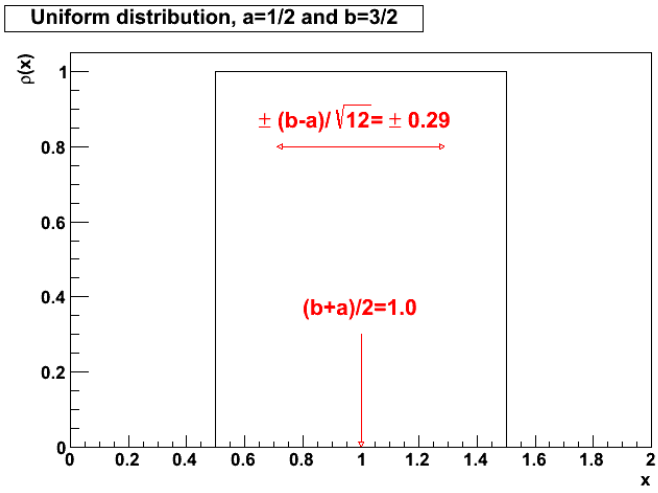
This is the continuous equivalent of the discrete uniform probability distribution which we met in Lecture 1. The continuous uniform probability density function has the same value ρ_0 anywhere within the allowed range of x. Note, the x range cannot be $\pm\infty$ as it would not be possible to normalise the PDF. If the range of x is $a \leq x \leq b$ then we have

$$\begin{aligned} \rho(x; a, b) &= \rho_0 \qquad \text{for } a \leq x \leq b \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

The normalisation requirement means

$$\int_{-\infty}^{\infty} \rho(x)\,dx = \int_a^b \rho(x)\,dx = \int_a^b \rho_0\,dx = \rho_0[x]_a^b = \rho_0(b-a) = 1$$

so $\rho_0 = 1/(b-a)$. The expectation value is intuitively the middle of the range, $(b+a)/2$ (for detail see the appendix) and the variance can be calculated to be $V(x) = (b-a)^2/12$ (see the appendix) so the standard deviation is $(b-a)/\sqrt{12}$. An example is shown below.



This gives a useful way to estimate an error when making a digitised measurement. For example, a digital voltmeter gives a reading of $V = 4.1\,\mathrm{V}$. In the absence of other errors, what this means is the the actual voltage is somewhere in the range $V = 4.05$ to $4.15\,\mathrm{V}$. Hence, if we treat it as a uniform distribution within that range, then the standard deviation would be

$$\text{Std. Dev.} = \frac{4.15 - 4.05}{\sqrt{12}} = \frac{0.1}{3.46} \sim 0.03\,\mathrm{V}$$

which could be used as the estimated measurement error.

17 Exponential density function

The exponential PDF appears in many places in physics. It arises whenever there is a Poisson distribution based on a constant rate, as discussed in the previous lecture. In the case of a constant rate in time, the probability density for the time between each occurrence is given by the exponential distribution.

To see why, say the average rate of occurrences is λ, so the average number in time t is λt. We know the probability of no occurrences between time 0 and t from the Poisson distribution to be $P(0;\lambda t) = e^{-\lambda t}$. The probability of one occurrence between time t and $t+dt$ is $P(1;\lambda dt) = (\lambda dt)e^{-\lambda dt}$. However, for infinitessimal dt, the exponential goes to $e^0 = 1$, so the probability is λdt, as you might expect. Hence, the probability of the next occurrence being between t and $t+dt$ is the probability of no occurrences up to t, times the probability of an occurrence in t to $t+dt$, which is $\lambda dt\, e^{-\lambda t}$. Therefore, the PDF is

$$\rho(t;\lambda) = \lambda e^{-\lambda t}$$

This was for the example of a rate in time; it could also be e.g. the distance between places where crashes happened on the M1, so generically, we will write

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

where λ is the average number per unit of x. The exponential distribution is valid for $0 \leq x \leq \infty$. It is also commonly expressed with a different parametrisation

$$\rho(x;a) = \frac{1}{a}\, e^{-x/a}$$

where a is the average separation in units of x, e.g. the average time between occurrences. Clearly $\lambda = 1/a$ and the average number expected in a period Δx is $\mu = \lambda \Delta x = \Delta x/a$. The factor in front of the exponential ensures the distribution is correctly normalised, so

$$\int_0^\infty \lambda e^{-\lambda x}\, dx = \left[-e^{-\lambda x}\right]_0^\infty = 1 \qquad \text{or} \qquad \int_0^\infty \frac{1}{a}\, e^{-x/a}\, dx = \left[-e^{-x/a}\right]_0^\infty = 1$$

For the exponential distribution, the expectation value is

$$E(x) = \int_0^\infty x\lambda e^{-\lambda x}\, dx$$

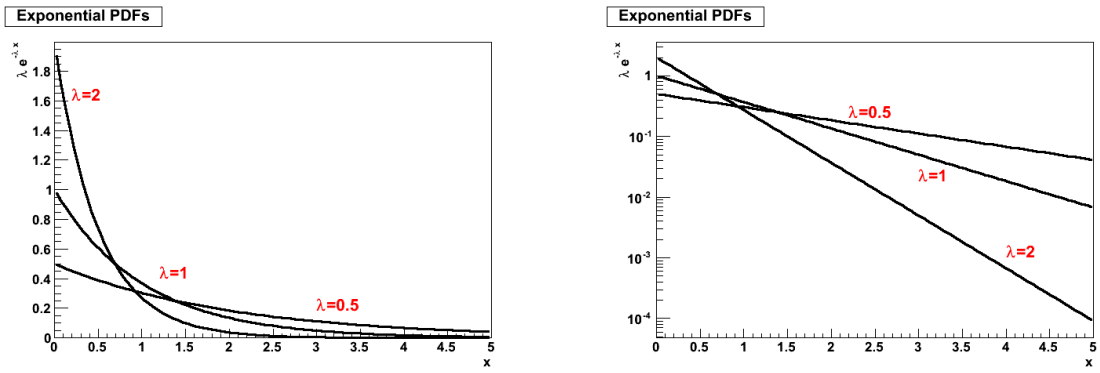
As shown in the appendix, integrating by parts gives

$$E(x) = \frac{1}{\lambda} = a$$

as expected if a is the average separation. The variance is also calculated by integrating by parts to give

$$V(x) = \frac{1}{\lambda^2} = a^2 \qquad \text{and} \qquad \sigma = \frac{1}{\lambda} = a = E(x)$$

Some examples of exponential PDFs with both a linear and logarithmic y axis are shown below. It can be seen how larger values of λ indicate faster rates and shorter intervals between occurrences.



Example: you are waiting at a bus stop. There are 2.5 buses per hour on average but the traffic variation means they turn up at random times. The average rate is $\lambda = 2.5$/hour (i.e. the average time between each is $a = 24$ minutes). What is the probability you would wait for at least 30 minutes? This is given by

$$P(0.5 \leq t \leq \infty) = \int_{0.5}^\infty \lambda e^{-\lambda t}\, dt = \left[-e^{-\lambda t}\right]_{0.5}^\infty = e^{-0.5\lambda} = e^{-1.25} = 0.287$$

18 Gaussian density function

The Gaussian PDF (also called the *normal* or *bell shape* PDF) is perhaps the most important continuous distribution. It is used in many situations involving continuous random variables.

The reason for this, and the justification for the functional shape, will be given in the next lecture when we study the **Central Limit Theorem**. The Gaussian PDF is given by

$$G(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

where the range of x is $-\infty < x < \infty$ and the parameters μ and σ directly control the mean and standard deviation, respectively. For a Gaussian, the probability is proportional to the exponential of the square of the difference from the mean. Therefore, the probability is highest at the mean value and then decreases moving away from the mean in either direction, where this decrease follows the exponential of the square of the difference from the mean. The variance appears in the exponent and governs the steepness of the decrease in probability moving away from the mean, as shown in the example plots below. The factors in front normalise the distribution such that the area under the curve is equal to one, as required for a probability distribution

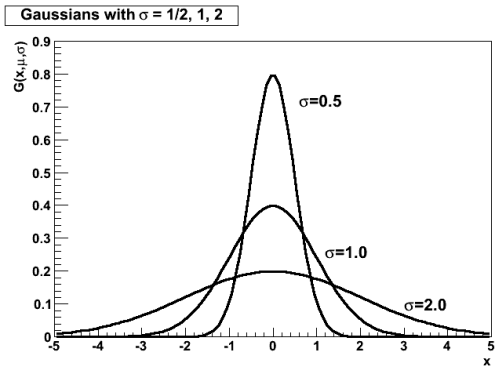
$$\int_{-\infty}^{\infty} G(x;\mu,\sigma)\, dx = 1$$

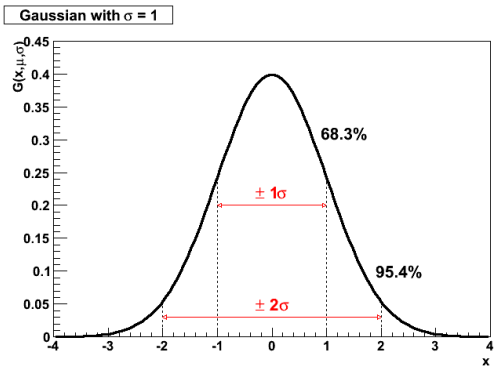
It is common to use the **standard score** $Z = (x-\mu)/\sigma$ such that $dx = \sigma\, dZ$ and hence

$$G(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}$$

Converting to Z makes the expression general, and its value indicates how far x is from the mean, relative to the standard deviation, e.g. for x within one standard deviation of the mean Z will be between -1 and 1. As stated above (and calculated in the appendix)

$$E(x) = \mu, \qquad V(x) = \sigma^2 \qquad \text{and} \qquad \text{Standard deviation} = \sigma$$

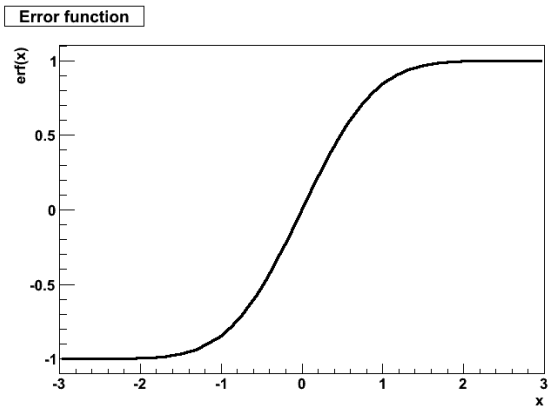




It is often the case that we are trying to measure something, e.g. the length of a rod, and we get slightly different values each time we do it. These most commonly have a Gaussian distribution with a mean at the true value of the length (if our ruler is unbiased), but each individual measurement is spread around the mean with the standard deviation σ. The measurements are then said to have an uncertainty (or error) and the size of the uncertainty reflects the spread of the values. By convention, the uncertainty is specified by the standard deviation σ. Note, however, that this does not absolutely contain all possible values of x; indeed, the tails of the Gaussian go out to infinity so no finite range can contain all values. To find the probability of x being within $\pm 1\sigma$ of μ, we would need to calculate $\int_{\mu-\sigma}^{\mu+\sigma} G(x;\mu,\sigma)\, dx$. Unfortunately, the integral over a finite range of a Gaussian cannot be done analytically. However, the closely related 'error function' (written as 'erf') is defined to be

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2}\, dy$$

and this is normally implemented in most software packages, e.g. in `scipy.special` in python. With a change of variables, it is possible to express Gaussian integrals in terms of $\mathrm{erf}(x)$ and evaluate them numerically. A plot of the erf function is shown below.



As an example of using the error function, again consider calculating the fraction of the Gaussian contained within the $\pm 1\sigma$ range. This is

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{\mu-\sigma}^{\mu+\sigma} e^{-(x-\mu)^2/2\sigma^2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-Z^2/2}\, dZ$$

Changing variables from x to $y = (x-\mu)/\sqrt{2}\sigma = Z/\sqrt{2}$ means $dZ = \sqrt{2}\, dy$ and the limits are $y = \pm 1/\sqrt{2}$, so the integral becomes

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-1/\sqrt{2}}^{1/\sqrt{2}} e^{-y^2} \sqrt{2}\sigma\, dy = \frac{1}{\sqrt{\pi}} \int_{-1/\sqrt{2}}^{1/\sqrt{2}} e^{-y^2}\, dy = \frac{2}{\sqrt{\pi}} \int_{0}^{1/\sqrt{2}} e^{-y^2}\, dy$$

where the last equality follows as the function being integrated is even. Hence this gives

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{\mu-\sigma}^{\mu+\sigma} e^{-(x-\mu)^2/2\sigma^2}\, dx = \mathrm{erf}(1/\sqrt{2}) = \mathrm{erf}(0.707) = 0.683$$

This means that the probability of x being within this range of $\pm 1\sigma$ is 68.3%. This is true of all Gaussian distributions and one of the few numbers you are expected to remember from this course. In addition to numerical calculations using the error function, tables of numerical values relating to the Gaussian distribution are widely available.

Now that we know the probability of x being within the range of $\pm 1\sigma$ is 68.3%, we can reflect on the convention of using the standard deviation for uncertainty. When we quote the uncertainty as $\pm 1\sigma$ we are quoting the range for which the true value of x is within with probability 68.3%, so a pretty good chance but still a substantial chance it could be outside this range. This convention is somewhat arbitrary and if we wanted to be more conservative, then we could specify the uncertainty as $\pm 2\sigma$, which would give a probability of 95.4%, or $\pm 3\sigma$ for 99.7%. Alternatively, we could set the probability we want and then calculate the range as a multiple of the standard deviation. Some common choices are given in the table below. These are all examples of 'confidence intervals' which we will discuss further in lecture 8; for example, we would say we are 68.3% confident that x will be in the $\pm 1\sigma$ interval. Note that quoting $\pm 1\sigma$ as the uncertainty is by far the most common, but it is good practice to always say what the uncertainty refers to. This is particularly important when using something other than $\pm 1\sigma$.

$\pm\sigma$	Fraction	Fraction	$\pm\sigma$
1	0.683	0.90	1.64
2	0.954	0.95	1.96
3	0.997	0.99	2.58

19 Cumulative distribution function

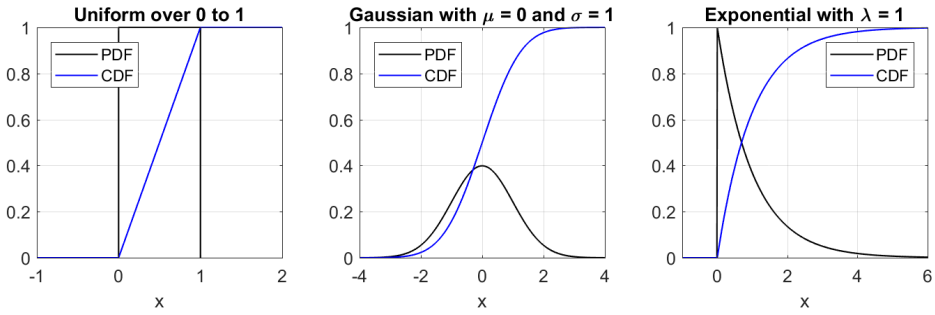
The cumulative distribution function (CDF) is another useful way of representing a probability distribution. The cumulative distribution function $F(x)$ is the probability that the random variable X takes on a value that is less than or equal to x

$$F(x) = P(X \leq x)$$

If X is a continuous random variable then the CDF is given by the integral of the PDF $\rho(x)$ from $-\infty$ (or the lower limit of the distribution) to x

$$F(x) = \int_{-\infty}^{x} \rho(t)\, dt$$

It can be seen that $F(x)$ is an anti-derivative of $\rho(x)$. As x tends to $-\infty$ (or the lower limit) then $F(x)$ tends to zero and as x tends to ∞ (or the upper limit) then $F(x)$ tends to one, i.e. the entire pdf is integrated. The plots below give examples of CDFs for the uniform, exponential and Gaussian distributions. Note that the CDF for the Gaussian (given by $\frac{1}{2}(1+erf(\frac{x-\mu}{\sigma\sqrt{2}}))$) resembles the error function above, as expected. CDFs can also be created for discrete distributions by summing the probabilities for all possible values less than x.



20 Non-examinable: Appendix

20.1 Convolution

With discrete distributions the outcome is unambiguous; one of the discrete x_i is observed. However, with continuous distributions, then the variables can have a measurement uncertainty in addition to the 'underlying' distribution.

For example, the exponential distribution is used to measure the rate of nuclear decays. This involves measuring the time t until the next decay. If the measurement of t itself has a negligible uncertainty, then we will see a pure exponential. However, if the measured value of t has a Gaussian uncertainty with a significant width compared with the lifetime, then the measurements will have a significant smearing and the exponential will be distorted. Indeed, for a very short lifetime, the uncertainty may mean the apparent measured time is negative!

Consider a small part of the exponential, between t' and $t' + dt'$. If there was no uncertainty on the time measurement, then we would observe t' directly. However, t' is smeared by the

Gaussian to give us a measured time t. The probability of the pure exponential being between t' and $t' + dt'$ is $\lambda e^{-\lambda t'}\, dt'$. This will be smeared by a Gaussian, which will of course be centred on t'. Hence, the probability density for observed time t around the exponential region at t' is

$$G(t; t', \sigma)\lambda e^{-\lambda t'}\, dt'$$

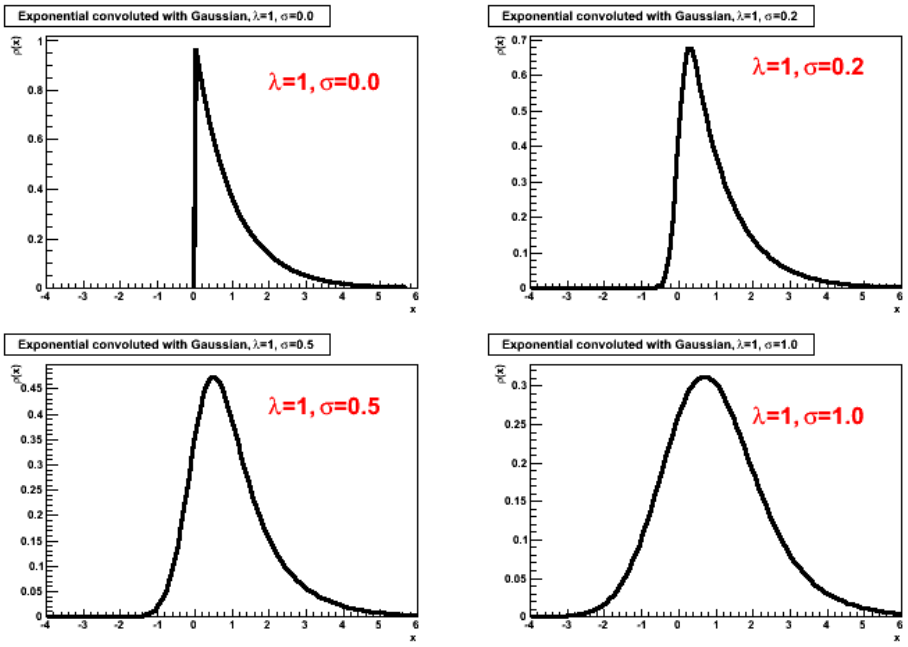
We only see t, so we need to integrate over all possible t' to get the PDF in terms of t only, i.e.

$$\rho(t; \lambda, \sigma) = \int_0^\infty G(t; t', \sigma)\lambda e^{-\lambda t'}\, dt'$$

This is messy but doable and gives

$$\rho(t; \lambda, \sigma) = \frac{\lambda e^{\lambda^2\sigma^2/2}}{2} e^{-\lambda t}\left[1 - \mathrm{erf}\left(\frac{\lambda\sigma^2 - t}{\sigma\sqrt{2}}\right)\right]$$

This process is called 'convolution'. Convolutions of an exponential distribution with a Gaussian distribution are shown below for various values of σ.



We can also do a convolution of a Gaussian with another Gaussian, as would be seen if there are two sources of uncertainty in a measurement, e.g. when measuring a circuit voltage and this depending on a fluctuating temperature as well as on the uncertainty from the voltmeter. In general the two Gaussians will have different widths σ_1 and σ_2. This convolution is also messy but the answer is very simple; it is just another Gaussian. The mean is unchanged (as you would expect, given that you are convoluting two symmetric distributions) and the overall width is

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$$

Clearly, the total width is larger than either contribution. This should look familiar from the propagation of errors formula for $z = x + y$; it is effectively the same thing because for every value of x (which has a Gaussian distribution), the sum (i.e. z) is also smeared by the Gaussian in y.

20.2 Expectation value and variance of a uniform distribution

The expectation value of a continuous uniform distribution is

$$E(x) = \int_a^b \frac{x}{b-a}\,dx = \frac{1}{2(b-a)}\left[x^2\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}$$

i.e. the centre of the distribution, as would be expected.

For the variance, we need the average of x^2. This is

$$\langle x^2\rangle = \int_a^b \frac{x^2}{b-a}\,dx = \frac{1}{3(b-a)}\left[x^3\right]_a^b = \frac{b^3-a^3}{3(b-a)}$$

Hence the variance is

$$\begin{aligned} V(x) &= \langle x^2\rangle - \langle x\rangle^2 = \frac{b^3-a^3}{3(b-a)} - \frac{(b+a)^2}{4} = \frac{4(b^3-a^3)-3(b+a)^2(b-a)}{12(b-a)} \\ &= \frac{4b^3-4a^3-3b^3-6b^2a-3ba^2+3b^2a+6ba^2+3a^3}{12(b-a)} \\ &= \frac{b^3-3b^2a+3ba^2-a^3}{12(b-a)} = \frac{(b-a)^3}{12(b-a)} = \frac{(b-a)^2}{12} \end{aligned}$$

and hence the standard deviation is $(b-a)/\sqrt{12}$.

20.3 Expectation value and variance of an exponential distribution

Integrating by parts, the expectation value is

$$E(x) = \int_0^\infty \lambda x e^{-\lambda x}\,dx = \int_0^\infty e^{-\lambda x}\,dx - \left[xe^{-\lambda x}\right]_0^\infty = -\frac{1}{\lambda}\left[e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda}$$

For the variance, we need the average of x^2. Again by parts, this is

$$\langle x^2\rangle = \int_0^\infty \lambda x^2 e^{-\lambda x}\,dx = 2\int_0^\infty xe^{-\lambda x}\,dx - \left[x^2e^{-\lambda x}\right]_0^\infty = \frac{2}{\lambda^2}$$

and so

$$V(x) = \langle x^2\rangle - \langle x\rangle^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

and hence the standard deviation is $1/\lambda$.

20.4 Expectation value and variance of a Gaussian distribution

Consider the expectation value of the standard score Z

$$E(Z) = \int_{-\infty}^\infty Z\frac{1}{\sqrt{2\pi}}e^{-Z^2/2}\,dZ = \frac{1}{\sqrt{2\pi}}\left[-e^{-Z^2/2}\right]_{-\infty}^\infty = 0$$

The expectation value of x is

$$\begin{aligned} E(x) &= \int_{-\infty}^\infty x\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}\,dx = \int_{-\infty}^\infty (\sigma Z+\mu)\frac{1}{\sigma\sqrt{2\pi}}e^{-Z^2/2}\,\sigma\,dZ \\ &= \sigma\int_{-\infty}^\infty Z\frac{1}{\sqrt{2\pi}}e^{-Z^2/2}\,dZ + \mu\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-Z^2/2}\,dZ \\ &= \mu \end{aligned}$$

since the first integral is $E(Z) = 0$ and the second is the PDF in terms of Z and so integrates to one.

Since $E(Z) = \langle Z \rangle = 0$, the variance of Z is simply

$$V(Z) = \int_{-\infty}^{\infty} Z^2 \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}\, dZ - E(Z)^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} Z^2 e^{-Z^2/2}\, dZ = \langle Z^2 \rangle$$

Using integration by parts, this is

$$V(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-Z^2/2}\, dZ - \frac{1}{\sqrt{2\pi}} \left[Z e^{-Z^2/2} \right]_{-\infty}^{\infty} = 1$$

again using the normalisation of the PDF. The variance of x is then

$$\begin{aligned} V(x) &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\, dx - E(x)^2 = \int_{-\infty}^{\infty} (\sigma Z + \mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-Z^2/2}\, \sigma\, dZ - \mu^2 \\ &= \sigma^2 \int_{-\infty}^{\infty} Z^2 \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}\, dZ + 2\sigma\mu \int_{-\infty}^{\infty} Z \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}\, dZ + \mu^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}\, dZ - \mu^2 \\ &= \sigma^2 \end{aligned}$$

since the first integral is $V(Z) = 1$, the second is $E(Z) = 0$ and the third is the normalisation $= 1$.

Statistics of Measurement
Lecture 4 - The Central Limit Theorem

Heather Graven, 18 May 2023

21 Introduction

We have seen the expectation values and variances of the various discrete and continuous distributions we have met so far; these are summarised in the table below, using the notation used in the previous lectures.

		Expectation value	Variance	Standard deviation
Discrete	Binomial	pN	$Np(1-p)$	$\sqrt{Np(1-p)}$
	Poisson	μ	μ	$\sqrt{\mu}$
Continuous	Uniform	$(b+a)/2$	$(b-a)^2/12$	$(b-a)/\sqrt{12}$
	Exponential	$1/\lambda = a$	$1/\lambda^2 = a^2$	$1/\lambda = a$
	Gaussian	μ	σ^2	σ

In practice, it is typical that a variable will have a probability distribution that does not follow an ideal theoretical distribution because there may be multiple influences on that variable that have different distributions or multiple sources of error in the measurement. In addition, it is often the case that an experiment will take several measurements and compute the average. The average is clearly found from the sum of the measurements. The expected probability distribution of summed quantities can look very different from the underlying distribution from which the individual measurements are drawn. In this lecture, we will look at probability distributions for the sums of random variables.

22 Sums of random variables

Consider that we are going to sample N independent random variables x_i where the x_i are each from a separate PDF $\rho_i(x)$ which has an expectation value E_i and a variance V_i. Note, these individual distributions do not have to be Gaussian or even the same type of distribution. We will calculate a sum of them

$$X = \sum_i x_i$$

X is an example of a variable which is a function of only random variables, the x_i. Clearly, if we sampled another set of x_i, then since the x_i are random, we would have a different value of X. Hence, X, is itself a random variable.

We would like to know something about the probability distribution of X. For simplicity, we will initially look at summing just two random variables x_1 and x_2 with a PDF of $\rho(x_1, x_2)$. Consider the expectation values of the sum

$$\begin{aligned} E(X) &= \langle x_1 + x_2 \rangle = \int\int (x_1 + x_2)\rho(x_1, x_2)\, dx_1 dx_2 \\ &= \int\int x_1 \rho(x_1, x_2)\, dx_1 dx_2 + \int\int x_2 \rho(x_1, x_2)\, dx_1 dx_2 \\ &= \langle x_1 \rangle + \langle x_2 \rangle = E_1 + E_2 \end{aligned}$$

i.e. the expectation value of the sum is simply the sum of the individual expectation values. Similarly, for the variance

$$\begin{aligned} V(X) &= \langle (X - \langle X \rangle)^2 \rangle = \left\langle [(x_1 + x_2) - (E_1 + E_2)]^2 \right\rangle = \left\langle [(x_1 - E_1) + (x_2 - E_2)]^2 \right\rangle \\ &= \left\langle (x_1 - E_1)^2 \right\rangle + \left\langle (x_2 - E_2)^2 \right\rangle + 2 \left\langle (x_1 - E_1)(x_2 - E_2) \right\rangle \end{aligned}$$

The variance of the sum is the sum of the variances (first two terms), plus the covariance (third term). If we have two independent (i.e. uncorrelated) variables, then the joint probability density is the product of the two separate probability densities

$$\rho(x_1, x_2)\, dx_1 dx_2 = \rho_1(x_1)\rho_2(x_2)\, dx_1 dx_2$$

In this case, the covariance term becomes

$$\begin{aligned} \langle (x_1 - E_1)(x_2 - E_2) \rangle &= \int\int (x_1 - E_1)(x_2 - E_2)\rho_1(x_1)\rho_2(x_2)\, dx_1 dx_2 \\ &= \int (x_1 - E_1)\rho_1(x_1)\, dx_1 \int (x_2 - E_2)\rho_2(x_2)\, dx_2 \end{aligned}$$

Each of these integrals is

$$\int (x_i - E_i)\rho_i(x_i)\, dx_i = \int x_i \rho_i(x_i)\, dx_i - E_i \int \rho_i(x_i)\, dx_i = E_i - E_i = 0$$

since E_i is the average of x_i. Conceptually, you can think of the covariance term being zero for independent random variables because a positive $(x_{1,i} - E_1)$ has an equal chance of being multiplied by a positive or negative $(x_{2,j} - E_2)$ and the average will be zero. Hence, following the derivation above and the conceptual reasoning, for independent variables

$$V(X) = \left\langle (x_1 - E_1)^2 \right\rangle + \left\langle (x_2 - E_2)^2 \right\rangle = V_1 + V_2$$

i.e. the variance of the sum is simply the sum of the variances. Note, the covariance term is only zero for two independent variables.

These results generalise to summing more than two independent random variables, such that

$$E(X) = \sum_i E_i, \qquad V(X) = \sum_i V_i$$

Hence, in general, **both the expectation value and the variance of the sum is just the sum of the individual expectation values and variances** for independent random variables.

It turns out there are other quantities besides the expectation value and variance which can be used to characterise a distribution. The expectation value is linear in the x_i, while the variance is quadratic. There are further similar quantities for higher orders; the third order (or cubic) measure is called **skewness** S, which relates to the asymmetry of the distribution, while the fourth order (or quartic) is **(extended) kurtosis** K, which compares the density in the peak of the distribution vs the tails of the distribution relative to the Gaussian distribution, and so on. The exact definitions are given in the appendix. These higher order quantities are also all the sums of the corresponding quantities for the individual distributions for independent random variables, i.e. $S(X) = \sum_i S_i$, $K(X) = \sum_i K_i$, etc.

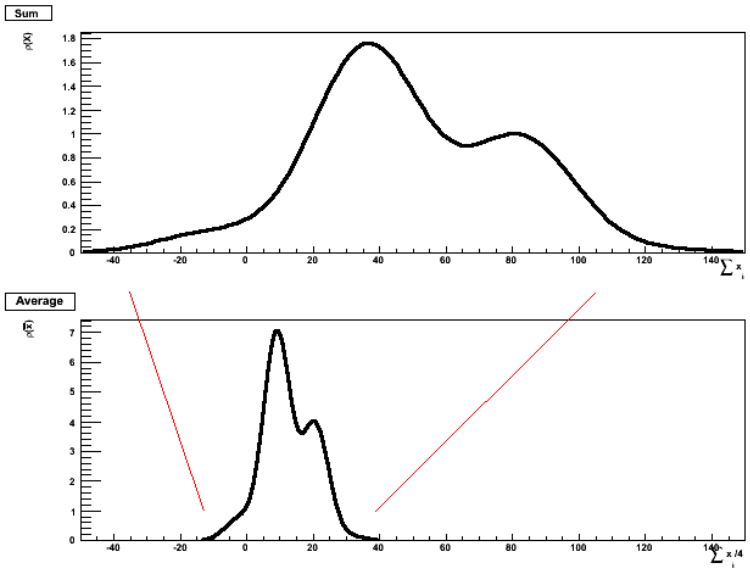
23 The Central Limit Theorem

The Central Limit Theorem (CLT) is a very general theorem which applies to any such cases. It basically says that a sum (and hence also the average) of N independent random variables will tend to have a Gaussian PDF as N gets large, irrespective of the original distributions of the measurements. This may seem bizarre, but you will see it is true.

We are interested in the limit of adding a very large number N of independent random variables x_i. For the sum X, all the quantities mentioned (expectation value, variance, skewness, etc) will generally keep getting bigger and bigger as N increases, as there are more and more contributions to the sums. Hence, these quantities all go to infinity as N gets large, which can obscure what the result is. It is mathematically less tricky to see what is happening if we consider the average of the x_i rather than the sum. The average is

$$\overline{x} = \frac{1}{N}\sum_i x_i = \frac{X}{N}$$

It is therefore just the sum divided by N. This means whatever the probability density distribution is for X, then $\overline{x}$ has the same PDF, but with the x axis scaled down by N, as shown in the figure.



This means the expectation value of $\overline{x}$ is the expectation value of X divided by N, as shown explicitly by changing variables

$$E(\overline{x}) = \int \overline{x}\rho(\overline{x})\, d\overline{x} = \int \frac{X}{N}\rho(X)\, dX = \frac{E(X)}{N}$$

It also means the standard deviation of $\overline{x}$ is that of X divided by N. However, this results in the variance of $\overline{x}$ being that of X divided by N^2. In fact, this should be clear as the variance is quadratic in X. Similarly, the skewness is divided by N^3, etc. Hence, when adding more and more values of x_i, then the N dependences of the various quantities go as

$$E(\overline{x}) = E\left(\frac{X}{N}\right) = \frac{E(X)}{N} = \frac{\sum_i E_i}{N} \qquad \text{goes as} \sim \frac{N}{N} \sim 1$$

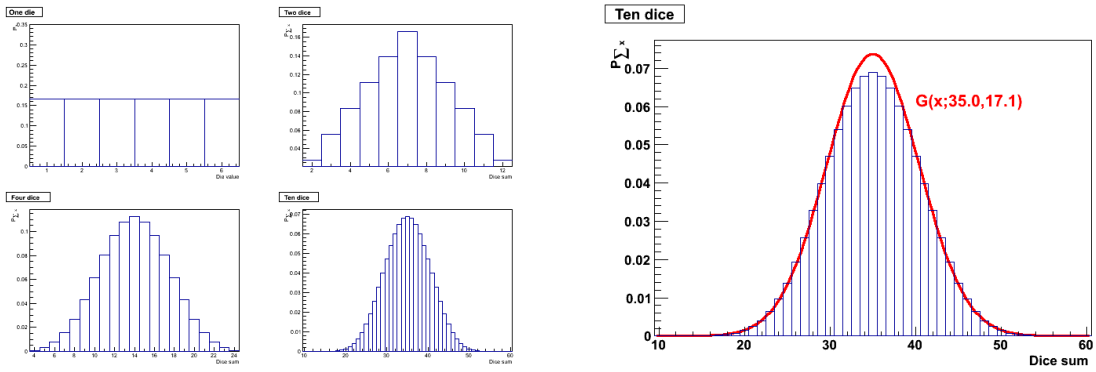
$$V(\overline{x}) = V\left(\frac{X}{N}\right) = \frac{V(X)}{N^2} = \frac{\sum_i V_i}{N^2} \qquad \text{goes as} \sim \frac{N}{N^2} \sim \frac{1}{N}$$

$$S(\overline{x}) = S\left(\frac{X}{N}\right) = \frac{S(X)}{N^3} = \frac{\sum_i S_i}{N^3} \qquad \text{goes as} \sim \frac{N}{N^3} \sim \frac{1}{N^2}$$
$$K(\overline{x}) = K\left(\frac{X}{N}\right) = \frac{K(X)}{N^4} = \frac{\sum_i K_i}{N^4} \qquad \text{goes as} \sim \frac{N}{N^4} \sim \frac{1}{N^3}$$

and so on. Therefore, in the large N limit, the leading term is the expectation value, and all other terms go to zero. This means as $N \to \infty$, then the only value for the $\overline{x}$ we can get will be the expectation value, as the width (and higher terms) are all zero. This is expected; we should get a measurement average equal to the expectation value in the limit of an infinite number of measurements. This means that in this limit, all measurement averages go to a single universal PDF shape, which is basically an infinitely thin spike (called a 'Dirac delta function') at the distribution average.

If we step back from the extreme limit for a moment, the next order term is the variance (since all other terms have a higher $1/N$ dependence). Hence, to this level of approximation, the distribution is that of a mean and a (small) variance, with all higher order terms zero. Again in this limit, all measurement averages have a universal PDF shape with only the expectation value and variance non-zero. The question is then; what distribution has the property that the skewness, kurtosis, etc, are all zero, as this will be the resulting distribution in all cases? The function which has this property turns out to be the Gaussian; this can be considered to be the defining property of the Gaussian distribution. Hence, no matter what the original distributions were, in the large N limit where the higher order terms above the variance can be ignored, then everything will look Gaussian.

As stated above, the sum X distribution and average $\overline{x}$ distribution are simply related by an N scaling. Hence, if the average is Gaussian, then so is the sum (albeit with a mean and width which go very large in the large N limit). As an example, the plot shows the sum of multiple dice throws (one, two, four and ten), and a comparison of the sum of ten dice throws with a Gaussian of the appropriate mean and variance.



Finally, note that if we are adding Gaussian-distributed random variables, then all the S_i, K_i, etc, higher order quantities are zero by definition. This means $S(X)$, $K(X)$, etc, are also all exactly zero, even for small N. Hence, the resulting distribution for X is by definition *exactly* Gaussian and the overall width will satisfy

$$\sigma^2 = \sum_i \sigma_i^2$$

This is another occurrence of **adding in quadrature**.

24 Sums of exponentially distributed random variables

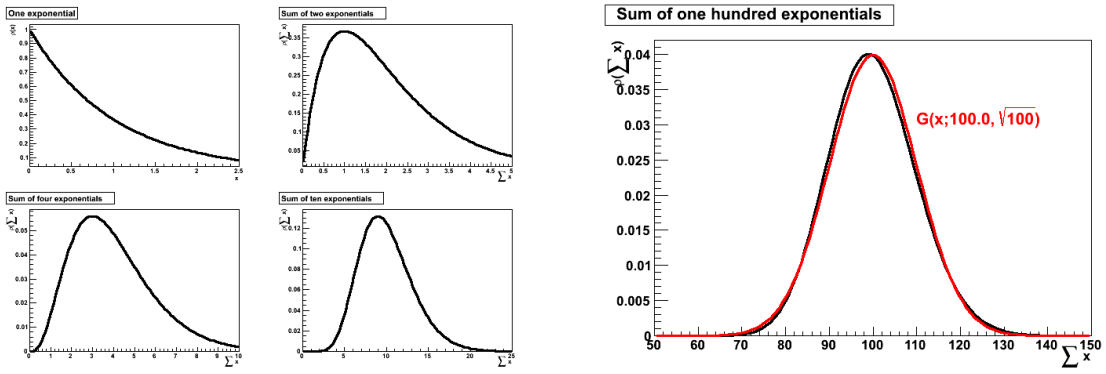
As another example of the Central Limit Theorem, consider summing two random variables drawn from exponential distributions. The sum can be anything from 0 to ∞ but there are many different combinations of the two values which will give the same sum. Since we are only interested in the sum, we have to integrate over all these combinations. The chance of getting a low value of the sum is relatively small as it requires both values to be small. Very large values in the sum are also unlikely, particularly as the exponential distribution falls off for large values. The resulting PDF for the sum of two random variables with the same exponential PDF is (see the appendix)

$$\rho(X)\,dX = \lambda^2 X e^{-\lambda X}\,dX$$

This generalises for the sum of N exponential random variables to give the PDF

$$\rho(X)\,dX = \lambda^N X^{N-1} e^{-\lambda X}\,dX$$

As N increases, this gives a PDF for the sum which looks more and more Gaussian; this is shown below. Again, the red curve gives the Gaussian with the expected mean and width.



One issue is that a Gaussian is not limited in range, and has a tail which goes to $\pm\infty$. However, the exponential (and hence sum of exponentials) has a range 0 to ∞ and so it can never truely duplicate the Gaussian distribution for negative x. It is also the case for both of the discrete distributions we have studied, the binomial and the Poisson, that n is limited to $0 \leq n$ (and in the binomial case $n \leq N$). However, in all cases, with a standard deviation much smaller than the mean, then the probabilities near zero are very small and this effect is not important. For the binomial, we would also need the mean to be far from the upper limit N.

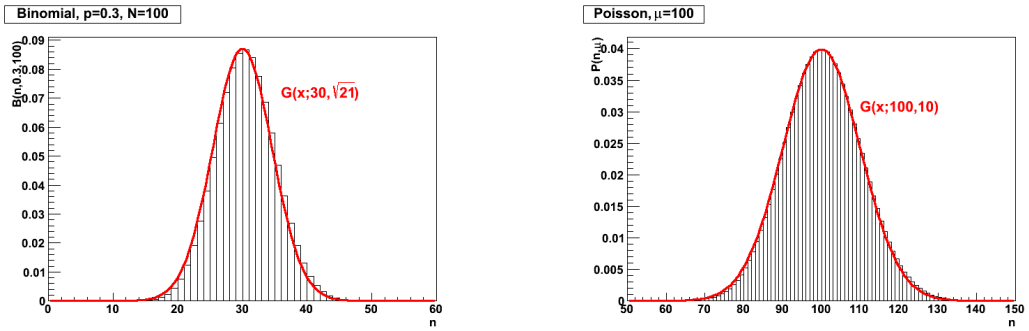
25 CLT for discrete distributions

The CLT applies to discrete distributions but here a bit of care is needed. The sum of the numbers resulting from a binomial or Poisson distribution must be an integer. Indeed, the same is true of the dice case discussed above. Hence, this sum can never be truly a Gaussian distribution as it is still discrete. However, with a large enough sum, then the standard deviation can be much bigger than 1, so the integerisation becomes effectively negligible and the distribution looks almost continuous.

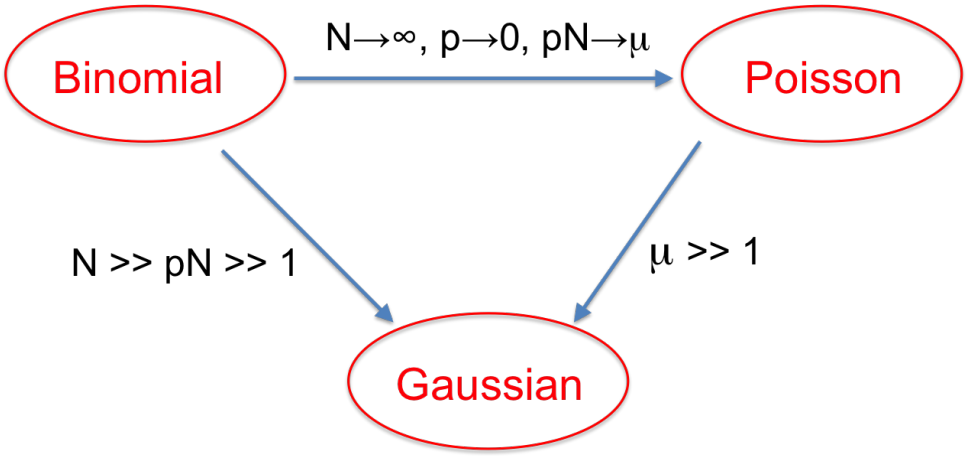
Both the binomial and Poisson are in fact very simple to sum. Specifically, the sum of the outcomes of a binomial is another binomial, and the sum of Poissons is another Poisson. This can be shown mathematically, but in fact, it is straightforward to understand. For the binomial, doing N trials and then another N trials is equivalent to $2N$ trials. This is still binomial,

i.e. the probability distribution of the sum will be given by $B(n; p, 2N)$. For the Poisson, the average number expected is μ and so if the number observed is measured N times, then the total average number expected over these N times is $N\mu$. Summing the n_i observed will be equivalent to having a Poisson with an average $N\mu$ sampled once. Hence, the sum is also a Poisson, explicitly its probability distribution is given by $P(n; N\mu)$.

Therefore, to see how sums of these distributions tend to a Gaussian just requires us to take the number of trials N to be large for the binomial (while ensuring that $1 \ll pN \ll N$ to avoid the problems with discreteness and range), or the average μ to be large for the Poisson. These are shown below for $p = 0.3$ and $N = 100$ for the binomial, and $\mu = 100$ for the Poisson, with the corresponding Gaussian approximations superimposed. It is seen that the Gaussian approximations are very reasonable. The width of the corresponding Gaussian for the Poisson case is seen to be $\sim \sqrt{n}$.



Note, this is a different limit from that discussed in lecture 2 where the Poisson and binomial become the same. The limit there was for a large number of trials, but not necessarily a large mean. The CLT limit of them both becoming Gaussian-like is only when the mean is large (and for the binomial, when the mean is also small compared with N).



26 Non-examinable: Appendix

26.1 Skewness and kurtosis

The explicit forms for expectation value, variance, skewness and kurtosis of a discrete distribution are

$$\begin{aligned} E(x) &= \frac{1}{N}\sum_{i=1}^{N} x_i P_i \\ V(x) &= \frac{1}{N}\sum_{i=1}^{N} \left[x_i - E(x)\right]^2 P_i \\ S(x) &= \frac{1}{NV(x)^{3/2}}\sum_{i=1}^{N} \left[x_i - E(x)\right]^3 P_i \\ K(x) &= -3 + \frac{1}{NV(x)^2}\sum_{i=1}^{N} \left[x_i - E(x)\right]^4 P_i \end{aligned}$$

with equivalent definitions for continuous distributions. The -3 in the definition of kurtosis is purely to give zero for a Gaussian distribution.

26.2 PDF for the sum of two exponential random variables

The PDF for two exponential random variables is

$$\rho(x_1, x_2)\, dx_1 dx_2 = \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2}\, dx_1 dx_2 = \lambda^2 e^{-\lambda(x_1+x_2)}\, dx_1 dx_2 = \lambda^2 e^{-\lambda X}\, dx_1 dx_2$$

To get the PDF in terms of X, we need to change variables from x_1 and x_2 to $X = x_1 + x_2$ and another variable; a simple choice is the difference $D = x_2 - x_1$. Inverting these gives $x_1 = (X - D)/2$ and $x_2 = (X + D)/2$. The Jacobian for this change is

$$\begin{vmatrix} \partial x_1/\partial X & \partial x_1/\partial D \\ \partial x_2/\partial X & \partial x_2/\partial D \end{vmatrix} = \begin{vmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{vmatrix} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

so that

$$dx_1 dx_2 = \frac{1}{2}\, dX dD$$

Hence, the PDF in terms of X and D is

$$\rho(X, D)\, dX dD = \frac{1}{2}\lambda^2 e^{-\lambda X}\, dX dD$$

We only want the PDF in terms of X (as we don't care about D) so we need to integrate over D for fixed X. The minimum possible value of D is when $x_2 = 0$, for which $x_1 = X$ and hence $D = -X$. Similarly, the maximum possible value of D is when $x_1 = 0$, for which $x_2 = X$ and hence $D = X$. Therefore

$$\rho(X) = \int_{-X}^{X} \frac{1}{2}\lambda^2 e^{-\lambda X}\, dD = \frac{1}{2}\lambda^2 e^{-\lambda X} \left[\, D \,\right]_{-X}^{X} = \frac{1}{2}\lambda^2 e^{-\lambda X}\, 2X = \lambda^2 X e^{-\lambda X}$$

Statistics of Measurement (Lectures 5-9)
Lecture 5 - Hypothesis testing

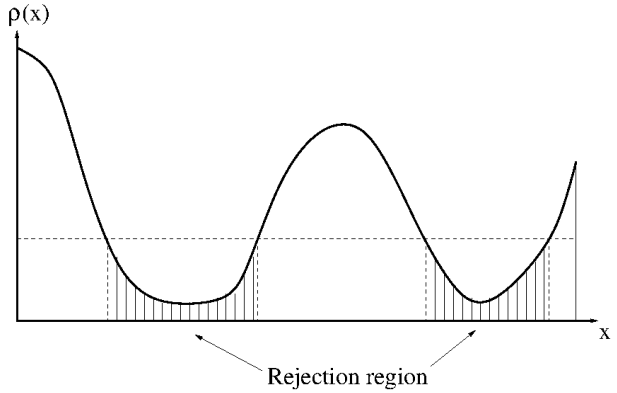
Heather Graven, 19 May 2023

1 Introduction

Now we turn to the estimation part of the course. The term 'estimation' means to deduce the values of parameters using data taken in a experiment. However, before we look at methods of estimation of the values of parameters, we will first look at a simpler problem. We will ask whether some data we have measured are consistent with a particular probability distribution. The probability distribution may be parameterless and hence fixed. However, it is more usual that there is a parameter with a particular value which we are interested in and we want to see if the data support it having this value. Hence, rather than trying to estimate the value of a parameter, we are seeing if this value is experimentally allowed. For example, we want to test if the parameter θ has the value θ_0. This is our hypothesis and this is often labelled as H_0. This is called 'hypothesis testing' as we are testing whether the hypothesis that the parameter has a particular value is consistent with data.

2 Hypothesis testing

The first problem we hit is what do we mean by 'consistent'? Take the simplest case of having a hypothesised PDF $\rho_0(x)$ (which may depend on some parameter for which we have assumed a specific value) and making a single measurement, x_m. Basically, if the hypothesis is correct, we would expect the value we measure usually to be in the regions where the PDF is large. Hence, we would reject the hypothesis if the measured value is in regions where it is small. But how small is small? The absolute value of the PDF is not a useful measure; e.g. the value of the PDF in a uniform distribution is the inverse of the range so it can be arbitrarily small. We need an absolute probability so we must integrate the PDF over the regions where it is small. Quantitatively, a common way to do this is to choose a horizontal line, as shown in the figure, and integrate over ranges of the PDF which are smaller than the line. There can be one or several such ranges for a given horizontal line and together, all the values of x which lie within those ranges are said to be in the 'rejection region' (sometimes called the 'critical region').



The actual height of the horizontal line is not a particularly meaningful value, but it is usually chosen so that there is something like $\alpha = 10\%$, 5% or even 1% of the total probability

in the rejection region, where

$$\alpha = \int_{\text{Rejection region}} \rho_0(x)\, dx$$

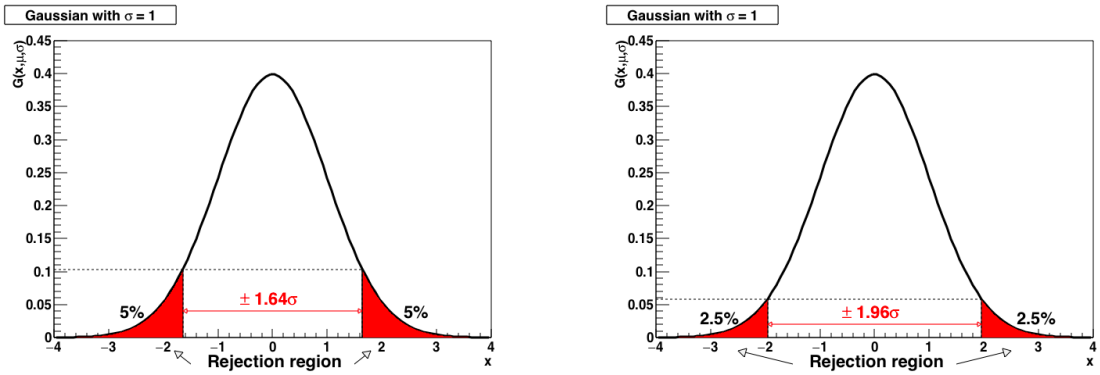
The fraction α is called the 'significance level' (for reasons we will discuss later) and is entirely up to us; there is no 'right' value. (As stated previously, estimation can be somewhat arbitrary.) This means whenever we quote the result of a hypothesis test, it is essential we give the significance level used.

Therefore, if we chose $\alpha = 0.05$ and measured x_m in the rejection region, we would say the hypothesis is false at the 5% significance level. Of course, we could have just been unlucky; by definition, 5% of the time we would expect to get x_m in the rejection region even if the hypothesis is true. Such an outcome, i.e. rejecting a true hypothesis, is called a 'Type I error' and is unavoidable for any finite significance level. The easy way to remember that this is one meaning of the significance level is that significance level equals 'screw-up level'; we will screw-up 5% of the time if the hypothesis is true. Even at a significance level of 1%, then with 100 physicists testing the hypothesis of the theory of relativity, it is likely one would reject the theory. The other 99 might not bother publishing as their result is boring and expected, but you can be sure the one would (try to) publish their evidence that Einstein was wrong.

Therefore, we might want to reduce the significance level to avoid this. However, that makes the rejection region smaller and so it is less likely we will be able to reject the hypothesis if it is false. Hence, we have to choose values which set a balance between these effects.

Note, we can never prove a hypothesis is true. Even if x_m was right at the peak of the PDF, it does not mean that the PDF we have chosen is correct. Hence, all we can do is reject what we consider to be false hypotheses. If we get an x value which will not reject H_0, even though it is actually false, it is called 'Type II error'. Unless we actually know the true PDF (given that the hypothesis under test is false in this case), we cannot evaluate the probability of a Type II error.

Because of the CLT, the Gaussian distribution is the most common, and the integrals over various ranges were given in Lecture 3. Some plots are shown below for a total rejection region of 10% and 5% (which have 5% and 2.5% in the tail on each side). Hence, if you had a theory which predicted a Gaussian distribution centred on zero for some quantity, you would reject the theory at 5% significance level if the value you measured was more than 1.96σ away from zero.



In fact, you will often hear the significance level quoted in terms of number of sigma rather than percentages, as a Gaussian PDF is so common. Indeed, this explains why it is called 'significance level'. Clearly, if we pick a 5% significance level, which is outside 1.96σ, and then find x_m is in the rejection region, this is a more meaningful result than if we had picked a 10% significance level at 1.64σ and found x_m in the rejection region in that case. Hence, a rejection with significance level 1.96σ is more important, i.e. statistically more significant, than one at

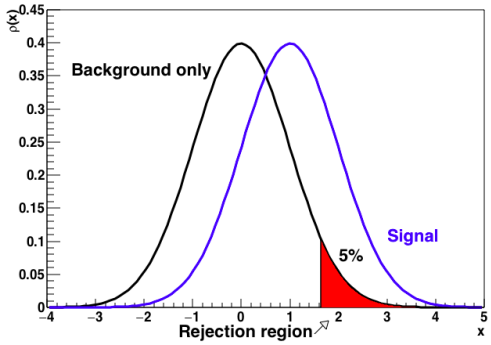
1.64σ. Note, this is even done for non-Gaussian PDFs, where a significance level is converted into an equivalent of sigmas using the Gaussian integrals, even though the actual PDF has no parameter called sigma.

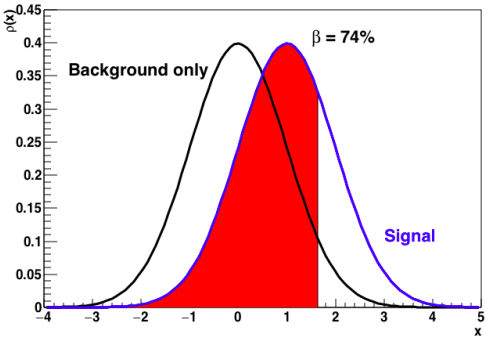
3 The alternative hypothesis

So far, so straightforward. However, even in the above, which seems intuitive, there are some hidden assumptions which we need to understand. The basic issue is that, if we have a hypothesis which we are testing to see if it is false, then we have to think about what could be the true hypothesis. There must always be *some* PDF which is true, if it is not the one we are testing. Hence, it is necessary to have an 'alternative hypothesis' which we compare against. The choice of the alternative can change the rejection regions for our hypothesis under test completely.

A very simple case; if we measure x_m to be in a range which has a low PDF for our hypothesis, but it has an even lower value for the alternative, then is it still correct to reject the hypothesis? H_0 is unlikely, but more likely than the other possibility. Hence, for such a case, we would probably not want the rejection region to contain those values of x.

A specific example; we want to see if there is a signal of some physical process predicted by a new theory. If there is no signal, then x will have a Gaussian distribution centred at 0. However, if the new theory is true, then x will have a Gaussian distribution of the same width centred at 1. The former is often called the 'null hypothesis' (hence H_0) as it means nothing new is happening, and the other is the alternative hypothesis H_1 with PDF $\rho_1(x)$. The PDFs for these are shown below. Note, the null hypothesis is the one we are testing, as we cannot prove the new theory is correct; there could be many other theories which give something similar. All we can do is try to reject the null hypothesis and so say there is a need for a some new theory. Clearly, if we measure a large positive value of x_m, then we can reject H_0. However, if we measure a large negative value, then even if it is many σ away from zero, there is still a higher probability for H_0 than H_1. Hence, we must pick our rejection region only to include positive values of x. Therefore, the 5% rejection region is now everything above 1.64σ; note *not* 1.96σ as we have a one-sided integral, not a two-sided one in this case. Hence, this alternative hypothesis has changed our rejection region significantly.





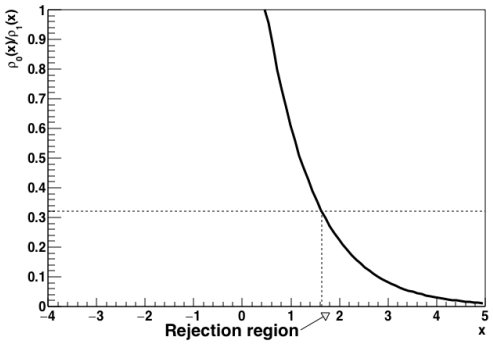
The other thing a well-defined alternative hypothesis allows us to do is find out how often we will not reject H_0, even if it is false, i.e. how often we will have a Type II error. If H_0 is false, then by definition, we assume H_1 is true. Therefore, we can integrate the PDF for H_1 over all values of x which are *not* in the rejection region. This tells us the probability β of getting an x value which will not reject H_0, even though it is false, i.e. the probability of a Type II error.

$$\beta = \int_{\text{Not rejection region}} \rho_1(x)\, dx$$

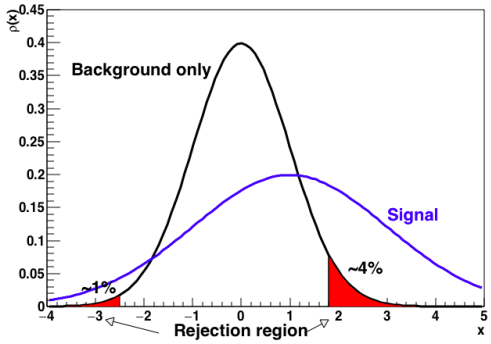
The probability $1 - \beta$ is called the 'power' of the test. For the above example

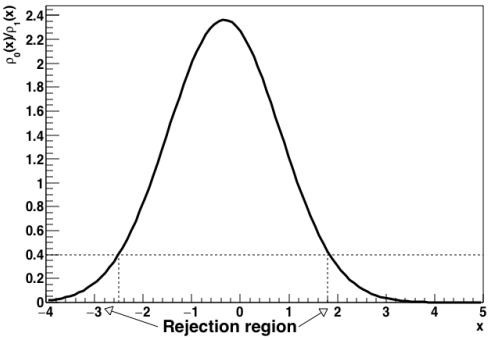
$$\beta = \int_{-\infty}^{1.64} G(x; 1.0, 1.0)\, dx = 0.74$$

Ideally, we would like to minimise both α and β. It turns out for cases where both the hypothesis under test and the alternative hypothesis are well-defined, as above, then the best rejection region selection is not $\rho_0(x) <$ some value, but is $\rho_0(x)/\rho_1(x) <$ some value, i.e. the ratio of the two PDFs. By 'best' it mean the selection which minimises α and β as far as possible. This can be proved in general for any two PDFs and is called a 'Neyman-Pearson test'. Hence, our intuitive use of which PDF was higher is indeed a sensible choice.



A final comment; most text books simply state that the alternative hypothesis defines whether you do a one-sided or two-sided integral. These are indeed the most common cases. However, there can be others; for example if the signal Gaussian PDF above had a bigger width (due to extra noise predicted as occuring in the new signal process) then at some large negative value of x, H_1 would become more probable than H_0 again. Hence, the rejection region would be two-sided, but would be a very uneven, with most at large positive x and a smaller amount at large negative x.





4 Composite alternative hypotheses

It is often the case that the alternative hypothesis is not a well defined PDF. For the example above, the new theory could predict a signal will be higher than zero, but the actual size of the signal is not known. Hence, the Gaussian for H_1 could have any mean as long as $\mu > 0$. We

cannot therefore write down an explicit PDF for H_1 and so we cannot find the power β, nor do an optimal Neyman-Pearson test. This is called a composite hypothesis.

In this case, the signal is predicted only to be higher than zero, not lower, so we know the PDF for H_0 is higher than for H_1 when $x \leq 0$, so that we would again use a one-sided test. This is often the case in physics. However, for some cases, the alternative hypothesis can be equally likely to shift the value up or down, and then we need the two-sided rejection regions shown initially. Hence, for the first part of the lecture we had been implicitly assuming a rather ill-defined and very general alternative hypothesis; in these cases, we can take a uniform distribution as the alternative hypothesis to define our rejection region.

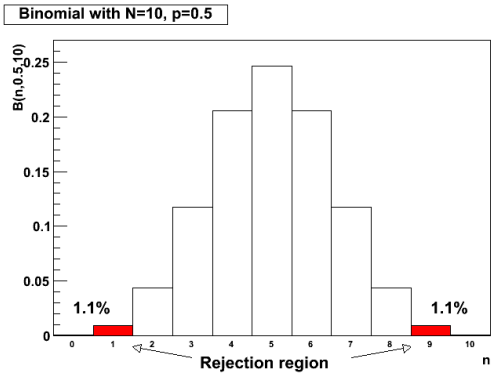
5 Discrete distributions

All of the above maps over from continuous PDFs to discrete distributions except for one complication. Due to the non-continuous nature of the probabilities, then it is usually not possible to find a region which contains exactly our chosen significance level.

For example, I flip a coin 10 times; is the result consistent with the coin being unbiased? The distribution is binomial and unbiased means the probability of heads p is equal to the probability of tails, so $p = 0.5$. Hence, I have chosen a particular value of the binomial parameter for my null hypothesis H_0. My alternative hypothesis is that the coin is biased, either towards heads or tails, so $p > 0.5$ or $p < 0.5$. Hence, I want a two-sided rejection region and I choose a significance level of 5%. Assuming H_0 is true, the probability of getting ten heads or ten tails is $p^{10} = 1/1024$, of nine heads or nine tails is $10p^9(1-p) = 10/1024$, and of eight heads or eight tails is $45p^8(1-p)^2 = 45/1024$. Hence, a rejection region of ten heads, ten tails, nine heads and nine tails totals $22/1024 = 2.1\%$, which is below 5%, while also including eight heads and eight tails would make it $112/1024 = 10.9\%$, which is above 5%.

How do we choose the rejection region in this case? Consider the continuous case again. Say we choose a 5% rejection region and find x_m falls within it. If we had instead chosen a 10% rejection region, this would completely include our previous 5% rejection region and so the measured x_m would also reject the hypothesis at 10%. Hence, a rejection at a lower significance level can also be interpreted as a rejection at a higher significance level. Another way to see this is that for the 5% rejection region, the actual x_m could be within e.g. the range which would be within a 1% rejection region. Hence, it is conservative but correct to label a lower significance level with a larger value.

Hence, for our coin example, we would choose the smaller rejection region, i.e. nine and ten heads and tails, for our 5% significance level, even though it contains only 2.1% of the probability.



6 Setting significance from the data

The handling of significance levels for discrete distributions discussed above is clearly conservative and so in reality, you might think it would be more sensible to choose our significance level to correspond to one of the possible values, e.g. either 2.1% or 10.9%, so that we are not being too cautious. Indeed, this is very common in practise; rather than setting the rejection region before doing the experiment, we do the measurement and then find out the smallest significance level which would include x_m in the rejection region. We then say we can reject the hypothesis at that significance. The significance level determined by the data is called the p value, and so we set α equal to the p value. (Obviously, if it is large, then the result does not have much to say about the hypothesis.) For example, if we got ten heads from ten coin tosses, then we would have a p value of 0.002 and we could reject the unbiased coin hypothesis at 0.2% significance level. Using the convention of quoting the number of Gaussian standard deviations corresponding to the integrated probability, this corresponds to 3.1σ. As we can see, significance levels of 3σ or more are very unlikely and hence rejections based on these are considered very significant.

Statistics of Measurement
Lecture 6 - Point estimation and maximum likelihood

Heather Graven, 22 May 2023

7 Introduction

A common issue in physics is that we have a distribution for which the parameters are not known and we do an experiment to try to determine them. This deduction of the values of the underlying parameters is called "estimation" and is the main aim of this course; it is what is often meant by the study of statistics. In the last lecture, we saw that a hypothesis test often asks the question "Is a particular value of the parameter consistent with the data?" We will now ask the question "What value of the parameter is most consistent with the data?", i.e. we want the 'best estimate' of the parameter. We saw in the last lecture that we would expect measurements to be in the regions where the PDF is largest. This leads us to the concept of maximum likelihood. Note, "likelihood" here is used in a technical sense; as we will see, it is different from probability, even though the two are used interchangeably in non-technical everyday speech. This is similar to "energy"; it means something a lot more specific in physics than the way people use it normally.

8 Estimation

Finding the 'best' estimate is called 'point estimation' as in general, for many parameters, we are determining a best point in the multidimensional parameter space. Clearly, we will also be interested in the uncertainty on the best estimate and that is discussed in Lecture 8.

There are many ways to estimate a value for a parameter from a set of experimental measurements. At a basic level, any point estimation method boils down to finding a function M of the measurements x_i which gives an estimate of the value of the parameter θ

$$\hat{\theta} = M(x_1, x_2, \ldots, x_N)$$

We write the estimate of the parameter as $\hat{\theta}$ to distinguish it from the actual (but unknown) value θ. Note, $\hat{\theta}$ is also a random variable and hence has a PDF. All estimation comes down to determining what method to use, which defines what the function M is.

You will have already met some methods for estimation; one is so obvious that it is easy to forget that it is formally an estimate at all and that is the mean of a set of measurements x_i

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

so here the function M is simply the sum divided by N.

There are many methods of estimation; some are good and some are not. Good estimates are ones which are

1. **Consistent:** meaning, in the limit of large N, that $\hat{\theta} \to$ the true value of θ.
2. **Unbiased:** meaning that $E(\hat{\theta}) =$ the true value of θ.
3. **Efficient:** meaning that $V(\hat{\theta})$ is small.

There is often a trade-off between these characteristics, e.g. you can use a method which is biased but efficient, or unbiased but has a larger variance and so is less efficient.

The above estimate $\hat{\mu}$ of the average is very widely used and that is because it is consistent, unbiased, and has good efficiency. However, it is of course limited to finding the mean, not general parameters of a distribution. In this course, we will look at two other main methods of parameter estimation; the method of maximum likelihood and the chi-squared (χ^2) method. These are more general methods which can be used to find arbitrary parameters of a distribution. In this lecture, we will look at the maximum likelihood method, which can be applied to any probability distribution. In the next lecture, we will study the chi-squared method, which is applicable when measurements are made of variables with (at least approximately) Gaussian distributions; this is of course a very common case due to the CLT. While less general than the maximum likelihood, it is often simpler to implement. In both cases, these methods are usually consistent and have good efficiency, but are only unbiased for large N.

9 The likelihood function

It is very common that we believe we know the form of the probability (density) distribution $P(x;\theta_i)$ for an experiment but this depends on unknown parameters θ_i. The point of doing the experiment is then to determine these parameters. Here we will consider the one parameter case for clarity but the method is easily generalised to more than one parameter.

Imagine we have taken N independent data points with our experiment so our data values x_i are now fixed. Because they are uncorrelated, the total probability is the product of the separate probabilities, so we can write

$$L(\theta; x_i) = \prod_{i=1}^{N} P(x_i;\theta)$$

which defines the "likelihood" function L for our experiment. Since the data values x_i are fixed, L is purely a function of the parameters θ_j and indeed, it is often just written as $L(\theta)$ to emphasise this. The likelihood is effectively the probability of seeing the data values actually observed, as a function of the parameters. Note, it is *not* a PDF for θ.

For example; say we were performing 10 binary outcome trials and saw 3 of one of the two outcomes. The probability distribution for this experiment is the binomial and we want to estimate the probability p for a single trial (the equivalent to θ above). The likelihood for this experiment has only one term in the product (as we only did one set of trials) and so is

$$L(p;3,10) = B(3;p,10) = \frac{10!}{3!7!}p^3(1-p)^7 = 120p^3(1-p)^7$$

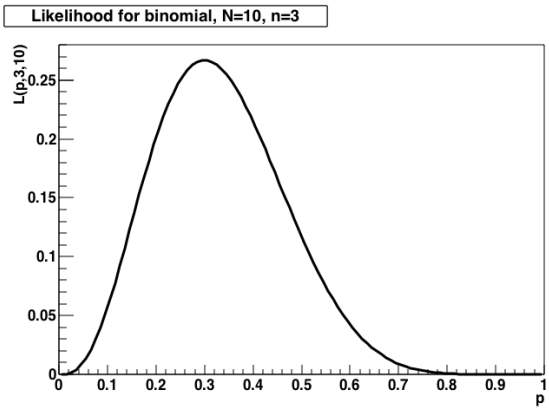
where is it clear it is a function of p only. This function is plotted below.

The obvious estimate for p would probably be the average fraction of the outcome which was observed, i.e. $\hat{p} = 3/10 = 0.3$. As is seen in the figure, the likelihood does indeed peak at that value. This implies that the value of p which gives a maximum in the likelihood function, i.e. which would make the probability of seeing the observed data the largest, is a sensible estimate.

In practice, for reasons which will become clear later, we always take logarithms of both sides of the original equation, giving the "log-likelihood". This turns the product into a sum

$$\ln[L(\theta)] = \ln\left[\prod_{i=1}^{N} P(x_i;\theta)\right] = \sum_{i=1}^{N} \ln[P(x_i;\theta)]$$

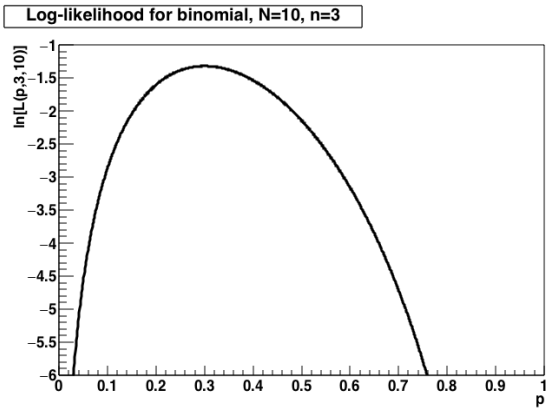
which is usually easier to handle mathematically. Since the logarithm is a monotonically increasing function, the maximum of L is also the maximum of $\ln(L)$ and so we can work with



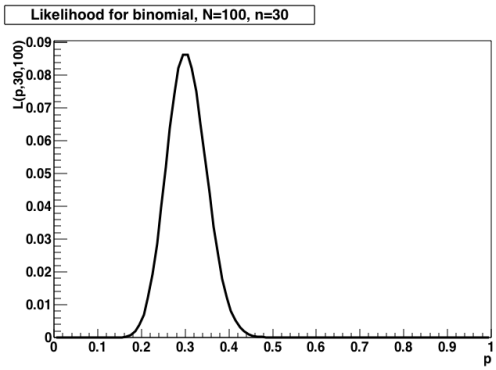
$\ln(L)$ for all calculations. The log-likelihood for the binomial above is

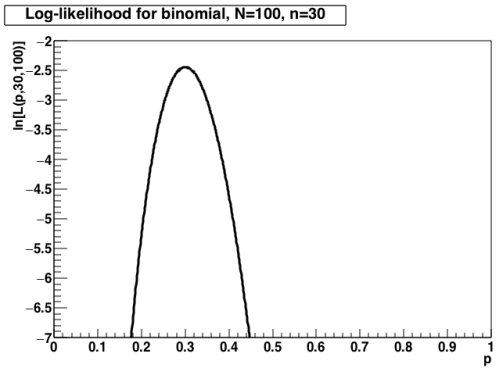
$$\ln[L(p;3,10)] = 3\ln(p) + 7\ln(1-p) + \ln(120)$$

This is shown below and it does indeed peak at the same place.



Furthermore, if we had seen 30 outcomes out of 100 trials (rather than 3 out of 10), then we would expect our estimate $\hat{p}$ to still be 0.3 but also to be more accurate. The likelihood and log-likelihood curves for this case are shown below. Again they peak at $p = 0.3$ but also it is clear they are more sharply peaked around this value. Hence, we would expect the width of the likelihood to reflect the error on the estimate. We will come back to this in Lecture 8.





10 The maximum likelihood principle

This idea can be elevated to the level of a principle; the values of the parameters which maximise the log-likelihood function are our best estimates of those parameters. Basically, we want the measurement value to be at the peak of the PDF so we adjust the PDF parameters to make this happen. By a good estimate here, we mean it turns out they are consistent and efficient, and unbiased at least in the limit of large numbers of data points. Hence, this gives us a general method for estimating parameters of a probability distribution.

Clearly we do not need to make a plot each time we have a log-likelihood function; to find the maximum we take a derivative and find the values of the parameters which make it zero. Any method of adjusting the parameters to maximise (or minimise) a value and hence make an estimation is generally called 'fitting' for the parameters. E.g for the binomial above

$$\frac{d\ln(L)}{dp} = \frac{3}{p} - \frac{7}{1-p}$$

The estimate $\hat{p}$ is given by this being zero, for which

$$\frac{3}{\hat{p}} = \frac{7}{1-\hat{p}} \qquad \text{so} \qquad 3 - 3\hat{p} = 7\hat{p} \qquad \text{so} \qquad \hat{p} = \frac{3}{10}$$

as we would expect.

11 Gaussian approximation

Let's look at the Gaussian case in some detail. The probability distribution for a single measurement x is

$$G(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad \text{so} \qquad \ln[G(x;\mu,\sigma)] = -\ln(\sigma\sqrt{2\pi}) - \frac{(x-\mu)^2}{2\sigma^2}$$

If σ is known, then one measurement x_1 allows us to estimate μ. The log-likelihood is

$$\ln[L(\mu)] = -\ln(\sigma\sqrt{2\pi}) - \frac{(x_1-\mu)^2}{2\sigma^2}$$

which obviously peaks at $\hat{\mu} = x_1$ as might be expected. The log-likelihood function is a quadratic in μ around x_1.

However, a Taylor expansion of any function around a maximum (where the first derivative is zero) will approximate to a constant plus a negative quadratic term, just as above. Hence, any PDF will look approximately Gaussian close to the peak. This 'Gaussian approximation' will be used in Lecture 8. An example of this for the binomial case discussed previously is shown in the plots below.

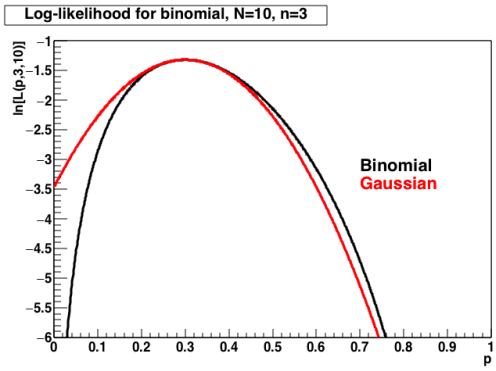
12 Gaussian parameter estimation

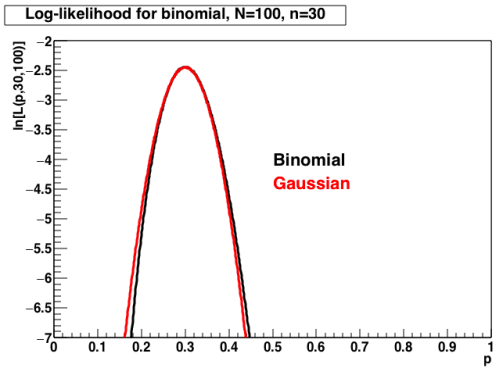
Now take the case of N measurements of x_i which has a Gaussian probability distribution with unknown mean μ and width σ. The probability distribution for a single measurement x_1 is

$$G(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad \text{so} \qquad \ln[G(x;\mu,\sigma)] = -\ln(\sigma) - \frac{1}{2}\ln(2\pi) - \frac{(x-\mu)^2}{2\sigma^2}$$

Hence, the total log-likelihood function is

$$\ln[L(\mu,\sigma)] = \sum_{i=1}^{N} \ln[G(x_i;\mu,\sigma)] = -N\ln(\sigma) - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$





We need to maximise with respect to both μ and σ in this case, so we need both derivatives

$$\frac{\partial \ln(L)}{\partial \mu} = \frac{1}{2\sigma^2}\sum_{i=1}^{N} 2(x_i - \mu), \qquad \frac{\partial \ln(L)}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i - \mu)^2$$

to be zero. We can use the first of these being zero to find the estimate on μ

$$\sum_{i=1}^{N}(x_i - \hat{\mu}) = 0 \qquad \text{so} \qquad \sum_{i=1}^{N} x_i = N\hat{\mu} \qquad \text{i.e.} \qquad \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

as we stated earlier. Solving for the estimate of σ using the other derivative

$$\sum_{i=1}^{N}(x_i - \hat{\mu})^2 = N\hat{\sigma}^2 \qquad \text{so} \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

You may know this is a biased estimate; it turns out the expectation value $E(\hat{\sigma})$ is less than the true value of σ. The unbiased estimate is

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

which differs by the 'Bessel correction' factor of $N/(N-1)$, which clearly goes to 1 for large N. Hence, this is an example of a case where the maximum likelihood gives an unbiased estimator only in the large N limit.

When using maximum likelihood estimation, one must be careful of biased estimates, particularly at small N. This can be explored by calculating the expectation value of your parameter and checking for bias.

Statistics of Measurement
Lecture 7 - The chi-squared estimation method

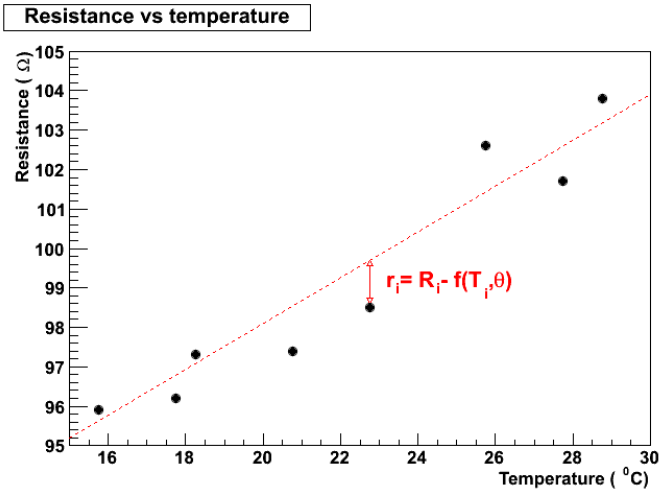
Heather Graven, 23 May 2023

13 Introduction

We saw the maximum likelihood method in the previous lecture. Today, we will look at the chi-squared method, which is applicable in the case of measurements made from an underlying Gaussian distribution. The chi-squared is closely related to the least squares method, so we will start with that.

14 Least squares

The chi-squared method is a more powerful generalisation of the method of least squares. The most common application of least squares is when we have a set of measured data values y_i which were taken while another variable x_i was varied. The idea is that the true values of y depend on x according to some functional form (or 'model') $y = f(x;\theta_1,\theta_2,\ldots) = f(x;\theta_j)$ and that we want to find the parameters θ_j of this function, where j runs over the number of parameters. One example would be to estimate the linear temperature dependence of a resistor. For this, the function we would use is a straight line $f = \theta_1 + x\theta_2$ and so there are two parameters. Measurements of the resistance are taken for various temperatures and the results of the resistance as a function of temperature are plotted below.



To do this estimation, we use the "residual" r which is the difference of each measured value from the function value

$$r_i = y_i - f(x_i;\theta_1,\theta_2,\ldots) = y_i - f(x_i;\theta_j)$$

We then take the sum of the squares of the residuals

$$S(\theta_j;y_i) = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} [y_i - f(x_i;\theta_j)]^2$$

The value of S is considered as a function of the parameters θ_j given that we have observed data values y_i. Note, the x_i are taken as known values with no (or negligible) uncertainties. We

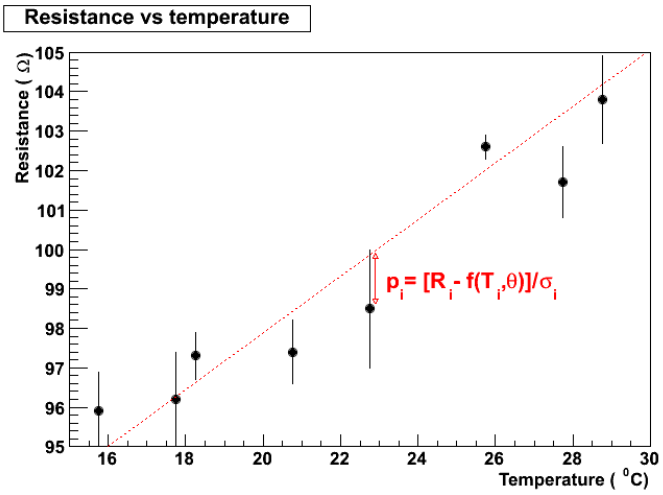
then adjust the parameters so as to minimise this sum; mathematically we need to solve the simultaneous equations

$$\frac{\partial S}{\partial \theta_j} = 0$$

where there is one equation for each parameter θ_j. This gives us the estimates $\hat{\theta}_j$. Clearly this minimisation will tend to make the function agree with the measured values as far as possible, so that the differences, i.e. the residuals, are small. Hence, it is intuitive that this will give a reasonable method for estimating the θ_j.

15 Chi-squared method

The chi-squared method is a more rigorous approach than least squares but is fundamentally similar. The difference is that it incorporates the uncertainties on the measurements y_i, here denoted by σ_i, which can be different for each point. Clearly, measurements with large uncertainties should have less power than those which are more accurate and have small uncertainties. Note, this of course requires that the uncertainties are already known.



The chi-squared method starts by dividing the residual by the uncertainty to form a quantity called the "pull" p_i

$$p_i = \frac{r_i}{\sigma_i} = \frac{y_i - f(x_i;\theta_j)}{\sigma_i}$$

If all the data points are independent, the sum of the squares of the pulls is then called the chi-squared

$$\chi^2(\theta_j; y_i) = \sum_{i=1}^{N} p_i^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i;\theta_j)}{\sigma_i}\right]^2$$

and we again want the minimum, i.e. to solve

$$\frac{\partial \chi^2}{\partial \theta_j} = 0$$

Although it is not obvious yet, it turns out that dividing the residuals by σ_i (as opposed to σ_i^2 or some other function) gives a good estimation method if the y_i have Gaussian distributions and this is where this assumption is implicitly needed.

Note, if all the uncertainties on the points are the same, i.e. $\sigma_i = \sigma$, then the chi-squared becomes

$$\chi^2(\theta_j; y_i) = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i; \theta_j)}{\sigma} \right]^2 = \frac{1}{\sigma^2} \sum_{i=1}^{N} [y_i - f(x_i; \theta_j)]^2 = \frac{S}{\sigma^2}$$

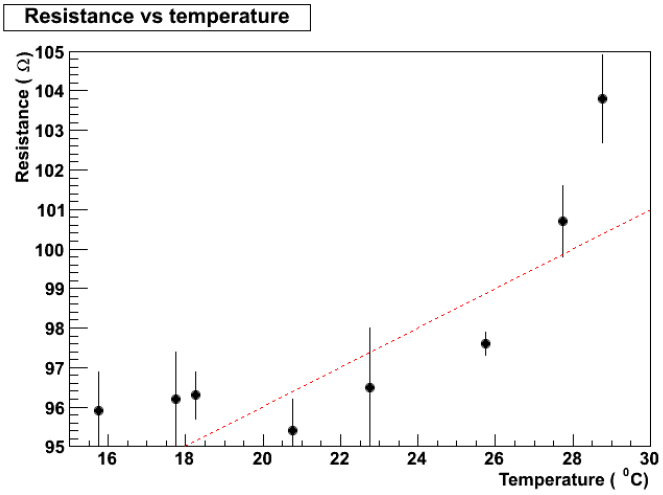
Hence the chi-squared is equal to the sum of squares of the residuals S multiplied by a constant factor. This means the values of the parameters θ_j which minimise the chi-squared will be identical to those resulting from the least squares method. Hence, the estimates of the parameters from the two methods will be the same in this particular case.

Solving for the minimum can be done analytically for simple functions, but in most cases must be done numerically on a computer. Such a process of finding the values of parameters which minimise (or maximise) a function is generally called 'fitting' the parameters. Both the chi-squared and the maximum likelihood methods are therefore examples of fitting the parameters.

16 Goodness of fit

Besides allowing for differing sizes of uncertainties on the y_i, the chi-squared method is more useful even if all the σ_i have the same value (when it gives identical estimates for the θ_j as the simple least squares method). This is because the minimum chi-squared value itself gives further information, specifically on what is called the "goodness of fit". This tells us if the function we have assumed is a sensible one or not.

It is clear that if the function *is* correct, then the values of the y_i and the function $f(x_i; \theta_j)$ for each x_i should not be too different. Conversely, a bad fit will have the points far from the function; see an example below.



Specifically, for a good fit, we would expect the values of the y_i to differ from the fitted function by amounts similar to the σ_i as they are (assumed to be) Gaussian distributed. Specifically, if the fitted function is sensible, it should be close to the mean for that x_i and hence the residual (i.e. the fluctuation around that mean) would be expected to have a standard deviation around the fitted function of roughly σ_i. Hence, each term in the chi-squared sum would be expected to be $(r_i/\sigma_i)^2 \sim 1$ and hence the chi-squared at the minimum should have a value of $\chi^2_{\text{min}} \sim N_{\text{data}}$. A much bigger value would imply that the wrong function is being used and so can help to identify if an underlying theory is incorrect.

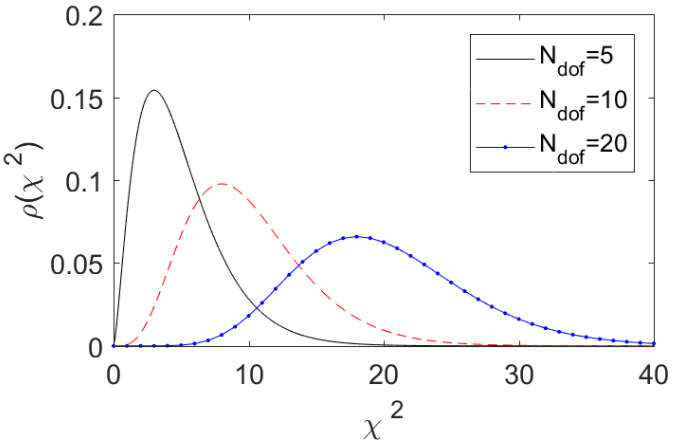
To be more precise, it is not actually correct to compare the chi-squared value to N_{data}, but a correction is needed. To see this, consider an extreme example; a straight line fit to two data

points will always give a minimum chi-squared of exactly zero. This does not mean the straight line fit is good; this is just saying there is a solution, not a fit, if the number of data points is equal to the number of parameters. It turns out the critical measure is the "number of degrees of freedom" $N_{\rm dof}$, which is defined to be the difference between the number of data points being used in the fit and the number of parameters being estimated in the fit

$$N_{\rm dof} = N_{\rm data} - N_{\rm paras}$$

If there are more data points than parameters, i.e. $N_{\rm dof} > 0$, then there are more constraints than there are parameters to be determined and a fit is needed. The $\chi^2_{\rm min}$ value then gives information on the goodness of fit. If the number of data points is equal to the number of parameters, i.e. $N_{\rm dof} = 0$, then the equations can be solved rather than fitted, and the chi-squared will be identically zero so we have no information on the goodness of fit. Finally, if the number of data points is less than the number of parameters, i.e. $N_{\rm dof} < 0$, the the system is underconstrained and no unique determination of the parameters is possible. In terms of the minimum value of the chi-squared, it turns out that for $N_{\rm dof} > 0$ we would expect $\chi^2_{\rm min} \sim N_{\rm dof}$ for a good fit, while values much bigger than this would imply the function being fitted does not actually describe the data well.

The chi-squared is actually a random variable with a PDF. At a basic level, given that all the measurement distributions are Gaussian, then we can calculate this PDF for the chi-squared value. It turns out this only depends on the number of degrees of freedom, not the number of data points or parameters separately, i.e. $\rho(\chi^2; N_{\rm dof})$. The PDF has an expectation value of $N_{\rm dof}$ and has a long tail out to infinity, as shown in the plot below for different $N_{\rm dof}$.



The goodness of fit is therefore effectively a statement about whether our fit gives us a chi-squared value near the PDF peak or in the tail at higher chi-squared. This is often quantified in terms of a "p-value" which gives the probability of getting the observed or a worse (higher) chi-squared value. There are tables of p-values in statistics books and on the web, and software packages may have functions to calculate the p-value from a given chi-squared and number of degrees of freedom. However, as the p-value corresponds to integrating the chi-squared distribution from the observed value up to infinity, then it corresponds exactly to what we previosuly called the one-sided significance level, set by the data value. Hence, stating that the goodness of fit is poor is precisely a hypothesis test, i.e. we say we reject the hypothesis that the data agree with the function chosen for the fit if the goodness of fit is bad.

17 Chi-squared method for binned distributions

We are often trying to find the parameters of a function which describes the shape of some data distribution (e.g. the shape of the histogram). By separating the data into small ranges ("bins") as in a histogram, we can count the number of times the variable falls into each bin.

This lends itself to a chi-squared treatment if the numbers of events in each bin are large, i.e. $n_i \gg 1$ for each bin i. This is because the probability distribution for the number n_i in a single bin is (to a good approximation) a Poisson, where the mean of the Poisson is just the integral of the function over the bin. Specifically

$$\mu_i(\theta_j) = \int_{\text{Bin } i} f(x;\theta_j)\, dx$$

It is also common in practise that μ_i is approximated to

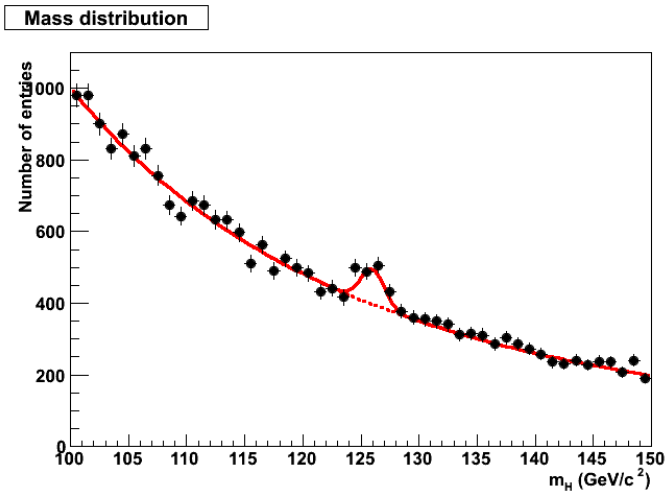
$$\mu_i(\theta_j) \approx f(x_i;\theta_j)\, \Delta x$$

where x_i is the value of x at the centre of each bin and Δx is the bin width. Given this, n_i has a distribution of $P[n_i;\mu_i(\theta_j)]$ Hence, because of the Central Limit Theorem for large n_i as seen in Lecture 4, this is approximately Gaussian, with $\sigma_i \approx \sqrt{n_i}$. Hence the chi-squared can be approximated by

$$\chi^2(\theta_j;\underline{n}) = \sum_{i=1}^{N} \frac{[n_i - \mu_i(\theta_j)]^2}{n_i}$$

where the sum is over all the bins. This method is widely used but it is important it is only done when all bins have a large number of entries, i.e. $n_i \gg 1$, which is typically at least 20 entries.

An example of this application is the discovery of the Higgs boson; it was discovered in 2012 by finding the mass of its decay products. There is a large background of random combinations which gave a smooth shape in the mass distribution, while the Higgs itself gives a narrow Gaussian peak in mass. By fitting for the background shape parameters, the background-only hypothesis could be rejected due to a poor goodness of fit; in the Higgs case the significance level was 5σ which is very strong evidence for rejection. Knowing that the smooth background function was not sufficient for a good fit to the data, including parameters for the mean (i.e. the Higgs mass) and width of the Gaussian allowed its existence to be established and its mass to be determined.



18 Connection to maximum likelihood estimation

Let's consider the case of a set of measurements y_i of Gaussian distributed random variables, each with a different mean μ_i and width σ_i. We will assume the means depend on some parameters θ_j so $\mu_i(\theta_j)$, but that the widths σ_i are known. The likelihood for one such measurement is

$$L_i = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-(y_i-\mu_i)^2/2\sigma_i^2}$$

so the log-likelihood is

$$\ln(L_i) = -\frac{(y_i-\mu_i)^2}{2\sigma_i^2} - \ln(\sigma_i\sqrt{2\pi})$$

Hence, for all the measurements, the total log-likelihood is

$$\ln(L) = \sum_i \ln(L_i) = -\sum_i \frac{(y_i-\mu_i)^2}{2\sigma_i^2} - \sum_i \ln(\sigma_i\sqrt{2\pi}) = -\sum_i \frac{(y_i-\mu_i)^2}{2\sigma_i^2} + C$$

where C is a term which does not depend on the μ_i and hence also not on the θ_j parameters. This means

$$2C - 2\ln(L) = \sum_i \frac{(y_i-\mu_i)^2}{\sigma_i^2}$$

The term on the right can be seen to be what we defined as the chi-squared, i.e.

$$\chi^2 = 2C - 2\ln(L)$$

When we considered the chi-squared, we discussed the values as being described by some function but effectively all this does is to define a different mean for each measurement, i.e.

$$\mu_i(\theta_j) = f(x_i;\theta_j)$$

In these terms, then we would maximise the log-likelihood to find the estimates $\hat{\theta}_j$. However, since they differ by a sign, *maximising* the log-likelihood is equivalent to *minimising* the chi-squared and hence the estimates from either method will be identical. In fact, it is common to work with $-2\ln(L)$ and find a minimum rather than a maximum, as this is then directly comparable to the chi-squared in the Gaussian case.

In general however, there is no goodness-of-fit measure for a likelihood fit, unlike the chi-squared method. In the Gaussian case, the constant C above is calculable and so the chi-squared can be found from the maximum likelihood. However, for non-Gaussian cases (which are exactly when the maximum likelihood is most useful), the above form is not found and there is no such goodness-of-fit measure.

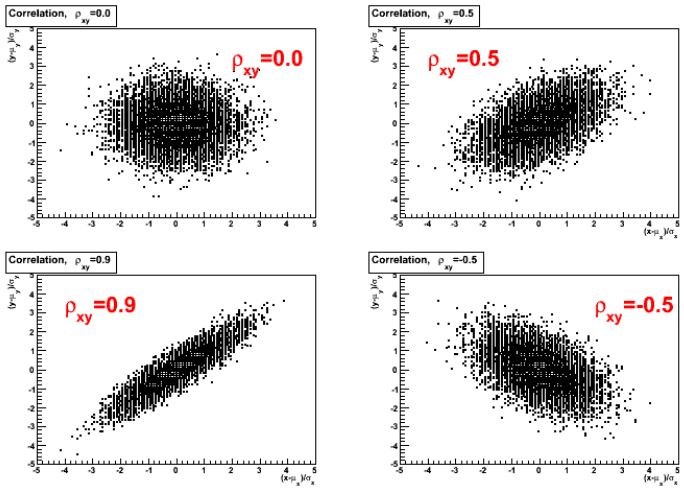
19 Non-examinable: the error matrix

The general formulation for uncertainties with more than one variable is using the 'error matrix' $\underline{\underline{E}}$, sometimes called the 'covariance matrix'. For independent (so-called 'uncorrelated') variables, the error matrix is diagonal and equal to

$$\underline{\underline{E}} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots \\ 0 & \sigma_2^2 & 0 & \dots \\ 0 & 0 & \sigma_3^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where σ_i is the uncertainty on variable i.

This can be generalised; the zeros off-diagonal indicate that there are no correlations between the measurements. However, conversely, if the variables *are* correlated, then these off-diagonal elements are no longer zero, but have the form $\rho_{ij}\sigma_i\sigma_j$, where ρ_{ij}, the 'correlation coefficient', must be in the range $-1 \leq \rho_{ij} \leq 1$. Clearly, $\rho_{ij} = 0$ corresponds to the uncorrelated case. For non-zero ρ_{ij}, then if one variable fluctuates higher than the average, then the other will also tend to fluctuate higher (if they are correlated, i.e. $\rho_{ij} > 0$) or lower (if they are anticorrelated, i.e. $\rho_{ij} < 0$), as shown in the figure below.



Hence, the general error matrix is

$$\underline{\underline{E}} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \cdots \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \cdots \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Knowing about the error matrix allows us to write down the most general form for the propagation of errors formula. If a quantity z depends on several variables x_i, i.e. $z = z(\underline{x})$, then by forming a 'derivative vector' $\underline{d}$

$$\underline{d} = \begin{pmatrix} \partial z/\partial x_1 \\ \partial z/\partial x_2 \\ \partial z/\partial x_3 \\ \vdots \end{pmatrix}$$

the error on z is given by

$$\sigma_z^2 = \underline{d}^T \underline{\underline{E}}\, \underline{d}$$

For example, with two variables $x_1 = x$, $x_2 = y$, then generally

$$\begin{aligned} \sigma_z^2 &= (\,\partial z/\partial x \quad \partial z/\partial y\,) \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \begin{pmatrix} \partial z/\partial x \\ \partial z/\partial y \end{pmatrix} \\ &= (\,\partial z/\partial x \quad \partial z/\partial y\,) \begin{bmatrix} (\partial z/\partial x)\sigma_x^2 + (\partial z/\partial y)\rho_{xy}\sigma_x\sigma_y \\ (\partial z/\partial x)\rho_{xy}\sigma_x\sigma_y + (\partial z/\partial y)\sigma_y^2 \end{bmatrix} \\ &= \left(\frac{\partial z}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial z}{\partial y}\right)^2 \sigma_y^2 + 2\left(\frac{\partial z}{\partial x}\right)\left(\frac{\partial z}{\partial y}\right)\rho_{xy}\sigma_x\sigma_y \end{aligned}$$

This reduces to the form given previously for the propagation of errors formula in the uncorrelated case, i.e. when $\rho_{xy} = 0$.

20 Non-examinable: chi-squared method for correlated uncertainties

There is an even more complete version of the chi-squared method, which occurs when the measurements y_i are correlated. The general formulation for uncertainties with more than one variable is using the error matrix $\underline{\underline{E}}$, as discussed in the previous section.

The expression for the chi-squared can be written in terms of vectors and matrices. Defining the "residual vector" $\underline{r}$ as

$$\underline{r} = \begin{bmatrix} y_1 - f(x_1;\theta_j) \\ y_2 - f(x_2;\theta_j) \\ y_3 - f(x_3;\theta_j) \\ \vdots \end{bmatrix}$$

so its transpose is

$$\underline{r}^T = [\, y_1 - f(x_1;\theta_j) \quad y_2 - f(x_2;\theta_j) \quad y_3 - f(x_3;\theta_j) \quad \ldots\,]$$

then the chi-squared can be written as

$$\chi^2 = [\, y_1 - f(x_1;\theta_j) \quad y_2 - f(x_2;\theta_j) \quad y_3 - f(x_3;\theta_j) \quad \ldots\,] \begin{pmatrix} 1/\sigma_1^2 & 0 & 0 & \ldots \\ 0 & 1/\sigma_2^2 & 0 & \ldots \\ 0 & 0 & 1/\sigma_3^2 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{bmatrix} y_1 - f(x_1;\theta_j) \\ y_2 - f(x_2;\theta_j) \\ y_3 - f(x_3;\theta_j) \\ \vdots \end{bmatrix}$$

The matrix in the middle is called the "weight matrix" $\underline{\underline{W}}$ and the chi-squared can be written compactly as

$$\chi^2 = \underline{r}^T \underline{\underline{W}}\, \underline{r}$$

The inverse of the weight matrix is obviously

$$\underline{\underline{E}} = \underline{\underline{W}}^{-1} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \ldots \\ 0 & \sigma_2^2 & 0 & \ldots \\ 0 & 0 & \sigma_3^2 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

From above, this will be recognised as the form of the error matrix for uncorrelated variables. As discussed previously, the more general error matrix, allowing for possible correlations, is

$$\underline{\underline{E}} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \ldots \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \ldots \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Inverting this matrix to get the weight matrix then allows the chi-squared to be expressed in the same way as before

$$\chi^2 = \underline{r}^T \underline{\underline{W}}\, \underline{r}$$

This is the general form; even for the correlated case, then combining the residual vectors and weight matrix as above gives the chi-squared.

Statistics of Measurement
Lecture 8 - Confidence intervals and parameter uncertainties

Heather Graven, 25 May 2023

21 Introduction

We have discussed estimating parameters in the last two lectures, using either the maximum likelihood or chi-squared methods. However, we have not discussed how to find the uncertainty on the parameters we estimate. We have also not defined precisely what we mean by an uncertainty on a parameter. This is where the real differences between frequentist and Bayesian definitions of probability arise. We shall discuss the frequentist approach here and then discuss Bayesian estimation in the next lecture.

The first thing to do is understand how to find the uncertainties on the parameters when using the two methods we have studied. In principle there is nothing new here. Both methods can be considered as giving an equation for the parameter(s) θ as a function of the measurements x_i

$$\hat{\theta} = M_\theta(x_1, x_2, \ldots, x_N)$$

so, at least in a linear approximation, we can just apply the propagation of errors formula to get the uncertainty on each θ_i, i.e.

$$\sigma^2_{\theta_i} = \sum_i \left(\frac{\partial M_\theta}{\partial x_i}\right)^2 \sigma^2_{x_i}$$

However, in practical terms this can often be inconvenient to calculate, particularly when the maximum or minimum is only found numerically. Also, the propagation of errors equation is only itself an approximation. Let's consider each of the two methods in turn.

22 Maximum likelihood parameter uncertainties

The central limit theorem says for large N, the mean and variance of the sum of repeated measurements of random variables are distributed according to a Gaussian of mean $\mu = \sum_i E_i$ and width $\sigma^2 = \sum_i V_i$. With enough measurements, many PDFs approximate to Gaussians. It turns out that the likelihood's dependence on the parameters is then also Gaussian. It is therefore very common to take an approximation that the likelihood is close to a Gaussian.

For simplicity, assume there is only one parameter θ. We can do a Taylor expansion of the log-likelihood in terms of θ around the maximum, which is where the parameter is $\theta = \hat{\theta}$. The Taylor expansion then gives

$$\ln[L(\theta)] = \ln[L(\hat{\theta})] + \left.\frac{d\ln(L)}{d\theta}\right|_{\hat{\theta}} (\theta - \hat{\theta}) + \left.\frac{d^2\ln(L)}{d\theta^2}\right|_{\hat{\theta}} \frac{(\theta-\hat{\theta})^2}{2!} + \ldots$$

Because $\hat{\theta}$ is defined to be at the maximum, the first derivative is zero. The second derivative is negative (since it is a maximum) but is some constant when evaluated at $\hat{\theta}$; let this be

$$\frac{1}{\Sigma^2} = -\left.\frac{d^2\ln(L)}{d\theta^2}\right|_{\hat{\theta}}$$

where the negative sign means Σ^2 is positive and hence Σ is real. In terms of Σ, then

$$\ln[L(\theta)] = \ln[L(\hat{\theta})] - \frac{(\theta-\hat{\theta})^2}{2\Sigma^2} + \ldots$$

If the higher order terms are now ignored, then taking exponentials gives

$$L(\theta) \approx L(\hat{\theta}) e^{-(\theta-\hat{\theta})^2/2\Sigma^2}$$

which is clearly a Gaussian function for θ. It is also seen that the quantity Σ generally is the width of the Gaussian, which is therefore the uncertainty on $\hat{\theta}$. Hence, in this approximation (or exactly if the function is truely a Gaussian), then the uncertainty is given by the equation above involving the second derivative, evaluated with $\theta = \hat{\theta}$. Warning: the above equation for $1/\Sigma^2$ is only valid for the one parameter case. For more parameters, this gives the inverse of a matrix, which has to be inverted to get the uncertainties (see the appendix).

Example: We have a large number N of measurements of an exponential random variable. We will express the exponential PDF as $e^{-x/a}/a$, where the parameter a (to be estimated) is the average, as shown in Lecture 3. The likelihood and hence log-likelihood for a single measurement is

$$L_i(a) = \frac{1}{a} e^{-x_i/a} \qquad \text{so} \qquad \ln[L_i(a)] = -\ln(a) - \frac{x_i}{a}$$

Hence, the total log-likelihood is

$$\ln[L(a)] = \sum_i \ln[L_i(a)] = -N \ln(a) - \frac{1}{a} \sum_i x_i$$

The derivative is

$$\frac{d \ln(L)}{da} = -\frac{N}{a} + \frac{1}{a^2} \sum_i x_i$$

so the estimate is given by this being zero, for which

$$\frac{N}{\hat{a}} = \frac{1}{\hat{a}^2} \sum_i x_i \qquad \text{so} \qquad \hat{a} = \sum_i x_i / N$$

as would be expected. The second derivative is

$$\frac{d^2 \ln(L)}{da^2} = \frac{N}{a^2} - \frac{2}{a^3} \sum_i x_i \qquad \text{so} \qquad \left. \frac{d^2 \ln(L)}{da^2} \right|_{\hat{a}} = -\frac{N}{\hat{a}^2}$$

and so is negative as required for a maximum. Therefore, the uncertainty on the estimate is approximately

$$\sigma_{\hat{a}} \approx \frac{1}{\sqrt{-d^2 \ln(L)/da^2|_{\hat{a}}}} = \frac{\hat{a}}{\sqrt{N}}$$

The same result would be obtained from using the propagation of errors formula on the result for $\hat{a}$, remembering that the standard deviation of an exponential random variable is a. Explicitly

$$\frac{\partial \hat{a}}{\partial x_j} = \frac{1}{N}$$

so

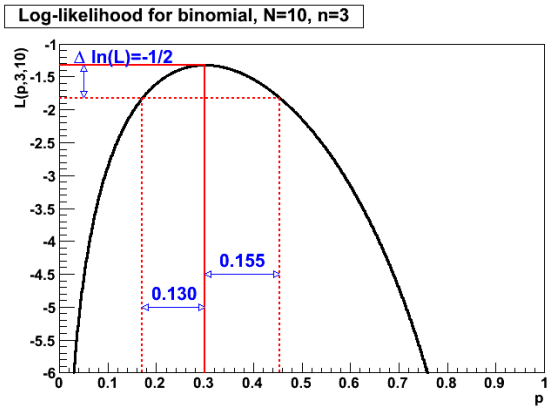
$$\sigma_{\hat{a}}^2 = \sum_j \left(\frac{\partial \hat{a}}{\partial x_j} \right)^2 \sigma_{x_j}^2 = \sum_j \frac{1}{N^2} \hat{a}^2 = \frac{\hat{a}^2}{N}$$

giving the same result for $\sigma_{\hat{a}}$.

What if the PDF is not even approximately Gaussian or/and propagation of errors is not a good approximation? The uncertainty can still be evaluated but by using a different method. For the Gaussian case, then the value of the log-likelihood at $\theta = \hat{\theta} \pm \Sigma$ is

$$\ln[L(\theta)] = \ln[L(\hat{\theta})] - \frac{(\hat{\theta} \pm \Sigma - \hat{\theta})^2}{2\Sigma^2} = \ln[L(\hat{\theta})] - \frac{\Sigma^2}{2\Sigma^2} = \ln[L(\hat{\theta})] - \frac{1}{2}$$

Hence, the uncertainty is the range which makes the log-likelihood change by $-1/2$. It turns out this is true for non-Gaussian likelihoods too; the uncertainty can be evaluated by changing the parameter and seeing what range gives $\Delta \ln(L) \geq -1/2$. (In principle, this is only exact in the large N limit, but it is normally a very good approximation anyway.) In general, this can require a different shift in θ when going up compared to down, and so this gives asymmetric uncertainties. An example for a binomial measurement is shown below. In this case, you will see them written as e.g. $0.30^{+0.15}_{-0.13}$ for this example.



23 Chi-squared parameter uncertainties

As discussed in lecture 7, the chi-squared is a function of the parameters

$$\chi^2(\theta; y_i) = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i;\theta)}{\sigma_i} \right]^2$$

and the estimates of the parameters are taken to be the values that give the minimum chi-squared.

However, consider expanding the chi-squared around the minimum using a Taylor expansion in the parameters. We will again consider only one parameter for clarity. Since it is a minimum, there is no linear term and so we know immediately it must be of the form

$$\chi^2(\theta) \approx \chi^2(\hat{\theta}) + \left. \frac{d^2\chi^2}{d\theta^2} \right|_{\hat{\theta}} \frac{(\theta - \hat{\theta})^2}{2} + \ldots$$

Since the best-fit chi-squared is a minimum, then the second derivative must be positive. Hence, let us express the second derivative as

$$\left. \frac{d^2\chi^2}{d\theta^2} \right|_{\hat{\theta}} = \frac{2}{\Sigma^2} \qquad \text{i.e.} \qquad \frac{1}{\Sigma^2} = \frac{1}{2} \left. \frac{d^2\chi^2}{d\theta^2} \right|_{\hat{\theta}}$$

such that we can write

$$\chi^2(\theta) \approx \chi^2(\hat{\theta}) + \frac{(\theta - \hat{\theta})^2}{\Sigma^2}$$

We can interpret this as being some sort of 'residual' of the parameters around the best estimate of $\hat{\theta}$. With this interpretation, then it is clear the uncertainty on the best estimate is Σ. It can be

shown (see appendix) that Σ is indeed the uncertainty on the parameter that would be obtained from using the propagation of errors formula, i.e.

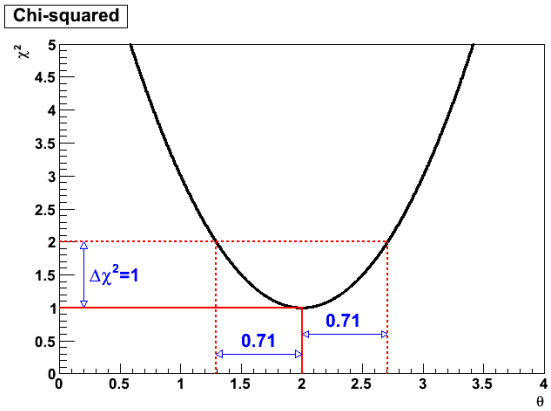
$$\Sigma^2 = \sum_i \left(\frac{\partial\theta}{\partial y_i}\right)^2 \sigma_i^2$$

so that our interpretation is indeed correct.

It is sometimes the case that the second derivative is easy to calculate and so the uncertainty can be found using the above method. However, the chi-squared method can be used for arbitrarily complicated functions $f(x_i;\theta)$ and in many cases it is impracticable to find the uncertainty this way. Luckily, there is a simple numerical method, which is to change the parameter from its best estimate and see what happens to the value of the chi-squared. If we shift the parameter θ from $\hat{\theta}$ (the value which gives the minimum chi-squared) to $\hat{\theta} \pm \Sigma$, then the chi-squared becomes

$$\chi^2(\hat{\theta} \pm \Sigma) \approx \chi^2(\hat{\theta}) + \frac{\Sigma^2}{\Sigma^2} \approx \chi^2(\hat{\theta}) + 1$$

Hence, if we numerically evaluate the chi-squared for various different values of θ, then the range for which the chi-squared changes by up to one unit gives the uncertainty on $\hat{\theta}$. An example is shown below, where the result would be quoted as $\hat{\theta} = 2.0 \pm 0.7$.



We showed in the last lecture that for the Gaussian case, the chi-squared is related to the log-likelihood by $\chi^2 \sim -2\ln(L)$. Hence, changing the chi-squared by $+1$ is equivalent to changing the log-likelihood by $-1/2$ and so these two uncertainty estimates are in agreement.

24 Confidence intervals for random variables

Let's think about confidence intervals in relation to random variables, then we will go back to estimated parameters in the next section. Back in lecture 3, we discussed the probability contained within certain ranges of the Gaussian PDF, e.g. the region $\pm 1\sigma$ of the mean contained 68.3% of the integrated PDF

$$\int_{\mu-\sigma}^{\mu+\sigma} G(x;\mu,\sigma)\,dx = \frac{2}{\sqrt{\pi}} \int_0^{1/\sqrt{2}} e^{-y^2}\,dy = \mathrm{erf}(1/\sqrt{2}) = 0.683$$

The same could be done for other ranges; e.g. the probability within $\pm 2\sigma$ is

$$\int_{\mu-2\sigma}^{\mu+2\sigma} G(x;\mu,\sigma)\,dx = \frac{2}{\sqrt{\pi}} \int_0^{2/\sqrt{2}} e^{-y^2}\,dy = \mathrm{erf}(2/\sqrt{2}) = 0.954$$

Considering the $\pm 1\sigma$ case, this means that a random measurement of x arising from a Gaussian distribution has a probability of 68.3% of lying within this range. This is referred to as the 'confidence interval', i.e. we have 68.3% confidence that the value will lie in the interval within $\pm 1\sigma$ of the mean. Hence, if we take many such measurements, then 68.3% of them (which is roughly 2/3) will be within $\pm 1\sigma$ of the mean.

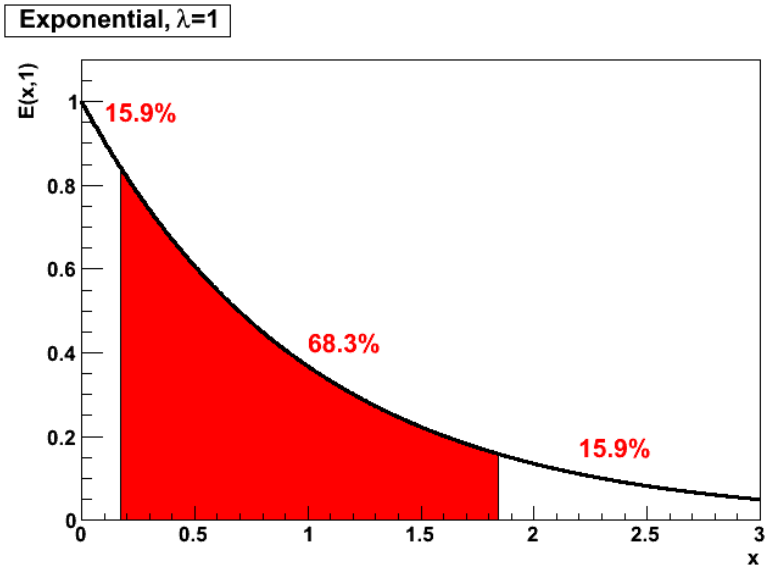
Ranges can also be defined for non-Gaussian distributions. Even though the specific value of 68.3% arises from the integration of the Gaussian, it is still very common to use a range containing this amount of the integral, even for non-Gaussian distributions. It is even common to refer to this 68.3% confidence range as being the 'one sigma' range, even though there may be no such parameter σ in the distribution, and the standard deviation may represent a different fraction of the total probability. However, even in trying to calculate the 68.3% range, there is complication, namely non-Gaussian distributions can be asymmetric. In this case, there is no unique answer as to how to define this range. We need

$$\int_a^b \rho(x)\, dx = 0.683$$

but this only gives one constraint on the two limits a and b. For example, for an exponential

$$\int_a^b \lambda e^{-\lambda x}\, dx = [-e^{-\lambda x}]_a^b = e^{-\lambda a} - e^{-\lambda b} = 0.683$$

we have two unknowns in the equation. A common way to handle this is to say that the amount outside the range, which is $100\% - 68.3\% = 31.7\%$, has to be equally divided above and below the range, so there is 15.9% on either side. This is shown for an exponential below.



25 Confidence intervals for estimated parameters

The meaning of confidence intervals for random variables should hopefully be clear. The confidence interval is associated with the specific probability of the random variable being within that interval, for example, the 68.3% confidence interval for a random variable with a Gaussian PDF is $\pm 1\sigma$.

Now back to parameters. Parameters are *not* random variables but have a exact value, even if this is unknown to us. E.g. the electron mass is a fixed value, even if we are trying to measure it in an experiment.

So what do we really mean by uncertainties on parameters, e.g. by finding $\Delta \ln(L) = -1/2$ or $\Delta\chi^2 = +1$? What we are in fact doing is giving a range. Clearly, the true value is either within this range or is not, and so any particular experiment may be right or wrong about the true value being within the range. However, the range chosen is such that if we did a large number of experiments to find the parameter value, then 68.3% (or 95% or whatever the quoted confidence interval) of those experiments will contain the true value within the derived confidence interval. This is the meaning of a confidence interval for a parameter in the frequentist interpretation and this is what we actually mean when we quote the 'uncertainty' on a parameter. This is subtly different from that for a random variable and can take a while to understand. In practise, we handle the parameter errors in the same way as random variable errors, e.g. using the propagation of errors formula, but there is an underlying conceptual difference.

Furthermore, sometimes we are only interested in one side of the confidence interval, e.g. if we are trying to determine a rate of a very rare process, we know a rate cannot be negative so we only want to set an upper limit. In this case, it is common to use the 90% or, more usually, 95% confidence level, but only on one side of the range. Hence, just like hypothesis testing, we can set one-sided or two-sided confidence intervals.

26 Physical constraints

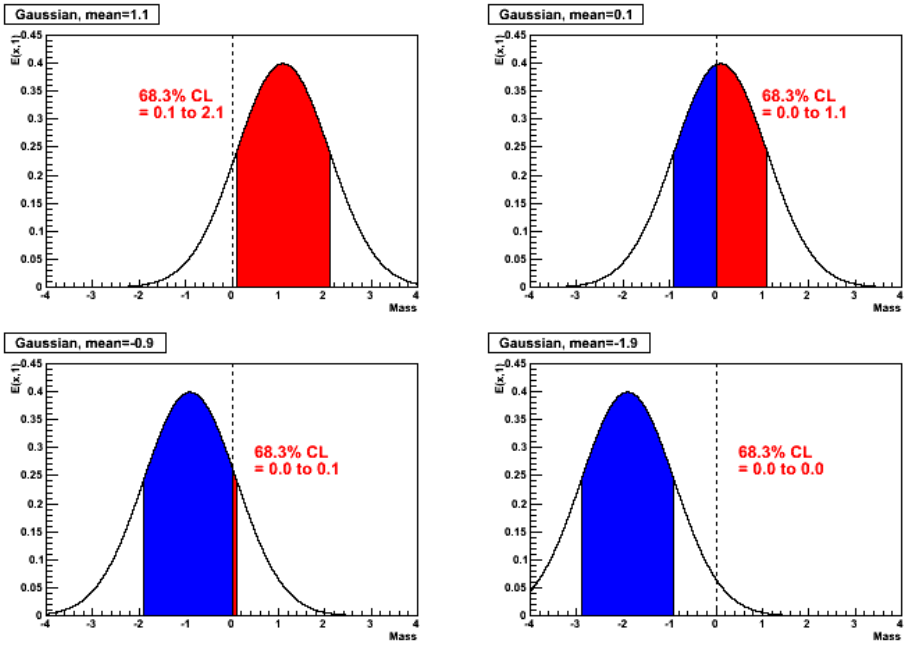
One of the problems with frequentist methods is that we can end up with results outside the physical region. For example, we want to measure the mass of some powder of around 1 g using a set of weighing scales with 1% accuracy. The dish containing the powder is known to weigh 100 g so the uncertainty on the weight is 1 g. If we happened to measure 101.1 g then in the frequentist approach, we would set the 68.3% confidence interval for the mass of the powder to be 0.1 to 2.1 g, which is fine. However, it is not unlikely that the measurement could give something like 100.1 g, which would give a range from -0.9 to 1.1 g. However, the mass cannot be negative so the 68.3% confidence interval would have to be quoted as 0.0 to 1.1 g, which looks like it is more accurate, even though it uses the same apparatus. A bigger problem arises if the measurement happens to yield 99.1 g; the range would be -1.9 to 0.1 g so the apparent confidence interval would be 0.0 to 0.1 g. However, saying the mass is below 0.1 g when the scales are only accurate to 1 g is clearly not right. A final example might be a fluctuation where the measurement yields 98.1 g; the range would be -2.9 to -0.9 g which means there is no 68.3% confidence interval in the physical region. These cases are illustrated in the plots below.

We know these low measurements must have been 'unlucky' in some sense. Hence, we would expect them to be more likely to be some of the 31.7% of confidence intervals which do *not* contain the true value. However, it is very hard to quantify this in a consistent way in the frequentist approach. We will see how the Bayesian approach handles this in the next lecture.

27 Non-examinable: Appendix

27.1 Single parameter uncertainty from chi-squared

This section goes through a derivation of the uncertainty of a single parameter and compares it with a propagation of errors approach like we did before with the likelihood, but now using χ^2. Consider a simple chi-squared case with one parameter θ, where the function is linear in the



parameter (although it can be arbitrarily complicated in x), so

$$f(x;\theta) = a(x) + b(x)\theta$$

for any functions $a(x)$ and $b(x)$. Hence,

$$\chi^2(\underline{\theta};\underline{y}) = \sum_{i=1}^{N}\left[\frac{y_i - f(x_i;\theta)}{\sigma_i}\right]^2 = \sum_{i=1}^{N}\left[\frac{y_i - a(x_i) - b(x_i)\theta}{\sigma_i}\right]^2$$

and the estimate of the parameter is given by solving

$$\left.\frac{d\chi^2}{d\theta}\right|_{\hat{\theta}} = 0 = -2\sum_{i=1}^{N}\frac{[y_i - a(x_i) - b(x_i)\hat{\theta}]b(x_i)}{\sigma_i^2}$$

This gives

$$\sum_{i=1}^{N}\frac{[y_i - a(x_i)]b(x_i)}{\sigma_i^2} = \hat{\theta}\sum_{j=1}^{N}\frac{b(x_j)^2}{\sigma_j^2}$$

so

$$\hat{\theta} = \frac{\sum_{i=1}^{N}[y_i - a(x_i)]b(x_i)/\sigma_i^2}{\sum_{j=1}^{N} b(x_j)^2/\sigma_j^2}$$

Note, the σ_i^2 terms do *not* cancel out here as they are inside separate summations.

The uncertainty on $\hat{\theta}$ arises as the y_i have uncertainties σ_i. Hence, we can use propagation of errors to find this uncertainty. From above

$$\frac{\partial\hat{\theta}}{\partial y_i} = \frac{b(x_i)/\sigma_i^2}{\sum_{j=1}^{N} b(x_j)^2/\sigma_j^2}$$

so that by propagation of errors

$$\sigma_{\hat{\theta}}^2 = \sum_{i=1}^{N}\left(\frac{\partial\hat{\theta}}{\partial y_i}\right)^2\sigma_i^2 = \sum_{i=1}^{N}\frac{b(x_i)^2/\sigma_i^4}{\left(\sum_{i=j}^{N} b(x_j)^2/\sigma_j^2\right)^2}\sigma_i^2 = \frac{\sum_{i=1}^{N} b(x_i)^2/\sigma_i^2}{\left(\sum_{j=1}^{N} b(x_j)^2/\sigma_j^2\right)^2} = \frac{1}{\sum_{i=1}^{N} b(x_i)^2/\sigma_i^2}$$

However, it is claimed in the main part of the lecture that this is equal to the uncertainty Σ given by the second derivative

$$\frac{1}{\Sigma^2} = \frac{1}{2}\left.\frac{d^2\chi^2}{d\theta^2}\right|_{\hat{\theta}}$$

The second derivative is

$$\frac{d^2\chi^2}{d\theta^2} = 2\sum_{i=1}^{N}\frac{b(x_i)^2}{\sigma_i^2} = \frac{2}{\sigma_{\hat{\theta}}^2}$$

and so $\Sigma = \sigma_{\hat{\theta}}$ as claimed. Effectively any function can be approximated to being linear in its parameter by doing a Taylor expansion around the estimate, so the above holds approximately in general, with the approximation being good for small uncertainties.

This means that for the chi-squared, and for the maximum likelihood in the Gaussian approximation, the uncertainty on a single parameter is given by either

$$\frac{1}{\sigma_{\hat{\theta}}^2} = \frac{1}{2}\left.\frac{d^2\chi^2}{d\theta^2}\right|_{\hat{\theta}} \qquad \text{or} \qquad \frac{1}{\sigma_{\hat{\theta}}^2} = -\left.\frac{d^2\ln(L)}{d\theta^2}\right|_{\hat{\theta}}$$

respectively.

27.2 Error matrix for more than one parameter

In the Gaussian approximation for more than one parameter, then we have to use the error matrix $\underline{\underline{E}}$, or more specifically the weight matrix $\underline{\underline{W}} = \underline{\underline{E}}^{-1}$ mentioned in the notes for Lecture 7. The generalisation of the above expressions is in terms of the elements of the weight matrix W_{ij}, where

$$W_{ij} = \frac{1}{2}\left.\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\right|_{\hat{\theta}} \qquad \text{or} \qquad W_{ij} = -\left.\frac{d^2\ln(L)}{\partial\theta_i\partial\theta_j}\right|_{\hat{\theta}}$$

Statistics of Measurement
Lecture 9 - Bayesian estimation

Heather Graven, 26 May 2023

28 Introduction

We saw in the previous lecture that a frequentist treatment of parameter estimates interprets them differently from random variables. As we will see in this lecture, the Bayesian approach allows us to handle these two cases in the same fashion. Sometimes this is at a cost of introducing a subjective element.

29 Bayesian probabilities

As mentioned briefly in Lecture 1, the Bayesian concept of a probability is subjective and depends on the available information. There is no need in the Bayesian framework to define multiple experiments which are repeatable. Hence, the Bayesian concept can be applied more widely.

In particular, in the Bayesian framework we can consider the parameters determined by our experiments as probabilistic variables. We are not saying e.g. that the true electron mass is not a precise value; what we mean is that the probability we assign to its possible values reflects the degree of our (lack of) knowledge. While this seems a minor change of philosophy, it means we can apply all the probability theory we have already learned for random variables directly to the parameter also. In particular, once we have the parameter probability distribution, we can find our 68.3% confidence interval (and hence uncertainties) just as we did for random variables.

In particular, we can apply Bayes' theorem of conditional probabilities to the parameters as well as the measurements. We saw Bayes' theorem in Lecture 1

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

where $P(x|y)$ is the probability for x given the value of y. Now, if we say x corresponds to the results from the experiment we have performed (the data) and y corresponds to the parameter θ, then this reads

$$P(\theta|\text{Data}) = \frac{P(\text{Data}|\theta)P(\theta)}{P(\text{Data})}$$

This allows us to calculate the conditional probability of the parameters given the data we have measured (the left-hand side) from the conditional probability of the data given the parameters (the numerator on the right-hand side). This term is just equivalent to the likelihood, which we met in Lecture 6. Multiplied by the likelihood is $P(\theta)$, the probability for the parameter, independent of the outcome of the experiment; i.e. before the experiment is done. (This is sometimes difficult to evaluate and this issue is discussed further below.) Finally, the denominator is the probability of seeing the data we measured, independent of any particular values of the parameters. This is sometimes called the 'evidence' but it effectively acts as a normalisation constant, i.e. for a discrete parameter

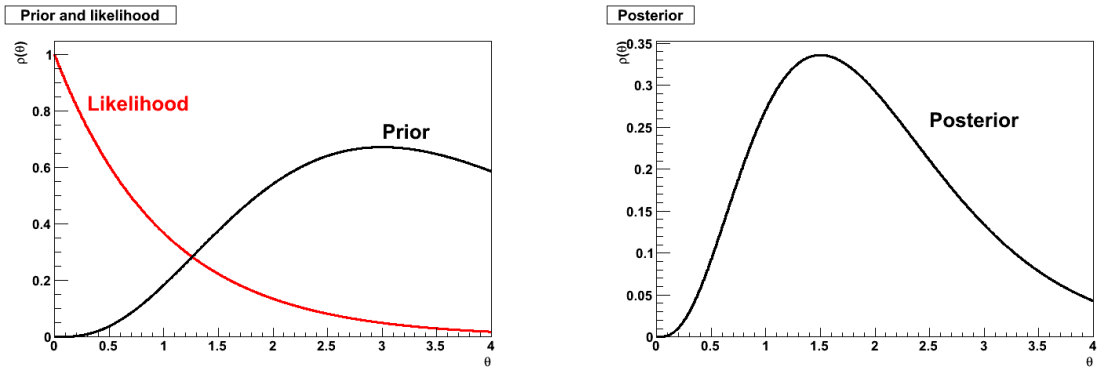
$$P(\text{Data}) = \sum_i P(\text{Data}, \theta_i) = \sum_i P(\text{Data}|\theta_i)P(\theta_i)$$

and for a continuous parameter

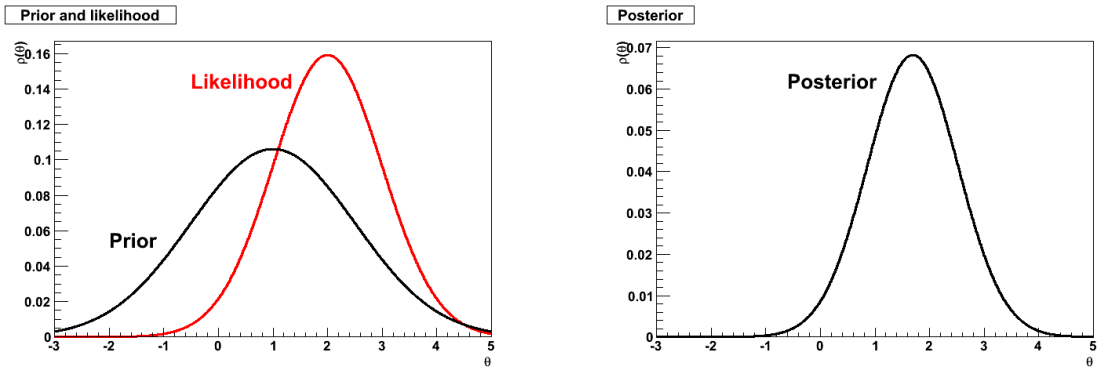
$$P(\text{Data}) = \int P(\text{Data}, \theta)\, d\theta = \int P(\text{Data}|\theta)P(\theta)\, d\theta$$

so it is simply the sum or integral of the numerator, which ensures that $P(\theta|\text{Data})$, the resulting probability distribution for θ, is normalised correctly. This property holds generally and is called the 'marginalisation rule' (also seen in Lecture 1).

You can think of the Bayesian equation as taking our existing knowledge of the parameter probability distribution $P(\theta)$, and updating it with an experiment to give a new probability distribution $P(\theta|\text{Data})$. The probability $P(\theta)$ therefore expresses our knowledge of the parameter before the experiment and hence is often called the 'prior', while the resulting probability $P(\theta|\text{Data})$ is called the 'posterior' and expresses our knowledge after the experiment. The likelihood can be thought of as updating our knowledge from the prior to give the posterior. A graphical example of the Bayesian equation is shown below.



A common case (due to the Central Limit Theorem) is that our prior knowledge (perhaps from a previous measurement) often has the form of a Gaussian distribution. The likelihood is often also Gaussian, as there is often a Gaussian distribution in the new data. Two Gaussian distributions multiplied together produce another Gaussian, so the posterior is also Gaussian distributed, as shown below.



The posterior is the probability distribution for the parameter after having seen the results of the experiment. If we did a second experiment later, then the posterior of the first experiment encapsulates the state of our knowledge about the parameter before the second experiment. Hence the first posterior becomes the second prior. This can be repeated many times.

Let's see an example. A friend of yours asks you to play a coin-tossing game where you bet on heads or tails, but he insists you use his coin and that he will bet heads. You suspect he will cheat and he has a double-headed coin. You play the game to determine if this is true or not (potentially at the cost of some money). The parameter to be determined is the number of heads on the coin, i.e. whether it is single-headed (SH) or double-headed (DH). You have to assign a prior probability for him cheating, i.e. for the coin being SH or DH. There is no 'right'

numerical answer; this is a subjective Bayesian prior and so it is up to your judgement. Let's take the initial probability you estimate for him to be cheating to be 20%, which means

$$P(\text{DH}) = 0.2 \qquad \text{so} \qquad P(\text{SH}) = 0.8$$

You play twice and he wins with two heads in a row. The probability of the coin being double-headed (DH) is given by

$$P(\text{DH}|\text{Two heads}) = \frac{P(\text{Two heads}|\text{DH})P(\text{DH})}{P(\text{Two heads})}$$

For the right-hand side, then the likelihood is

$$P(\text{Two heads}|\text{DH}) = 1$$

as a double-headed coin will always give two heads in two tosses. For the denominator, using the marginalisation rule

$$\begin{aligned} P(\text{Two heads}) &= P(\text{Two heads}|\text{DH})P(\text{DH}) + P(\text{Two heads}|\text{SH})P(\text{SH}) \\ &= 1 \times 0.2 + 0.25 \times 0.8 = 0.4 \end{aligned}$$

where the probability of getting two heads when using a normal (single-headed) coin is $P(\text{Two heads}|\text{SH}) = (1/2) \times (1/2) = 1/4$. Applying Bayes' theorem then gives

$$P(\text{DH}|\text{Two heads}) = \frac{P(\text{Two heads}|\text{DH})P(\text{DH})}{P(\text{Two heads})} = \frac{1 \times 0.2}{0.4} = 0.5$$

and so your initial guess (the prior) of 0.2 is modified to a posterior of 0.5 given the observed data, i.e. two heads. You get more suspicious that he is cheating.

You are not sure about your friend yet, so you decide to play for another two coin tosses, i.e. you repeat the experiment and (surprise) the result is another two heads. The prior for the second experiment is the posterior we just calculated. The likelihood is identical, so the evidence becomes

$$\begin{aligned} P(\text{Two heads}) &= P(\text{Two heads}|\text{DH})P(\text{DH}) + P(\text{Two heads}|\text{SH})P(\text{SH}) \\ &= 1 \times 0.5 + 0.25 \times 0.5 = 0.625 \end{aligned}$$

Applying Bayes' theorem then gives

$$P(\text{DH}|\text{Two heads}) = \frac{P(\text{Two heads}|\text{DH})P(\text{DH})}{P(\text{Two heads})} = \frac{1 \times 0.5}{0.625} = 0.8$$

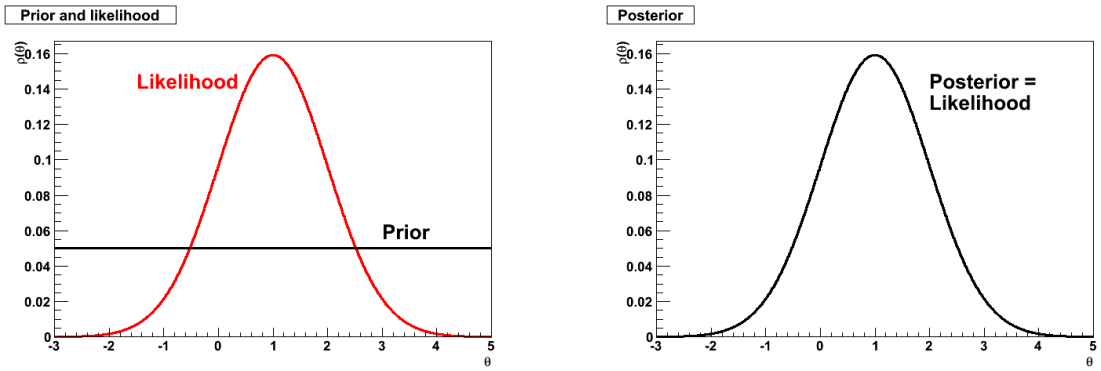
Now you definitely suspect him of cheating.

Note, if you calculated directly from your initial prior of $P(\text{DH}) = 0.2$, with the observed data of four heads (i.e. both experiments together), you would get exactly the same final answer.

30 Priors

It is with the prior that the problem with the Bayesian approach arises: we must have some knowledge of the parameter already. Sometimes this is OK; e.g. we may not be the first people to measure the electron mass. However, this is not always the case. In addition, it means we cannot give a result 'purely' from our own experiment to compare with other people's experiments.

If we can't, or don't want to, use previous results, then it is standard to use a 'flat' prior, i.e. a uniform distribution for $P(\theta)$. Here we just specify the upper and lower bounds, which



could be chosen conservatively to encompass a very large range. If we do choose a flat prior, then the posterior is the same shape as the likelihood.

Note that a flat prior might not always correspond to an unknown prior. In Lecture 3, we saw that a change in variables can change the shape of a distribution, so if we happened to use the square of the parameter instead, we would have to choose a different distribution in the parameter to get a flat distribution for the square of the parameter. Hence, the form of an unknown prior depends on the situation.

31 Relation to maximum likelihood

We discussed that the principle of maximum likelihood was a good way to get an estimate for a parameter. What does this estimation correspond to in terms of the Bayesian approach? As stated above, the likelihood appears directly in the expression we are using. Specifically

$$P(\theta|\text{Data}) \propto L(\theta)P(\theta)$$

where the likelihood has been evaluated for the particular experiment, as we did before. In particular, if we choose a flat prior so $P(\theta)$ is a constant, then

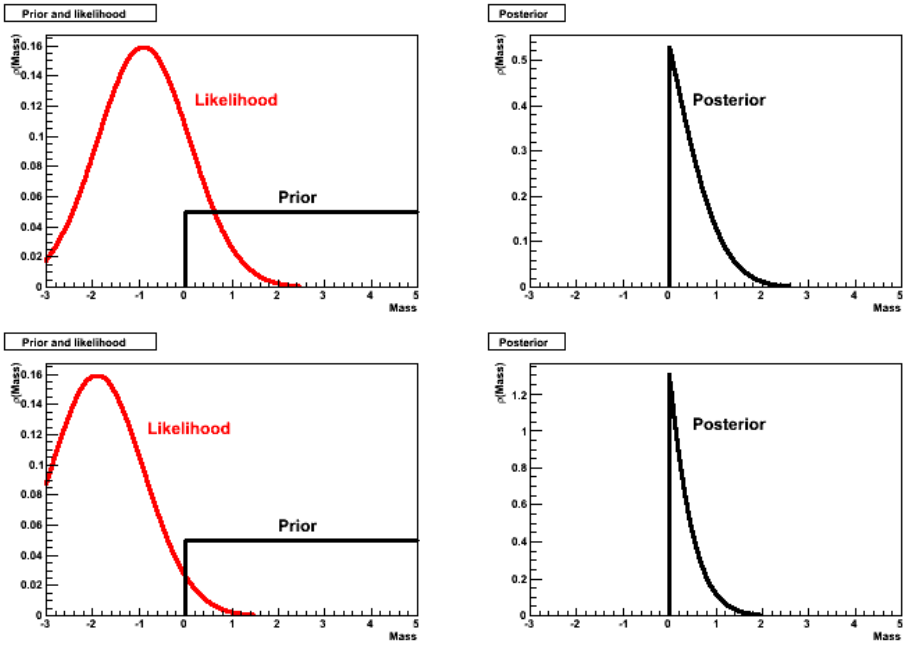
$$P(\theta|\text{Data}) \propto L(\theta)$$

Hence, the estimate obtained from the maximum likelihood approach corresponds to the most probable value of the parameter in the Bayesian approach when using a flat prior for the parameter. This in some sense 'justifies' the maximum likelihood principle, by making explicit which assumptions underpin it.

32 Physical constraints

We saw that frequentist methods have problems if they end up with results outside the physical region. We discussed an example of measuring the mass of a powder in the last lecture.

How does the Bayesian approach tackle this issue? The prior, whether flat or not, is the key; the probability of the mass genuinely being negative is of course zero, so that we simply make the prior be zero anywhere that is unphysical. This might leave only a small amount of the probability distribution in the physical region, but this is countered by the normalisation in the denominator (i.e. the evidence), which will also be small. Hence, it gives a sensible result overall, with the probability distribution for the parameter only being non-zero in the physical region. This is illustrated for the 99.1 g and 98.1 g mass measurements below (with a 100g weighing dish).



33 Frequentist or Bayesian?

This is a question which has vexed many people for a long time. Discussions can sometimes approach the intensity of religious wars; they are similar in that there is never any possibility of 'proving' one is better than the other.

Much of the time it doesn't matter. A Gaussian distribution well away from any physical constraints gives the same result in both frameworks. Since the Central Limit Theorem says many distributions look Gaussian, then in most cases the difference is minimal. Even when there is a difference, it is perfectly correct to use either method, as long as you state clearly, and give justification for, what you did.

34 Acknowledgment

Many thanks to Mark Richards and other previous lecturers for their development of the Statistics of Measurement course.