Statistical Methods for Experimental Physics
Part-I: Frequentist Statistics

Nicholas Wardle

October 2025

Contents

Introduction

These courses are supposed to give you a background to the kind of statistics issues that researchers in physics (and other scientific disciplines) face when analysing data. While I tend to use examples from my own field (particle physics), the fundamental concepts that we'll go through in this course are the same whether applied to data collected at the LHC or to satellite images of galaxy clusters (even if in practice they lead to very different approaches being used in the world of research).

These lectures are therefore not meant to be a complete course in statistics theory, but rather a more applied course focused on the common statistical methods used in data analysis. A lot of what we will deal with is "behind" the data analysis tools and methods that you'll be using in your practical course so hopefully this material will put that into context.

I'm basing a lot of this material on an excellent book by Frederick James, *"Statistical Methods in Experimental Physics: 2nd Edition,"* (see Figure 1 below). I strongly recommend this book for further reading on statistics, but there are of course more statistics books in the course reading list that you can find in the Library.

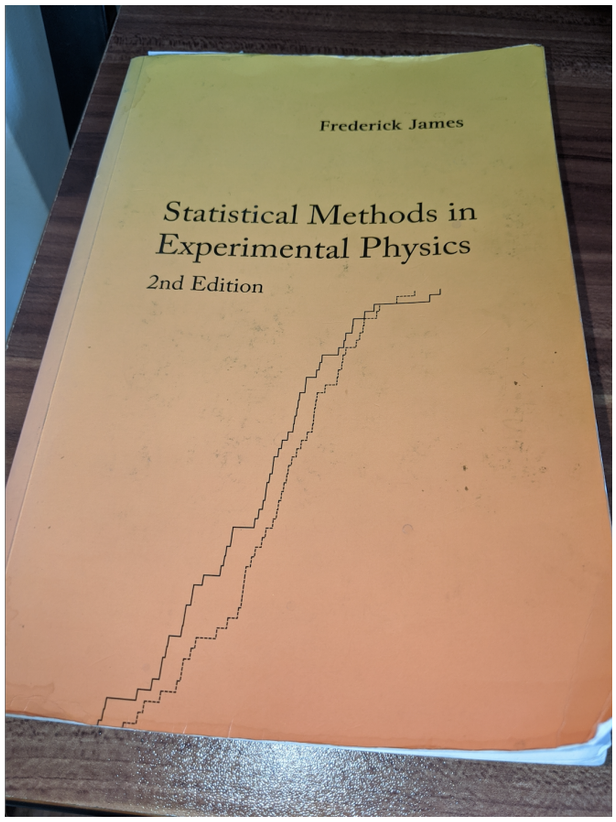

Figure 1: My beaten up copy of James, coffee stains and all.

While the theory behind the statistical methods we'll cover is vital to understand what we do and why we do it, the practicality of using those methods is equally important for successful research. These lectures will include some element of programming to give you an idea as to how to code up the methods for your own research.

I'll try to use very simple code without using too many fancy packages (not too much `Pandas`, `Keras` or `TensorFlow` here). You may find the examples I provide helpful when solving your problems, but feel free to use whatever you like concerning languages/packages to get the solutions. I'll be interested in seeing what you can come up with. A few examples of good languages for programming statistics are `R`, `Python`, `Julia` and `Mathematica`.

Whenever there is a snippet of code to explain something, you'll see a code box like the following.

```
# basic Python codeblock
nothing = [print("hello") for i in range(10)]

# or the for loop version
for i in range(10):
```

```
  print("hello")
```

I tend to use `python` when programming in statistics, but you should not let that discourage you from using another language if you prefer. We care about the *statistics* in the end. I personally like to use the *scientific python* (or `scipy`) package as I find it very intuitive and simple to use. You can find lots of helpful info on that package in the `scipy-lectures` website.

All of the code examples in these lectures can be found in full on the course page in Blackboard. You can follow along by downloading and running the notebooks. You might want to also download the '`environment.yml`' file to make setting up your python environment easier.

Use the code below to setup the right python environment and launch the jupyter notebooks – you need to make sure you have `conda` setup for this to work, which can be the same as the one you use for your practical course.

```
conda env create -f environment.yml
conda activate pystatistics
jupyter notebook
```

Which will open a web browser to run the notebooks. Once you have created the environment, you should be able to activate it in the future using.

```
conda activate pystatistics
```

By the end of this course, you should have an idea of what we mean by some basic concepts used to characterize data, like probability density, (co-)variance etc ... , and what terms like "significant result", and a "measured value of $X + \pm\sigma_X$" really mean. Furthermore, you should be able to calculate such things yourself using some simple programming tools.

We will go through the some of the notebooks during the lectures, but there will not be time to go through all of the code in detail. You should use time outside of the lectures to go through the notebooks and familiarise yourself with the material. At the end of each notebook, you will see a section called "Have a go yourself!" with some optional questions to help you understand the material in the notebook and get more practice with coding for statistics.

Please note that throughout this course, I tend to use the subscript i for various different things, e.g the index of a 3-element vector $\boldsymbol{X} = (X_i)_{i=1}^3$, or the index of a sequence of N outcomes (later this will be called "data") $\{X_i\}_{i=1}^N$ and so on. Hopefully the context makes it clear what I mean but please do ask if you are unsure.

Week 2 Probability

In this first lecture, we will review the basic concept of *probability*. You probably have a good idea already of what probability is, however the mathematical developed in 1933 by Kolmogorov treats probability as something which satisfies three axioms, and therefore no specific definition is given. The Kolmogorov axioms are,

- $P(X_i) \geq 0$ for all i
- $P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$
- $\sum_\Omega P(X_i) = 1$,

where Ω is the set of all possible, exclusive events X_i and $P(X_i)$ is the probability for X_i to occur.

We can see that probability should (probably) be at least a real number between 0 and 1, but otherwise we're free to choose the interpretation. For any meaningful statistics, we need a more practical definition and like all great constructs of society (democratic elections, quantum mechanics interpretations and "who shot first" in "Star wars IV: A new hope"), there are only two credible choices.

2.1 Frequentist probability

The concept of probability in the *frequensist* (or classical) paradigm is the one that you are most likely familiar with. As the name suggests, this definition of probability is related to the frequency with which an event occurs in repeated trials. More formally, imagine an experiment in which a series of events is observed and suppose that some (n) of these events can be labelled of type X. The probability that any single event out of a total number of N events will be of type X is given by the limit of the ratio,

$$P(X) = \lim_{N\to\infty} \frac{n}{N}. \tag{1}$$

Note that we inherit some limitations with this interpretation. This concept of probability can only be applied to repeated experiments, meaning the probability only makes sense when describing an ensemble of experiments. Furthermore, the conditions of these experiments must be essentially identical. As physicists, this shouldn't be an issue for you as good science requires reproducible results. On the other hand, this definition is very objective in the sense that if everyone agrees on the definition of X, no-one can object to the probability for an event to be labelled X - i.e the probability is independent of who is determining it.

2.2 Bayesian probability

The other commonly used definition of probability is *Bayesian probability*. This definition abandons the concept of frequency and instead defines probability as something which can be applied to non-repeatable experiments. The concept of Bayesian probability is based on the *degree of belief* in an event X occurring. Often, this is described in terms of the odds of winning a bet based on the outcome. Say for example that the odds on X occurring are 4:1, meaning if you place a bet of £1 that X occurs, you could either make a £4 profit or a £1 loss. If you are *willing to take those odds*, as a Bayesian, then you would ascribe a probability $P(X) = 1/(4+1)$ – the ratio between the amount you would bet to the total amount you stand to win (including the original bet) under those odds. Intuitively this makes sense, since $P(X)$ being very small means that you are almost certain X will not occur and therefore require very high odds before its worth betting any money. Instead $P(X) = 0.5$ means that you have have no strong belief about the outcome of X and therefore willing to accept odds of 1:1.

Note however that this definition of probability is a property of the observer and the system at the same time. It is also not a constant with time. As the observer gains more knowledge (perhaps by playing the game and placing a bet a few times), their value of $P(X)$ will change. This definition

of probability is therefore *subjective*. This is not necessarily a bad thing of course. Why shouldn't we update our believe about something after we gain knowledge and if you know more about how and when X will occur than I do, why should your probability be the same as mine? We'll see that this definition of probability has its advantages and at the end of the day, as physicists we want to use the definitions most convenient to the problem at hand.

2.3 Properties of probability

Let's quickly revisit the properties of any quantity that satisfies the Kolmogorov axioms. You would have seen these at school so there's nothing new here (hopefully). Let $P(A)$ be the probability that an exclusive event X_i occurs in A. Now consider two non-exclusive sets A and B of exclusive events X_i. Then the probability of an event occurring in either A, or B, or in both is given by,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B), \tag{2}$$

where $(A$ or $B)$ denotes the set of events X_i which belong to either set A or set B, or both. $(A$ and $B)$ denotes the set of events belonging to both.

Example: Let A be all even rolls of a six-sided die, and $\{X_i\}_1^6$ are the outcomes of a single roll of the die – i.e X_1 denotes that a 1 is rolled, X_2 denotes that a 2 is rolled etc. Now let B be all rolls which result in an outcome greater than or equal to 4. Then, according to Eqn. 2, $P(A \text{ or } B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{3} = \frac{2}{3}$. This matches what we'd expect given there are 4 outcomes that would be included in either of A or B, namely X_2, X_4, X_5 and X_6.

The other property you'll be familiar with is related to *conditional probability*, $P(A|B)$ – the probability of A given B. This is defined through the relationship,

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A). \tag{3}$$

We say that the sets A and B are *independent* if $P(A|B) = P(A)$. From the definition in Eqn. 3, we have that if A and B are independent, then,

$$P(A \text{ and } B) = P(A)P(B). \tag{4}$$

2.4 Bayes theorem

This definition brings us on to *Bayes theorem* for discrete events, which states that for sets A and B,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{5}$$

which follows from Eqn. 4. This theorem has important consequences for making decisions based on the probabilities of outcomes and can easily be forgotten when discussing them. For example, a simple pessimistic statement like "it always rains when I bring my umbrella", is easily checked with Bayes theorem. Let B be true whenever you take your umbrella with you and A be true when it's raining. Then the probability that it rains given you've brought your umbrella is the same as the probability you've brought your umbrella given that its raining multiplied by the ratio of probabilities that its raining to that of you bringing your umbrella. If it always rains where you live, then $P(A|B) = 1$ since $P(A) = 1$ and A and B are independent. However, if the probability for rain is very low, you better make sure you very rarely take your umbrella out to make such bold claim. Perhaps the most obvious but important consequence here is that the probability that it rains given you've taken your umbrella out is almost certainly not the same as the probability that you take your umbrella out given its raining – at least if you prefer to avoid getting wet!

There is another distinction which Bayesians make use of Bayes theorem and that is concerning statements about *hypotheses*. Hypothesis testing is a huge part of experimental physics (and all sciences), which we will cover in later lectures. For now, however its useful to bring up a distinction between random variables and hypotheses. To a Bayesian, $P(\theta_i)$ represents the degree of belief in the hypothesis $H(\theta = \theta_i)$, while for frequentists, θ_i is not a random variable and therefore a

probability cannot be assigned to it. To a Bayesian, the Bayes theorem can be directly applied to an experiment in which we have made a set of observations $\mathbf{X} = X_0, X_1, ...$ to assign probabilities to a given hypothesis θ_i,

$$P(\theta_i|\mathbf{X}) = P(\mathbf{X}|\theta_i)\frac{P(\theta_i)}{P(\mathbf{X})}. \tag{6}$$

$P(\theta_i|\mathbf{X})$ is called the *posterior probability* for hypothesis θ_i, given that we have observed $\mathbf{X}$ and $P(\mathbf{X}|\theta_i)$ is the probability to observe $\mathbf{X}$, if θ_i is true. These two are related by $P(\mathbf{X})$ which is the total probability to observe $\mathbf{X}$ for any hypothesis θ_i, which may or may not be known. Usually this can however be considered as a normalisation constant since summing over the LHS of Eqn. 6 for all i should yield 1 if all of the θ_i form a complete and exclusive set. Mathematically, if the hypotheses are exclusive and exhaustive (i.e they cover all possibilities), then we can calculate,

$$P(\mathbf{X}) = \sum_i P(\mathbf{X}|\theta_i)P(\theta_i) \tag{7}$$

Finally the quantity $P(\theta_i)$ is called the *prior probability* and represents the degree of belief in the different hypothesis *before* the experiment was conducted. Clearly this will depend on the individual and studying the effect changing priors is an important part of modern Bayesian statistics.

Example: The Monty Hall game show problem is a good example of using Bayes theorem to overcome our intuition. In case you don't know the problem, it goes like this. You are the contestant of a game show, presented with 3 boxes, labelled A, B and C and asked to choose one. You are told that one of the boxes contains a all-inclusive paid for trip to CERN (its an STFC funded game show) while the other two are empty. You pick box A but before opening it, the game show host opens one of the remaining two boxes that she *knows* doesn't contain the prize, revealing it to be empty. Given the choice of swapping your box A for the remaining box, do you stick with your original choice or switch? The answer may seem like the odds are 50:50 so it doesn't matter, but let's check.
Let $P(A)$, $P(B)$ and $P(C)$ represent the prior probabilities that the prize is inside box A, B and C, respectively. Sensible prior probabilities could be $P(A) = P(B) = P(C) = \frac{1}{3}$. We want to know the posterior probabilities, given that the host opens one of the other two boxes and its empty. Let H_B represent the case where the host opens box B and H_C the case where she opens box C. The probability that your box A contains the prize, given that the host opens box B is given by,

$$P(A|H_B) = \frac{P(H_B|A)P(A)}{P(H_B)} \tag{8}$$

The probability that the host opens B, given that you chose A is the same as the probability that the host opens box C in that case. The host knows neither box B or C contains the prize if the prize is in box A so can choose randomly. Therefore $P(H_B|A) = \frac{1}{2}$.
What about $P(H_B)$? From Eqn. 7, this must be the sum over the conditional probabilities $P(H_B) = P(H_B|A)P(A) + P(H_B|B)P(B) + P(H_B|C)P(C)$. The host will never pick the box that has the prize so $P(H_B|B) = 0$ and the host is forced to pick B if the prize is in C, given you've already picked A so $P(H_B|C)$=1. This means that $P(H_B) = \frac{1}{2}\cdot\frac{1}{3} + 0\cdot\frac{1}{3} + 1\cdot\frac{1}{3} = \frac{1}{2}$, and therefore

$$P(A|H_B) = \frac{1/2\cdot 1/3}{1/2}. \tag{9}$$

Instead, what about the other two cases where the prize is in one of the boxes you didn't pick? Again, using Bayes theorem,

$$P(C|H_B) = \frac{P(H_B|C)P(C)}{P(H_B)} = \frac{1\cdot 1/3}{1/2} = \frac{2}{3} \tag{10}$$

So we can see that if you pick A and the host opens box B, the probability that the prize is in box C is higher than the probability that its in box A, so you should switch!. The exact same arguments holds if instead the host opens box C and of course cycling around A, B and C doesn't change the problem so its *always* better to switch in this game show!

We can check our reasoning by simulating the game show multiple times employing the two strategies (stick with your original choice of box or swap). For this simulation (and quite often in these lectures), we need to use (pseudo) random generation. I fine the `random` module from `python`'s

`scipy` package to be very easy to use. Something like the snippet below can be used to choose a random element of a list (here just the boxes "A", "B" and "C").

```
import numpy
prize_boxes = ["A","B","C"]
chosen_box = numpy.random.choice(prize_boxes)
```

Have a look at the **MontyHall.ipynb** to see how the simulation compares to our calculation.

In the above example, we calculated a collection of $P(\theta_i|\mathbf{X})$ for all i, where $i = A, B, C$ and $\mathbf{X} = H_B$. We were calculating the posterior probability *distribution*. We also knew from the start that $P(\theta_i) = \frac{1}{3}$ for all i, which was our prior probability *distribution*. However, it may have been that we had a firmer belief, before the game show, that the prize is in a particular box (maybe the host was hovering over one of them more than the others). In general the this distribution can have a large impact on statistical (and experimental) results when using Bayesian statistics.

We've already gotten a bit ahead of ourselves since we've talked about hypotheses and distributions of a random variable/hypotheses are without first defining what a probability distribution is. We'll cover this in the next section.

2.5 Probability distributions

Random variables (as opposed to weather scenarios and umbrellas) are common place in experimental physics. For example. fundamental particle interactions (be it proton collisions at the LHC or photons interacting with a scintillator) are by their very nature probabilistic. Similarly, measurements of species populations or traffic flow will be distributed according to some underlying distribution of possible values. We should therefore expect that whether we are studying cosmic rays, nuclear decays or particle collisions, we will regularly encounter random variables (energies, momenta, number of events). Note that a random variable X does not need to be a single value, but may instead be multiple quantities – in particle physics for example, we tend to refer to this collection of quantities as an "event". For a sequence of events $X_1, X_2, X_3...$, we will often use vector notation $\mathbf{X}$. Note however that $\mathbf{X}$ can also refer to a repeated set of experiments and each experiment will have a single observation (e.g the total number of recorded nuclear decays in a given time window). Hopefully, it will be clear in the context as to which we mean, but the mathematics will be the same.

The corresponding probabilities $P(X)_\Omega$ over all possible values of $X \in \Omega$ form a *probability distribution*. As an example, if X is the outcome of a single die roll, then $\Omega = \{X = 1, X = 2, X = 3, X = 4, X = 5\}$ and $P(X) = \frac{1}{6}$ for every possible value of X, meaning the probability distribution is uniform. Note that of course we must have that,

$$\sum_\Omega P(X) = 1 \tag{11}$$

A single event (roll) can be then thought of a random draw from such a probability distribution, and successive rolls of the die will yield a distribution of values whose frequency converges to a uniform distribution (U). We will write this as $X \sim U(1, 6)$.

We'll look at a number of common probability distributions that you'll come across in experimental data analysis. You can take a look through the **CommonProbabilityDistributions.ipynb** notebook for examples of using these distributions as they are implemented in the `python` package `scipy.stats`.

2.5.1 The Binomial distribution

The binomial distribution describes the distribution of the number of successes (k) in a sequence of n independent trials, where the probability of success in any trial is p. More generally, any sequence of experiments, each of which results in a yes/no, 1/0 or other binary result with probabilities p and

$q = (1 - p)$ assigned to each outcome will be distributed according to the binomial distribution,

$$f(k; n, p) = \binom{n}{k} p^k q^{n-k}, \tag{12}$$

where k (the number of "yes", "1" etc...) $= 0, 1, 2...n$ and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

For any probability distribution in a single dimension, we can define the *cumulative* distribution as,

$$F(k; n, p) = \sum_{i=0}^{k} f(k; n, p) = \sum_{i=0}^{k} \binom{n}{k} p^k q^{n-k}, \tag{13}$$

Figure 2 shows the binomial distribution ($B(n, p)$) for several values of the parameters n and p. The binomial distribution in python is available using something like the code below;

```
from scipy.stats import binom

# Probability at k=4 for n=10,p=0.5
binom.pmf(4,10,0.5)

# Cumulative probability at k=4 for n=10,p=0.5
binom.cdf(4,10,0.5)
```

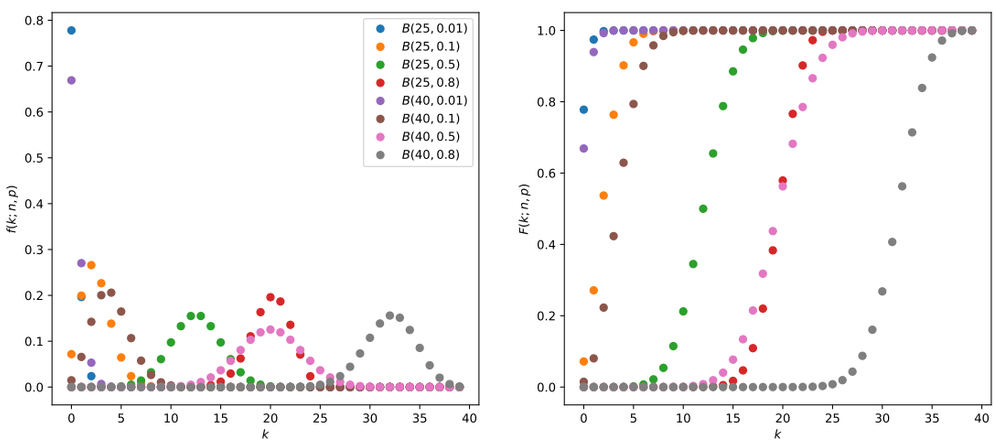

Figure 2: Binomial probability distribution $f(k; n, p)$ (left) for different values of the parameters n and p, and their cumulative distribution fuctions (right).

Note that I've introduced f here instead of P for the probability, and separated the random variable k from the parameters n, p with a semicolon ;. You may sometimes see a | in place of the ;, but in these lectures, I will only use this to mean "*given that*" – i.e for conditional probabilities (or related quantities). We'll use f whenever discussing a probability distribution (which will be more important to distinguish later when we get to *probability densities*.

2.5.2 The Poisson distribution

In experimental physics, we often deal with counting events that are very "rare". For example, the LUX experiment is designed to search for the presence of dark matter by counting the the number of events in which a dark matter particle interacts with a Xenon nucleus from a vast amount of liquid Xenon, which is shielded from other radiation sources. The number of dark matter particles passing through the earth at any given moment is expected to be relatively large, however the probability that one of these particles interacts is extremely small. In this circumstance, we can use a limit case of the binomial distribution where $n \to \infty$, and $p \to 0$ such that the product $\lambda = np$ is constant.

Let's re-write Eqn. 13 substituting $p = \frac{\lambda}{n}$,

$$f(k;n,p) = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \tag{14}$$

$$= \frac{\lambda^k}{k!}\frac{n!}{(n-k)!}\left(\frac{1}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \tag{15}$$

Now lets look at the middle terms in the product,

$$\frac{n!}{(n-k)!}\left(\frac{1}{n}\right)^k = \frac{n(n-1)(n-2)...(n-k+1)(n-k)(n-k-1)...(2)}{n^k(n-k)(n-k-1)...(2)} \tag{16}$$

$$= \frac{n}{n^k}(n-1)(n-2)...(n-k+1) \tag{17}$$

$$= \frac{\cancel{n^k}}{\cancel{n^k}}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{k+1}{n}\right) = A \tag{18}$$

Now lets look at the last term,

$$\left(1-\frac{\lambda}{n}\right)^{n-k} = \left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-k} = B \tag{19}$$

Now lets look at what happens when $n \to \infty$. The second term in B will tend to 1 since λ is constant. The first term, is the usual expression for the exponential function,

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n, \tag{20}$$

so $B \to e^{-\lambda}$ as $n \to \infty$. It should be easy to see also that $A \to 1$ as $n \to \infty$, so that finally,

$$f(k;n,p) \to f(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda} \tag{21}$$

as $n \to \infty$, which is the Poisson distribution. The Poisson distribution in python is available using something like the code below;

```
from scipy.stats import poisson

# Probability at k=3 for lambda=2
poisson.pmf(3,2)
```

Fig. 3 shows the Poisson distribution for different values of the parameter λ. You can see that as λ increases, the distribution looks more symmetric around its mode and that this mode gets closer to the value of λ. This is an important feature of the Poisson distribution that we'll come to later on in the lectures. In the Poisson distribution, our random variable $X = k$ and we write that k is distributed as a Poisson distribution with parameter lambda $k \sim \text{Poisson}(\lambda)$ (just to avoid writing $f(k;\lambda)$ all the time).

We can generalize probabilities (and distributions of probability) for events to probability distributions of *continuous* random variables. A continuous random variable is one whose possible values cover continuous intervals, for example $X \in \mathbb{R}$. In this case, we need to define something called a *probability density function*.

Suppose we measure the energy E of cosmic ray particles. Due to the underlying quantum mechanics of interactions with the atmosphere, there will be a probability that the particle has a particular energy within some finite range $P(E \in (E+\delta E)$. As the energy is a continuous variable, we are free to choose δE as small as we like. If there is a continuous function for all values of $E > 0$ which represents the limit as $\delta E \to 0$,

$$f(E) = \lim_{\delta E\to 0}\frac{P(E \in (E+\delta E))}{\delta E} \tag{22}$$

we call that function the *probability density function* or p.d.f. The p.d.f must be normalised analogously to the discrete case such that,

$$\int_\Omega f(E)dE = 1, \text{ where } E \in \Omega. \tag{23}$$

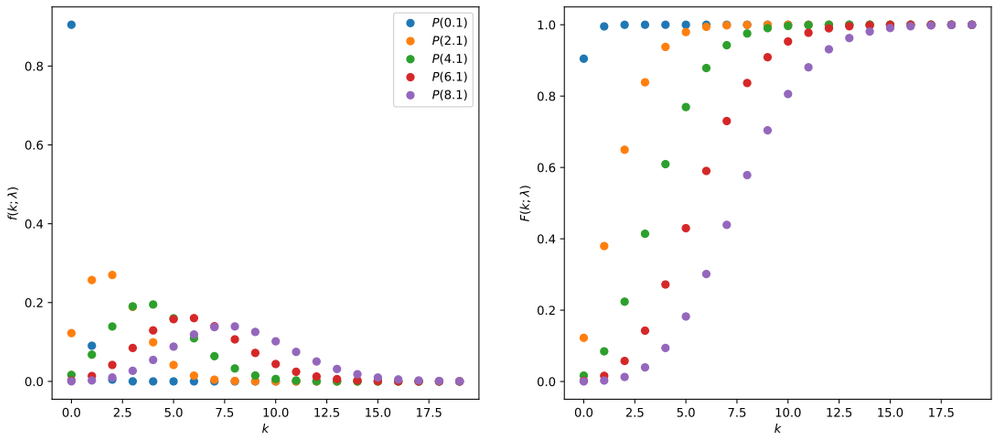

Figure 3: Poisson probability distribution $f(k;\lambda)$ (left) for different values of the parameter λ, and their cumulative distribution fuctions (right).

Of course, as with the discrete case, we can measure multiple continuous quantities $\mathbf{X}$ in which case $d\mathbf{X} = dX\,dY\,dZ, \ldots$ represents an infinitesimal volume.

We can define two more distributions related to the probability density. The first is the *cumulative distribution*, defined by the integral,

$$F(X) = \int_{X_a}^{X} f(X')dX' \tag{24}$$

where $X_a \leq X \leq X_b$, and $f(X)$ is the probability density distribution of X. Clearly then $F(X_a) = 0$ and $F(X_b) = 1$. We can interpret $F(X_1)$ as the probability that $X \leq X_1$. The cumulative distribution will be used later when we discuss p-values.

If the probability density function has more than one random variable, we can define a projection onto a fewer number of variables, which is called the *marginal density*. For example, if we have a probability density $f(X,Y)$, the projection onto the X variable is given by,

$$g(X) = \int_{Y_a}^{Y_b} f(X,Y)dY \tag{25}$$

where $Y_a \leq Y \leq Y_b$.

We can also use Bayes theorem for continuous random variables. If $f(X,Y)$ is the probability density for two variable X and Y, with marginal densities $g(X)$ and $h(Y)$, then Eqn. 5 becomes,

$$q(Y|X) = \frac{p(X|Y)h(Y)}{g(X)}, \tag{26}$$

where $q(Y|X)$ and $p(X|Y)$ are the conditional probabilities defined through the relationship

$$f(X,Y) = p(X|Y)h(Y) = q(Y|X)g(X). \tag{27}$$

In the next two sections, we'll take a look at two very common probability densities of continuous random variables.

2.5.3 Gaussian distribution

Probably (no pun intended), the Gaussian (or Normal) distribution is the most common probability distribution that you'll come across and use in your research. There is a good reason for this that we'll come to in couple of sections but for now, let's just refresh our memories about the Gaussian distribution.

If a continuous random variable X is distributed as a Normal distribution $\phi(\mu,\sigma)$, then ;

$$f(X;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2},$$

where μ and σ are the two parameters of the Gaussian probability denisity. In `scipy.stats` we can obtain the pdf and cdf of a Gaussian distribution using `scipy.stats.norm`, as for example;

```
from scipy.stats import norm

# Gaussian probability at x=5, for parameters mu=8, sigma=2
norm.pdf(5,8,2)
```

Figure 4 shows Gaussian probability density and cumulative distribution functions for different values of the parameters μ and σ.

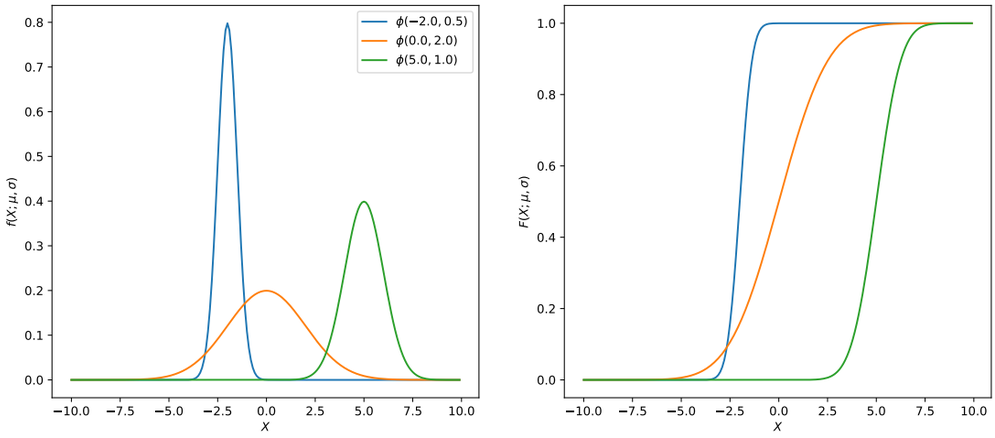

Figure 4: Gaussian probability density $f(X;\mu,\sigma)$ (left) for different values of the parameters μ and σ, and their cumulative distribution fuctions (right).

Multi-variate Gaussian distribution: In your research, you'll be dealing with datasets with many observables. While we don't expect each of these observables to be distributed as a Gaussian, the properties of multi-variate Gaussians will be useful when looking at properties of multivariable datasets and likelihood surfaces for multiple parameter estimations.

The multivariate Gaussian probability density can be written as,

$$f(\boldsymbol{X};\boldsymbol{\mu},\mathbf{V}) = \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{V})}} e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T(\mathrm{V}^{-1})(\boldsymbol{X}-\boldsymbol{\mu})} \tag{28}$$

Where $\boldsymbol{X} = (X_1, X_2, X_3, ..., X_N)$ is a vector of random variables (not to be confused with a sequence of random outcomes of X), and $\mathbf{V}$ is a $N \times N$ matrix of *co-variances* $V_{ij} = \mathrm{covariance}(X_i, X_j)$. We'll talk about the co-variance of two random variabes later, so don't worry about it for now.

In the special case where $N = 2$, we can write,

$$f(X,Y;\mu_X,\mu_Y,\sigma_X,\sigma_Y,\rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2 + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]\right) \tag{29}$$

where X and Y are the two random variabes and $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ and ρ are the parameters of the probability density function.

2.5.4 χ^2 distribution

The χ^2 distribution is related to the normal distribution. If $T \sim \phi(T; 0, 1)$ then $X = T^2$ will be distributed under the following probability density function,

$$f(X; 1) = \frac{1}{\sqrt{2\pi X}} e^{-\frac{X}{2}}$$

This distribution is known as a $\chi^2(1)$ – pronounced *chi-square* – distribution with 1 degree of freedom. In general, we can define χ^2 distributions with larger distributions – $\chi^2(n)$. The `scipy.stats.chi2` module can be used to calculate the probability density for any number of degrees of freedom,

```
from scipy.stats import chi2

# Chi2 probability at x=5, for n=3
chi2.pdf(5,3)
```

Figure 5 shows χ^2 probability density and cumulative distribution functions for different numbers of degrees of freedom n. We'll see later that the χ^2 distribution is very often seen in *hypothesis testing*.

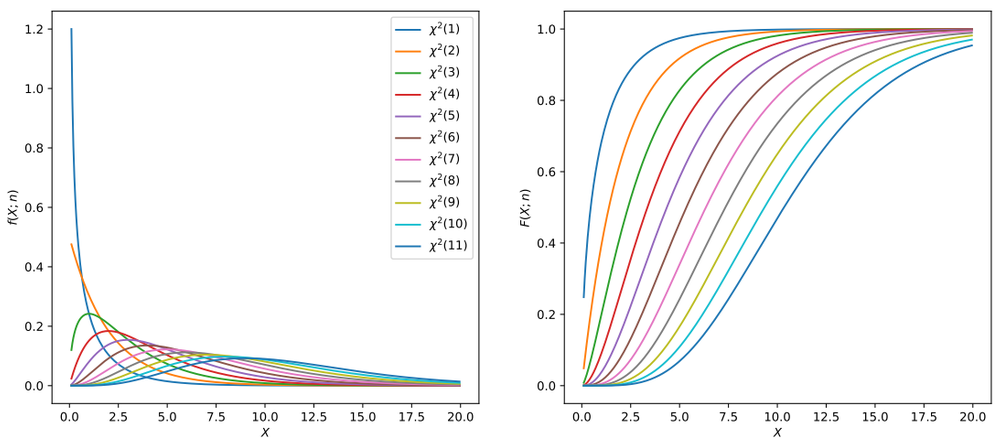

Figure 5: χ^2 probability density $f(X; n)$ (left) for different values of the parameter n, and their cumulative distribution fuctions (right).

2.6 Generating random variables

In statistics, we often need to generate *psuedo-data* (or "toys") in which we replicate hypothetical observations that we could have made in an experiment, to approximate moments distribtuions, calculate confidence intervals and determine marginalised posterior distributions. To do this we make use of random number generators and calculate quantities from the resulting random variables.

This approach to generating distribution of random variables is known as a 'Monte Carlo' approach and will be extremely useful as we'll see in these lectures. You'll learn much more about Monte Carlo methods and sampling in the second half of these lectures, but for now just know that its possible, using packages like `scipy.stats` and `numpy` to generate random variables according to some probability density.

Below is a snippet of code using the `numpy.random.poisson` function, which generates a vector of random integers ("toys") according to the Poisson distribution, for a given value of λ. By generating a large number of them we can estimate the frequency (i.e the probability) to get a given integer, and plot what that frequency is for each integer.

```
import numpy
import matplotlib.pyplot as plt
plt.rcParams.update({'font.size': 14})

# 5 different values of \lambda
lambdas = numpy.arange(0.1,10,2.0)
colors  = ["red","crimson","purple",
           "mediumslateblue","blue"]

# generate the distributions with MC
for lamb,col in zip(lambdas,colors):
 poisson_distribution = numpy.random.poisson(lamb,size=10000)
 plt.hist(poisson_distribution,bins=20,range=(0,20),\
    density=True,color=col,linewidth=2.\
    ,histtype='step',label="$\lambda=$%.1f"%lamb)

plt.xlabel("$k$")
plt.ylabel("$f(k,\lambda$)")
plt.legend()
plt.show()
```

The result of this code is shown in Figure 6.

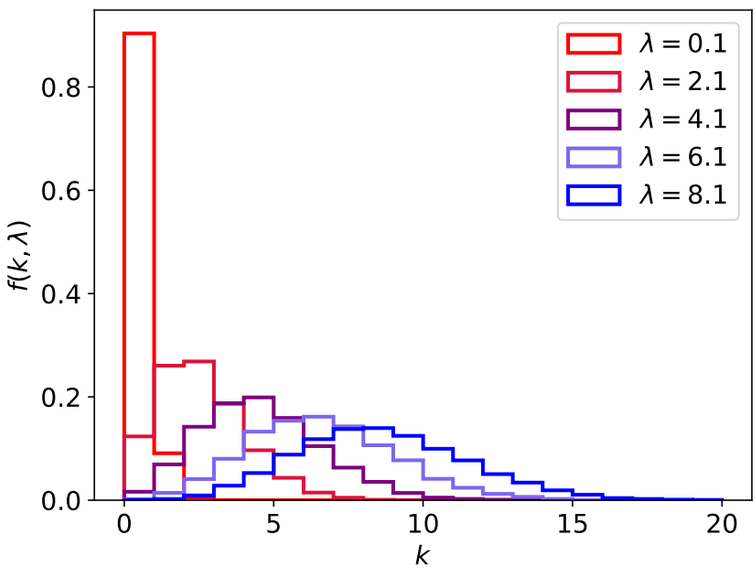

Figure 6: Poisson probability distribution $f(k;\lambda)$ for different values of the parameter λ. The distributions are produced by randomly generating 10,000 values (toys) of k according to a Poisson distribution with parameter k, and creating the histogram of the results.

As we'll see later, generating more toys provides a more accurate estimate of the probabilities.

2.7 Moments of probability distributions

Probability density distributions can be used to obtain useful information about a particular random variable, or functions of random variables. For example, the mean value of a random variable X (or its *expectation value* under $f(X)$ is given by,

$$E[X] = \int_{\Omega} X f(X) dX, \tag{30}$$

where $E[\cdot]$ is the expectation operator. Similarly, any function of X, $g(X)$ has an expectation value $E[g]$ under $f(X)$ of,

$$E[g] = \int_\Omega g(X)f(X)dX \tag{31}$$

It should be obvious that the expectation is a *linear* operator, since,

$$E[a \cdot g(X) + b \cdot h(Y)] = a \cdot E[g(X)] + b \cdot E[h(Y)] \tag{32}$$

If the two variables X and Y are *independent* then we also have that $E[X \cdot Y] = E[X]E[Y]$.

The expectation value (mean) is often referred to as the *first moment* of the distribution of X, but of course we can define higher moments too. For example, the expectation of the function $g(X) = (X - E[X])^2$ is called the *variance* of X under $f(X)$,

$$V(X) = E[(X - E[X])^2] = \int_\Omega (X - E[X])^2 f(X)dX \tag{33}$$

You will also have come across the term *standard deviation*, $\sigma = \sqrt{V(X)}$. Note that the variance (or the standard deviation) is *not* a linear operator. However, we do have the following properties,

$$V(X + a) = V(X) \text{ and } V(aX) = a^2V(X), \tag{34}$$

when a is a constant value.

There are equivalents of the expectation and variance for discrete random variables too, $E[X] = \bar{X} = \sum_\Omega Xf(X)$, and $V(X) = \sum_\Omega (X - E[X])^2 f(X)$.

Example: The expectation and variance of a discrete random variable k distributed as a Poisson distribution can be calculated using the discrete versions of the formula given in Eqns. 30 and 33.

$$\begin{aligned} E[k] &= \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = e^{-\lambda}\lambda e^{\lambda} && (35)\\ &= \lambda, && (36) \end{aligned}$$

and,

$$\begin{aligned} V(k) &= \sum_{k=0}^{\infty}(k-\lambda)^2 e^{-\lambda}\frac{\lambda^k}{k!} && (37)\\ &= e^{-\lambda}\sum_{k=0}^{\infty} k^2\frac{\lambda^k}{k!} - 2\lambda e^{-\lambda}\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} + \lambda^2 e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} && (38)\\ &= \lambda + \lambda^2 - 2\lambda^2 + \lambda^2 && (39)\\ &= \lambda && (40) \end{aligned}$$

using the identities $(x^2 + x)e^x = \sum_{k=0}^{\infty} k^2\frac{x^k}{k!}$, $xe^x = \sum_{k=0}^{\infty} k\frac{x^k}{k!}$, and $e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}$. So for $k \sim \text{Poisson}(\lambda)$, we have $E[k] = V(k) = \lambda$.

Of course, there are other moments we can calculate which describe more properties of the probability distribution we're interested in. Specifically,

$$\mu_n = E[X^n] \text{ is the } n^{\text{th}} \text{ algebraic moment,} \tag{41}$$

and,

$$\nu_n = E[(X - E[X])^n] \text{ is the } n^{\text{th}} \text{ central moment.} \tag{42}$$

Then we have that $E[X] = \mu_1$ and $V(X) = \nu_2$. Often you'll see quantities which are constructed from these moments. For example, the *skewness* of a probability distribution is given by,

$$\gamma = \frac{\nu_3}{(\nu_2)^{3/2}}, \tag{43}$$

which describes an asymmetry in the distribution. We can also determine moments of multidimensional distributions too. For example for a probability distribution $f(X, Y)$, the algebraic moment

of order m in X and n in Y is $\mu_{mn} = E[X^m Y^n]$. The most common moment you'll come across is the *co-variance*, which for an M dimensional probability distribution $f(\mathbf{X})$, is the central moment of order 1 in two variables X_i, X_j and order 0 for the remaining $M-2$ variables,

$$\text{covariance}(X_i, X_j) = \nu^{ij}_{1,1} = E[(X_i - E[X_i])(X_j - E[X_j])], \tag{44}$$

and the *correlation coefficient* is given by

$$\text{correlation}(X_i, X_j) = \rho(X_i, X_j) = \frac{\nu^{ij}_{1,1}}{\sqrt{\nu^i_2 \nu^j_2}}, \tag{45}$$

where ν^i_2, and ν^j_2 are the variance of X_i and X_j respectively. The correlation coefficient will be $-1 \leq \rho \leq 1$. Two variables X, Y which are independent under $f(X, Y)$ will have a correlation coefficient of 0. The reverse is however not necessarily true so you should remember that *independence* is a stronger statement than *uncorrelated* (see problems for an example).

You can calculate the moments of common probability distributions with the `scipy.stats` package using the examples below,

```
from scipy.stats import binom,poisson,norm

mean, var, skew = binom.stats(15, 0.4, moments='mvs')
print(mean,var,skew)

mean, var, skew = poisson.stats(2, moments='mvs')
print(mean,var,skew)

mean, var, skew = norm.stats(3, 2, moments='mvs')
print(mean,var,skew)
```

Sample moments At this point, it's worth clarifying that moments of probability distributions are not the same as sample moments. For a finite sample of a random variable, it is always possible to determine sample moments (unlike in the case of some probability densities).

For a sequence of a random variable $(X_1, X_2, ..., X_N)$ of size N, we can define the $n-$th sample moment as,

$$m_n = \frac{1}{N} \sum_i^N X_i^n.$$

You'll be familiar with the 1st such moment, which is the sample mean,

$$m_1 = \bar{X} = \frac{1}{N} \sum_i^N X_i,$$

and the second *central* moment, which is the sample variance,

$$\bar{V} = \frac{1}{N} \sum_i^N (X_i - \bar{X})^2.$$

We can also define the sample covariance of two random variables X and Y as,

$$\text{covariance}(X, Y) = \frac{1}{N} \sum_i^N (X_i - \bar{X})(Y_i - \bar{Y}).$$

We'll see later that the sample moments can be used to esimtate moments of probability distributions (if they exist) and more importantly, their parameters. For now however, just be aware that the sample moments are not the same as the moments of probability distributions.

2.8 Compositions of probability distributions

Compositions of random variables can be used to determine the distributions of derived quantities. We haven't covered what we mean by a "measurement" yet but you'll know that any measurement of a physical quantity is meaningless without an associated "uncertainty". Actually, uncertainty in the statistics sense won't be covered until the last sections of these lectures. However, you are probably aware of the concept at least in relation to the variance of a random variable. Often you'll have come across the idea of uncertainty being associated to the standard deviation – $\sigma_X = \sqrt{V(X)}$. Quite correctly, if X is a measured quantity, more often than not, it will be a random variable and we'll see that the standard deviation of this variable can be a pretty good approximation to a more concrete definition of uncertainty. You'll also have come across the concept of propagating uncertainties for quantities built from random variables such as the sum of two random variables X, Y. We already know that $E[X+Y] = E[X] + E[Y]$ by the linearity of expectation, but for the variance, the formula you know is,

$$V(X+Y) = V(X) + V(Y). \tag{46}$$

We'll see where this comes from.

Suppose $X \sim f(X)$ and $Y \sim g(Y)$ are two independent random variables and $U = X + Y$. What is the probability distribution $p(U)$ of U?. First, we change variables to,

$$U = X + Y, \; V = X. \tag{47}$$

The probability distribution $h(U,V)$ is defined by,

$$h(U,V)dUdV = f(X)g(Y)dXdY, \tag{48}$$

since the probability (not necessarily the density however) must be invariant under a change of variables. for a change of variables, the infinitesimal volume $dXdY$ changes as,

$$dXdY = \left| \det \begin{bmatrix} \frac{\partial U}{\partial X} & \frac{\partial U}{\partial Y} \\ \frac{\partial V}{\partial X} & \frac{\partial V}{\partial Y} \end{bmatrix} \right|^{-1} dUdV = dUdV \tag{49}$$

then we have

$$h(U,V)dUdV = f(V)g(U-V)dUdV \implies h(U,V) = f(V)g(U-V). \tag{50}$$

Recall that to obtain the marginal density $p(U)$, we integrate over V (see Eqn. 25) to obtain,

$$p(U) = \int_{-\infty}^{+\infty} h(U,V)dV = \int_{-\infty}^{+\infty} f(V)g(U-V)dV. \tag{51}$$

Note that we have relied on the fact that the change of variables $X, Y \to U, V$ is a *bijection* (1 to 1 mapping). If it were not the case, then we could split the integral into segments over which each transformation is a bijection.

Example: In the case that $X \sim \phi(X; \mu_X, \sigma_X)$ and $Y \sim \phi(Y; \mu_Y, \sigma_Y)$, where

$$\phi(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{52}$$

is the *Gaussian* (or normal) distribution, we can find the probability distribution of $U = X+Y$. Plugging $\phi(\cdot)$ into Eqn. 51 we find,

$$\begin{aligned}
p(U) &= \int_{-\infty}^{+\infty} \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left[-\frac{(V-\mu_X)^2}{2\sigma_X^2}\right] \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left[-\frac{(U-V-\mu_Y)^2}{2\sigma_Y^2}\right] dV && (53)\\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sqrt{2\pi}\sigma_X\sigma_Y} \exp\left[-\frac{(V-\mu_X)^2}{2\sigma_X^2} - \frac{(U-V-\mu_Y)^2}{2\sigma_Y^2}\right] dV && (54)\\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sqrt{2\pi}\sigma_X\sigma_Y} \exp\left[-\frac{\sigma_Y^2(V-\mu_X)^2 + \sigma_X^2(U-V-\mu_Y)^2}{2\sigma_X^2\sigma_Y^2}\right] dV && (55)\\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sqrt{2\pi}\sigma_X\sigma_Y} \exp\left[A\right] dV && (56)\\
& && (57)
\end{aligned}$$

Expanding the exponent,

$$\begin{aligned}
A &= \frac{-(\sigma_X^2+\sigma_Y^2)V^2 + 2\left(\sigma_X^2(U-\mu_Y) + \sigma_Y^2\mu_X\right)V - \sigma_X^2(U^2+\mu_Y^2-2U\mu_Y) - \sigma_Y^2\mu_X^2}{2\sigma_X^2\sigma_Y^2} && (58)\\
&= \frac{-\sigma_U^2V^2 + 2\left(\sigma_X^2(U-\mu_Y) + \sigma_Y^2\mu_X\right)V - \sigma_X^2(U^2+\mu_Y^2-2U\mu_Y) - \sigma_Y^2\mu_X^2}{2\sigma_X^2\sigma_Y^2} && (59)\\
&= \frac{-V^2 + 2\left(\sigma_X^2(U-\mu_Y) + \sigma_Y^2\mu_X\right)\frac{V}{\sigma_U^2} - \frac{\sigma_X^2}{\sigma_U^2}(U-\mu_Y)^2 - \frac{\sigma_Y^2\mu_X^2}{\sigma_U^2}}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_U}\right)^2}, && (60)
\end{aligned}$$

where we have defined $\sigma_U^2 = \sigma_X^2 + \sigma_Y^2$. Now we can complete the square for V, to get,

$$A = \frac{-\left(V - \frac{\sigma_X^2(U-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_U^2}\right)^2 + \left(\frac{\sigma_X^2(U-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_U^2}\right)^2 - \frac{\sigma_X^2(U-\mu_Y)^2+\sigma_Y^2\mu_X^2}{\sigma_U^2}}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_U}\right)^2}. \tag{61}$$

If we plug A back into the exponent we get,

$$\begin{aligned}
\exp[\mathrm{A}] &= \exp\left[\frac{\left(\frac{\sigma_X^2(U-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_U^2}\right)^2 - \frac{\sigma_X^2(U-\mu_Y)^2+\sigma_Y^2\mu_X^2}{\sigma_U^2}}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_U}\right)^2}\right] \times && (62)\\
&\quad \exp\left[\frac{-\left(V - \frac{\sigma_X^2(U-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_U^2}\right)^2}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_U}\right)^2}\right] && (63)
\end{aligned}$$

The first exponential is constant in V, so it can come outside of the integral. Furthermore, for the normalisation term, we have,

$$\frac{1}{\sqrt{2\pi}\sqrt{2\pi}\sigma_X\sigma_Y} = \frac{1}{\sqrt{2\pi}\sigma_U}\frac{\sigma_U}{\sqrt{2\pi}\sigma_X\sigma_Y} \tag{64}$$

so then,

$$\begin{aligned}
p(U) &= \frac{1}{\sqrt{2\pi}\sigma_U} \exp\left[\frac{\left(\frac{\sigma_X^2(U-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_U^2}\right)^2 - \frac{\sigma_X^2(U-\mu_Y)^2+\sigma_Y^2\mu_X^2}{\sigma_U^2}}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_U}\right)^2}\right] \times && (65)\\
&\quad \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\frac{\sigma_X\sigma_Y}{\sigma_U}} \exp\left[\frac{-\left(V - \frac{\sigma_X^2(U-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_U^2}\right)^2}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_U}\right)^2}\right] dV. && (66)
\end{aligned}$$

The term inside the integral is just a Gaussian distribution in V (with $\sigma = \frac{\sigma_X\sigma_Y}{\sigma_U}$), so the integral just yields 1. Therefore,

$$p(U) = \frac{1}{\sqrt{2\pi}\sigma_U}\exp\left[\frac{\left(\frac{\sigma_X^2(U-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_U^2}\right)^2 - \frac{\sigma_X^2(U-\mu_Y)^2+\sigma_Y^2\mu_X^2}{\sigma_U^2}}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_U}\right)^2}\right] \quad (67)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_U}\exp\left[\frac{\left(\sigma_X^2(U-\mu_Y)+\sigma_Y^2\mu_X\right)^2 - \sigma_U^2\left(\sigma_X^2(U-\mu_Y)^2+\sigma_Y^2\mu_X^2\right)}{2\sigma_U^2\left(\sigma_X\sigma_Y\right)^2}\right] \quad (68)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_U}\exp\left[\frac{\left(\sigma_X^2(U-\mu_Y)+\sigma_Y^2\mu_X\right)^2 - \sigma_U^2\left(\sigma_X^2(U-\mu_Y)^2+\sigma_Y^2\mu_X^2\right)}{2\sigma_U^2\left(\sigma_X\sigma_Y\right)^2}\right] \quad (69)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_U}\exp\left[-\frac{(U-(\mu_X+\mu_Y)^2)(\sigma_X\sigma_Y)^2}{2\sigma_U^2\left(\sigma_X\sigma_Y\right)^2}\right] \quad (70)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_U}\exp\left[-\frac{(U-(\mu_X+\mu_Y)^2)}{2\sigma_U^2}\right], \quad (71)$$

which is a Gaussian distribution with $\mu_U = \mu_X + \mu_Y$, and $\sigma_U^2 = \sigma_X^2 + \sigma_Y^2$ or $V(U) = V(X) + V(Y)$.

So there we have shown that the sum of two random normal variables yields another normally distributed random variable and we've verified Eqn. 46!

We can check that our calculation is correct using a Monte Carlo method. Again, the `random` module from `numpy` has several pdfs from which (pseudo)random numbers can be generated. Something like the snippet below will generate 1000 normally distributed random numbers $X \sim \phi(X; 5, 0)$, and save them in a `python` list called `toys_X`.

```
import numpy

muX, sigmaX = 5., 0.

toys_X = numpy.random.normal(muX,sigmaX,1000)
```

We can combine random numbers to generate new distributions and estimate properties of that new distribution, for example its moments, using the toys. Take a look at the **GaussianCompositions.ipynb** notebook. The code there will generate toys from two Gaussian distributions $X \sim \phi(X; 10, 1)$, $Y \sim \phi(Y; -6, 0.5)$ using the `numpy.random.normal` function. The generated values of X and Y are summed, and those values are filled into a histogram. We should be able to see that this distribution looks like another Gaussian distribution with $\mu = -6 + 10 = 4$ and $\sigma = \sqrt{0.5^2 + 1^2} = 1.1$.

The output of the notebook is shown in Fig. 7. Be warned however, that compositions of even Gaussian random variables doesn't always result in something Gaussian. For example, if $X \sim \phi(X; 0, 1)$ and $Y \sim \phi(Y; 0, 1)$ and $U = \frac{X}{Y}$, then (as you'll show in one of the problems) $p(U)$ is a Cauchy distribution,

$$p(U) = \frac{1}{\pi(1+U^2)}, \quad (72)$$

which has no mean value and infinite variance!. In this case, the usual error propagation formula can be quite misleading.

Week 3 Estimators and their properties

You are certainly aware of the concept of *convergence*. We used it already to demonstrate how the Poisson distribution arises from the binomial distribution when the probability of success is very small. In statistics, there are different kinds of convergence so we need to be careful to specify the sense in which something converges to another.

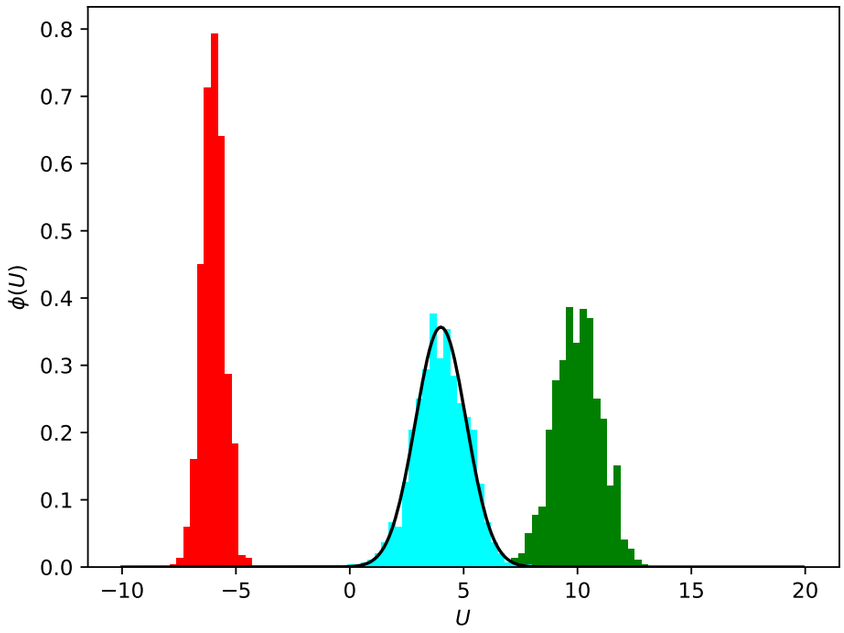

Figure 7: Distribution of toys generated from a Gaussian with $\mu = -6$, $\sigma = 0.5$ (red histogram) and $\mu = -10$, $\sigma = 1$ (green histogram). The cyan histogram shows the distrubtion of the sum of the toys, and the black line shows the Gaussian distribution with $\mu = -6 + 10 = 4$ and $\sigma = \sqrt{0.5^2 + 1^2} = 1.1$.

3.1 Types of convergence

The two most common senses of convergence in statistics are *convergence in distribution* and *convergence in probability*. First, we will define these two senses, without much discussion as to their application. This is because in the end, as particle physicists, the main result we care about is the *central limit theorem*, which turns out to be extremely useful for testing hypotheses and measuring physical properties – i.e for the main job of a particle physicist.

Convergence in distribution: Consider a sequence of random variables $\{X_1, X_2, ..., X_n\}$ $(X_i \varepsilon \mathbb{R})$ with cumulative distribution functions $\{F_1(x), F_2(x), ..., F_n(x)\}$. The sequence X_n converges in distribution as $n \to \infty$ to X, with cumulative distribution F if for every point $x \varepsilon \mathbb{R}$ where $F(x)$ is continuous,

$$\lim_{n\to\infty} F_n(x) = F(x), \tag{73}$$

where $F(x)$ is the CDF of X. What this says is that if the CDFs of a sequence of random variables converge to the CDF of X then X_n is said to tend to X in distribution as $n \to \infty$. What this means is that as n gets large, the distribution of of X_n will be better and better approximated by the distribution of X.

Convergence in probability: We say that the sequence of random variables $\{X_1, X_2, ..., X_n\}$, *converges in probability* to a random variable X if for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0. \tag{74}$$

This definition of convergence is stronger than convergence in distribution since nothing about the distance between X and X_n is implied in the other definition. Convergence in probability implies convergence in distribution.

The most important application of convergence is in the *law of large numbers*, which concerns the convergence of the sample mean to a fixed value. Let $\{X_1, ...X_N\}$ be a sequence of *independent* random variables, each having the same mean μ_1 and variances ν_2^i. We determine the *sample mean*

for this sequence, $\bar{X}_N$ as,

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i, \tag{75}$$

where i've explicitly added the subscript N to denote that this sample mean is itself a random variable calculated from the sequence of independent random variables. Note that this is not the same as μ_1 since this refers to the expectation value of X_i under their probability distributions. If μ_1 exists, then if either,

$$\lim_{N\to\infty}\left[\frac{1}{N^2}\sum_{i=1}^{N}\nu_2^i\right] = 0 \quad \textbf{or} \quad \lim_{N\to\infty}\left[\sum_{i=1}^{N}\frac{\nu_2^i}{i^2}\right] \text{ is finite,} \tag{76}$$

then $\bar{X}$ converges in probability to μ_1. In this case, rather than the sequence converging to a random variable, it converges to a fixed number (μ_1).

Example: To understand the law of large numbers, take a look at the **DemoLawOfLargeNumbers.ipynb** notebook. Here you will find a simple example showing how the sample mean of a Gaussian random variable converges to the μ parameter of the Gaussian (which is the mean of the Gaussian disribution). You can play with this notebook to understand how convergence in probability works in the law of large numbers. This will be particularly important when we talk about the method of moments in later lectures.

3.2 The central limit theorem

So far we have talked separately convergence of sequences of distributions and the convergence of the sample mean of independent random variables. The *central limit theorem* (CLT) states how the sum of independent random variables (for example as in sample means) is *distributed* in the limit of large N. Suppose we have a sequence of independent random variables X_i, each from a distribution with mean μ_1^i and variance ν_2^i. Now let,

$$T_N = \frac{\bar{X}N - \sum_{i=1}^{N}\mu_1^i}{\sqrt{\sum_{i=1}^{N}\nu_2^i}}. \tag{77}$$

Then for $T = \lim_{N\to\infty} T_N$, we have that $T \sim \phi(T; 0, 1)$ – meaning that the limit T is distributed as a standard normal distribution (a Gaussian with a width of 1 and a mean of 0) – T_N converges in distribution to a standard normal distribution.

We could attempt to demonstrate this by explicitly writing out the probability for the sum of the random variables to be contained in some small interval - but remember how long and tedious that was with just summing two normal distributed variables! Instead, we can rely on the *Paul Levy theorem* to do the heavy lifting. First, we define the *characteristic function* $\varphi_X(t)$ of a probability distribution function $f(X)$ by,

$$\varphi_X(t) = E\left[e^{itX}\right] = \int_{-\infty}^{\infty} e^{itX} f(X) dX. \tag{78}$$

The characteristic function completely determines the probability distribution of the random variable - i.e if we know $\varphi_X(t)$, then we know $f(X)$ too. The characteristic function has useful properties, for example, if X and Y are independent random variables, with characteristic functions $\varphi_X(t)$ and $\varphi_Y(t)$, then the characteristic function of the sum $X + Y$ is $\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t)$. Furthermore, the characteristic function of the sum of independent variables is the product of the individual characteristic functions of the variables, i.e,

$$\varphi_{X_1+\ldots+X_N}(t) = \prod_{i=1}^{N}\varphi_{X_i}(t). \tag{79}$$

We can use this to demonstrate the central limit theorem in the case of a sum of independent and identically distributed random variables.

Example: Let $\{X_1, ..., X_N\}$ be independent and identically distributed random variables with $E(X) = 0$ and $V(X) = \nu \in \mathbb{R}$. Then,

$$T_N = \frac{X_1 + ... + X_N}{\sqrt{\nu N}} = \sum_{i=1}^{N} \frac{Y_i}{\sqrt{N}}. \tag{80}$$

where $Y_i = X_i/\sqrt{\nu}$. We can say that the characteristic function of T_N is given by,

$$\varphi_{T_N} = \varphi_{Y_1}\left(\frac{t}{\sqrt{N}}\right) \cdot \varphi_{Y_2}\left(\frac{t}{\sqrt{N}}\right) \cdot ... \cdot \varphi_{Y_N}\left(\frac{t}{\sqrt{N}}\right) = \left(\phi_{Y_1}\left(\frac{t}{\sqrt{N}}\right)\right)^N \tag{81}$$

from Eqn. 79 and the linearity property of the expectation. Recall that the characteristic function is given by the expectation so,

$$\varphi_{Y_1}\left(\frac{t}{\sqrt{N}}\right) = E\left[e^{i\frac{t}{\sqrt{N}}Y_1}\right] = E\left[\sum_{r=0}^{\infty} \frac{\left(i\frac{t}{\sqrt{N}}Y_1\right)^r}{r!}\right] = \sum_{r=0}^{\infty} \frac{\left(i\frac{t}{\sqrt{N}}\right)^r}{r!} E\left[Y_1^r\right] \tag{82}$$

where we have Taylor expanded the exponential. Now, we know that the mean of Y_1 is 0 (since also $\mu = 0$) and the variance of Y_1 will be the same as the variance of X_1 divided by itself, hence $E[Y^2] = V(Y_1) = 1$. So we have,

$$\varphi_{Y_1}\left(\frac{t}{\sqrt{N}}\right) = 1 - \frac{t^2}{2N} + \mathcal{O}\left(\frac{t^3}{N^{\frac{3}{2}}}\right). \tag{83}$$

Substituting this back into Eqn. 81, we have that as $N \to \infty$, $\varphi_{T_N} \to e^{-\frac{1}{2}t^2}$, which is the characteristic function of a standard normal distribution. From the Paul Levy theorem, this means that the limit of T_N, T, is distributed as a standard normal distribution – $T \sim \phi(T; 0, 1)$.

Let's have a look at the CLT in action with a simulation. In the **GaltonMachine.ipynb** notebook, the idea is to simulate a Galton Machine, in which a counter is dropped through layers of pins which force the counter to go left or right. Figure 8 shows the setup with 5 layers of pins. In the code, for each layer, we'll randomly choose left or right for the counter direction and make a histogram of the position for many counters.

For different numbers of layers, we can see how the Gaussian distribution shape builds up as the number of layers increases in Figure 9. The final position is the sum of random integer variables, each of which has a distribution that is uniform between -1 and +1. This is the CLT in action! We can actually see two laws of large numbers at play - the CLT is what guarantees the converge in distribution to a Gaussian shape, while the law of large numbers is behind the fact that the mean value for each bin converges to the probability density as we increase the number of trials. The latter is convergence in probability and is what we rely on when using Monte Carlo (MC) simulations in particle physics event generators. This method of generating a Gaussian distribution is of course very inefficient – our modern day MC generators use much more sophisticated methods to generate probability (density) distribution functions for particle event simulations.

So far, we've only discussed probability densities for random variables, their properties and how these properties evolve with more and more data. In the data analysis world, this is extremely important as probability densities are what we use to construct models for understanding our data. Clearly this is where the statistics really comes in as extracting information from our data is the job of any experimental scientist. In order to do this properly, we need to understand the concept of *hypothesis testing* but before we do, we can first cover some methods of extracting information from data that will seem more intuitive and that you will hopefully be somewhat familiar with already.

The first thing we'll cover is the concept of an *estimator*. An estimator is simply an estimate of some parameter (could be the mean of a probability density function or the mass of the Higgs boson), using an observed data set. Necessarily, an estimator's value will depend on the data observed and as the data we observe are random variables, so too are the estimators.

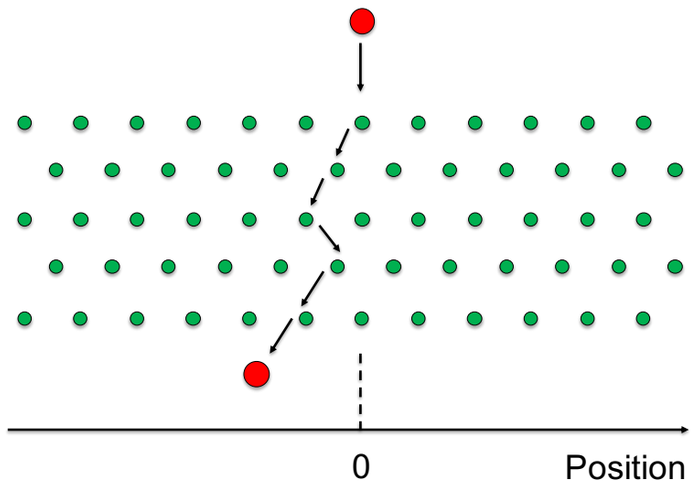

Figure 8: The Galton machine: The idea is to drop the red counter down and let it work its way to the bottom. Each green pin will knock the counter left (-1 to its position) or right (+1 to its position) with an equal probability for each. At the end we'll count how many counters are at each position.

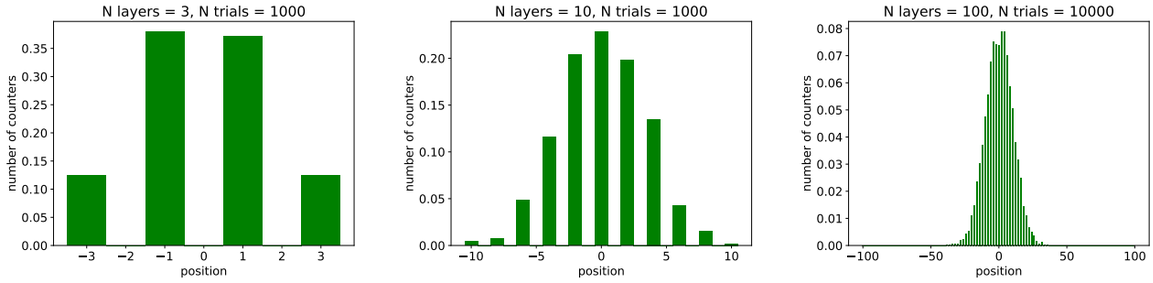

Figure 9: Galton machine simulation for 3, 10 and 100 layers. As the number of layers increases, we need more trials (counters) to fill in the possible positions but also we start to see the Gaussian distribution appear.

With that in mind, it would be also useful to know the variance of our estimators (some idea about varied our estimator might be) in order to check how stable they are with respect to the true values that we want to estimate. A desireable property of an estimator is that for large data sets (as the number of data increases), the estimator should converge to the true value of the parameter. Otherwise, we would end up with biased results even in the limit of infinite data (which of course is not ideal). We call this property *consistent*.

3.3 Method of moments

We'll start with a very simple estimator, being the method of moments. We can estimate parameters of probability densities from observations in data by using the method of moments. In this method, we try to match sample moments with (combinations of) parameters in the probability density model and use them as estimators of those parameters. To do this, we first need to define the moments of our probability density (provided they exist!) in terms of its parameters. We can either use the central or algebraic moments, depending on whichever is more convenient. The simplest way to explain this is with a couple of basic examples.

Example: In this example, we'll look at a Binomial distribution for a random variable k. Remember that k is the number of successful trials in n independent trials and that the Binomial parameter p is the probability for an individual trial to be successful. We can think of k as the sum over independent random numbers X_i who's value can be either 0 (not success) or 1 (success), then the expectation value of the Binomial distribution is,

$$E[k] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = np \tag{84}$$

We've used the fact that $E[X_i] = 0 \times (1-p) + 1 \times p$, from the definition of the expectation value and since the probability for success $(X = 1)$ is p.
If we observe a single value k successes in a given set of trials, then $\overline{k} = k$, and from the method of moments we have,

$$E[k] = \bar{k} \implies np = k \tag{85}$$

or the estimator for p is $\hat{p} = \dfrac{k}{n}$.

Example: In this example, we'll look at a Gaussian distributed random variable $X \sim \phi(3, 1.5)$ - i.e X is distributed as a Gaussian with $\mu = 3$ and $\sigma = 1.5$. For a Gaussian, we know that the first algebraic moment (the mean) is given by,

$$\mu_1 = E[X]_\phi = \mu$$

and the second central moment (variance), is

$$\nu_2 = E[(X - E[X])^2]_\phi = \sigma^2$$

Now in general for a sample of n observations, the sample mean is given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the sample variance is given by

$$\bar{V} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

In the method of moments then, we equate the sample mean with μ_1 and the sample variance with ν_2 to obtain our estimates $\hat{\mu}$ and $\hat{\sigma}$,

$$\mu = \mu_1 \implies \hat{\mu} = \bar{X}$$

and

$$\sigma^2 = \nu_1 \implies \hat{\sigma} = \sqrt{\bar{V}}$$

In both of these examples, our intuition should tell us what the estimators should be, but the method of moments makes this explicit for us and can be used in more complicated scenarios.

3.4 Properties of good estimators

What makes an estimator a good estimator? There are many different ways to define "good" estimators, which can include whether they can be easily used for your particular problem or how easy they are to implement in software for a particular data analysis problem. When choosing an estimator, a statistician will usually have in mind one or more property to be compare estimators. These can be the loss of information using a particular estimate, the variance of the estimator or even the simplicity of explaining it in a publication.

3.4.1 Consistency

We say that an estimator is *consistent* if as more data are added ($n \to \infty$), the estimator $\hat{\theta}$ for the parameter θ converges to the true value of θ. This is clearly a good property as we want our estimates to get more accurate with more and more data (right?). Formally, for a sequence of estimators $\hat{\theta}_n$, obtained from a sample of size n, the estimator is *consistent* if the sequence *converges in probability* to the true value of θ as $n \to \infty$. Remember, this means that for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(|\hat{\theta}_n - \theta)\right| > \epsilon) = 0. \tag{86}$$

Let's take a look at the method of moments estimator for the μ parameter of a Gaussian probability distribution. We can calculate the variance of the estimator $\bar{X}$ as $n \to \infty$.

$$V(\bar{X}_n) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n}\sigma^2. \tag{87}$$

which converges to 0 as $n \to \infty$. So for larger and larger values of n, the distribution of $\bar{X}$ gets more narrow - i.e the sequence $\bar{X}_n$ seems to be converging to some fixed value. This fact alone is not enough to tell us that the estimator is consistent as it only tells us something about the width of the distribution of the estimator as n gets large and not what the value of the estimator is converging to.

We can show explicitly that the method of moments estimator for the μ parameter of a Gaussian is a consistent estimator by using the law of large numbers. Remember that the law of large numbers says that if we have a sequence of random variables X_i distributed according to distributions with first algebraic moments all the same (μ_1) and second central moments ν_2^i, then if

$$\lim_{n\to\infty}\left(\frac{1}{n^2}\sum_{i=1}^{n}\nu_2^i\right) = 0, \tag{88}$$

we have that the sample mean converges in probability to μ_1. In our case every X_i is distributed according to the same distribution - the Gaussian distribution - and this this case we have $\mu_1 = \mu$ and the estimator $\hat{\mu} = \bar{X}$. If we can satisfy the law of large numbers requirement above, then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu_1 = \mu$, and that would mean $\hat{\mu} \to \mu$.

If we substitute the variance of the Gaussian random variable X into the equation 88 above - i.e $v_2^i = \sigma^2\ \forall i$ - we find that,

$$\lim_{n\to\infty}\left(\frac{1}{n^2}\sum_{i=1}^{n}\nu_2^i\right) = \lim_{n\to\infty}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\right) = \lim_{n\to\infty}\left(\frac{1}{n}\sigma^2\right) = 0, \tag{89}$$

so therefore by the law of large numbers we must have that $\hat{\mu}_n \to \mu$ as $n \to \infty$.

It can in general be shown that any of the method of moments estimators derived as a solution to the equations which match sample moments to moments of the distribution will yield consistent estimators (but we won't do this in these lectures).

3.4.2 Bias

The bias of an estimator $\hat{\theta}$ is defined as the difference between the expectation value of the estimator and the true value of the parameter θ, ie,

$$\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta. \tag{90}$$

We say that an estimator is *unbiased* if the bias is zero.

In the method of moments for the mean and standard deviation (eg when estimating the parameters of a Gaussian distribution), we can find the bias of the estimators for μ and σ. First, the estimator $\hat{\mu}$ is unbiased as can easily be seen from the definition of the bias. For a random variable X,

$$E[\hat{\mu}] = E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n} n\mu = \mu, \tag{91}$$

since the expectation operator is linear.

Instead, the estimator for σ is biased. We can see this by first showing that the estimator for σ^2 is biased. The expectation of the estimator $\hat{\sigma^2}$ is given by,

$$\begin{aligned}
E\left[\hat{\sigma^2}\right] = E\left[\bar{V}\right] &= E\left[\frac{1}{n}\sum_i (X_i - \bar{X})^2\right] && (92)\\
&= E\left[\frac{1}{n}\sum_i \left((X_i - \mu) - (\bar{X} - \mu)\right)^2\right] && (93)\\
&= E\left[\frac{1}{n}\sum_i \left((X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu)\right)\right] && (94)\\
&= E\left[\frac{1}{n}\sum_i (X_i - \mu)^2 + (\bar{X} - \mu)^2 - \frac{2}{n}\sum_i (X_i - \mu)(\bar{X} - \mu)\right] && (95)\\
& && (96)
\end{aligned}$$

Now notice that

$$\bar{X} - \mu = \frac{1}{n}\sum_i (X_i - \mu), \tag{97}$$

and hence

$$\begin{aligned}
E\left[\hat{\sigma^2}\right] &= E\left[\frac{1}{n}\sum_i (X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(\bar{X} - \mu)^2\right] && (98)\\
&= E\left[\frac{1}{n}\sum_i (X_i - \mu)^2 - (\bar{X} - \mu)^2\right] && (99)\\
&= E\left[\frac{1}{n}\sum_i (X_i - \mu)^2\right] - E\left[(\bar{X} - \mu)^2\right] && (100)\\
&= \frac{1}{n}\sum_i E\left[(X_i - \mu)^2\right] - V(\bar{X}) && (101)\\
&= \frac{n}{n}V(X) - V(\bar{X}) && (102)\\
&= \sigma^2 - \frac{1}{n}\sigma^2 < \sigma^2. && (103)
\end{aligned}$$

where in the last line we used the fact that $V(\frac{1}{n}X) = \frac{1}{n^2}V(X)$ (which we covered earlier in the lectures) and $V(\sum_i X_i) = \sum_i V(X_i) = n\sigma^2$ since the variance of the sum of *independent random variables* is the sum of the variance of those random variables (can you show this?).

Finally, we can use a result known as Jensen's inequality, that states for any *convex* function $g(x)$, $E[g(X)] \geq g(E[X])$. We won't prove this, but take a look at the example, $g(x) = x^2$ which is a convex function. Figure 10 shows how a variable X, which is normal distributed, has its distribution *stretched* into a new distribution for Y when $Y = X^2$. The expectation value of X does not map onto the expectation value of Y under the same function, i.e $E[Y] = E[X^2] \neq (E[X])^2$.

Since $g(x) = x^2$ is a convex function, using Jensen's inequality we have that,

$$\left(E\left[\hat{\sigma}\right]\right)^2 \leq E\left[\hat{\sigma}^2\right] < \sigma^2 \implies E\left[\hat{\sigma}\right] < \sigma \tag{104}$$

so the bias of the estimator $\hat{\sigma}$ is clearly not zero proving that the estimator for the σ parameter is a biased estimator.

Take a look at the notebook **MethodOfMomentsConsistencyAndBias.ipynb**. In this notebook, we use monte carlo simulation to demonstrate the consistency and bias of the estimators $\hat{\mu}$ and $\hat{\sigma}$ of the parameters of a Gaussian distribution. The output will look like Figure 11.

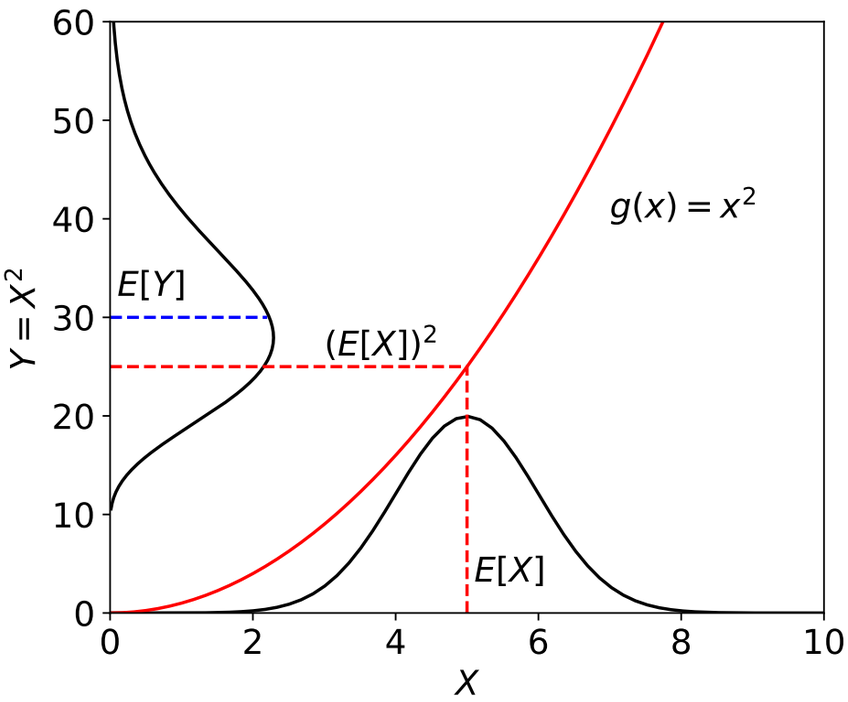

Figure 10: Cartoon of Jensen's inequality for the function $g(x) = x^2$ The variable X is normally distributed and when applying $Y = X^2$, the distribution of Y is stretched so that it has a longer tail. This means that $E[Y] = E[X^2] > (E[X])^2$ in this case.

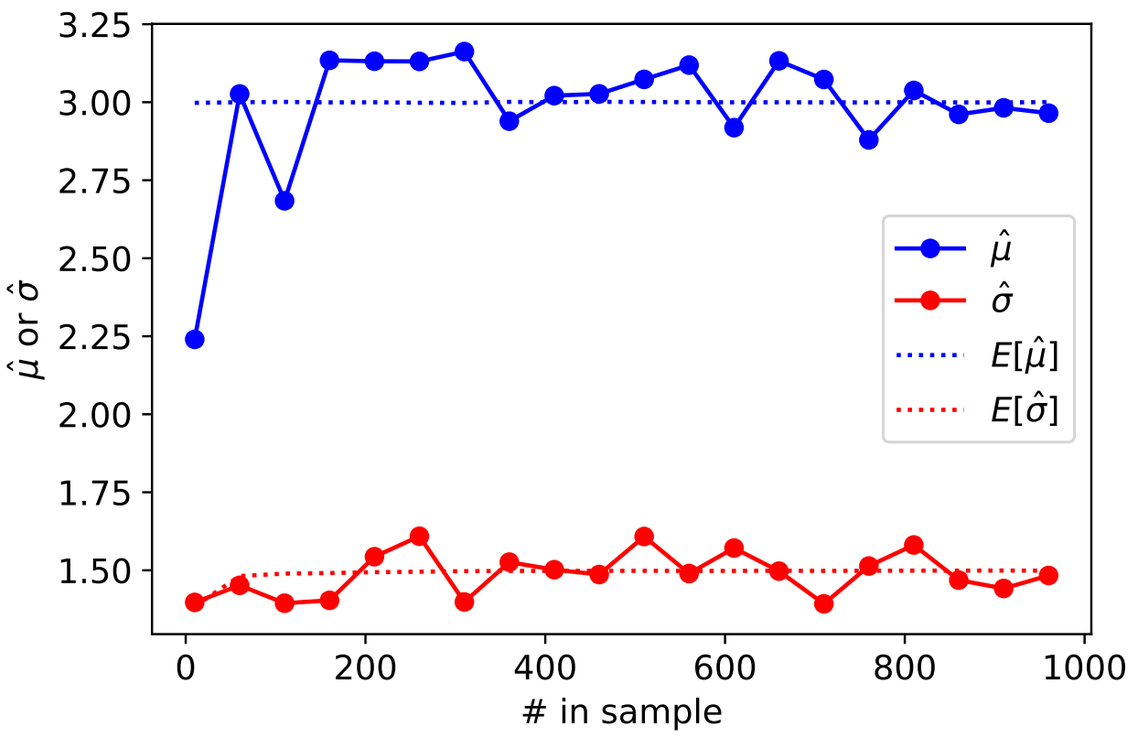

Figure 11: Method of moments estimators $\hat{\mu}$ and $\hat{\sigma}$ for the parameters of a Gaussian with true parameter values $\mu = 3.0$, $\sigma = 1.5$ for different sample sizes. The points show a particular outcome of the estimators while the lines show the expectaction values of these estimators.

As can be seen in the figure, both $\hat{\mu}$ and $\hat{\sigma}$ are consistent since they look to converge to their true values. However, while the estimator $\hat{\mu}$ is unbiased, the estimator $\hat{\sigma}$ has a non-zero bias.

3.5 Bootstrapping

In statistics, it is not always the case that you will have knowledge of the exact, or even approximate probability distributions for your data. In this case, its extremely difficult to just write down a likelihood and proceed to determine estimates, confidence/credible intervals for various parameters that you might be interested in.

The method of bootstrapping instead, only relies on re-sampling from your existing dataset. Suppose you have a dataset $X_1, X_2, ..., X_N$ of random, *independent observations*. A bootstrap sample is produced by selecting uniformly at random one of the observations N times, replacing them each time one is selected. From multiple such samples, you can calculate some quantity and use them to estimate properties of the distribution of that quantity.

We can use python's `random.choice` module for this, as in the snippet below which picks 6 random elements from a python list of 10 integers,

```
import random
random_selection = random.choices([1,4,2,5,6,3,8,7,9,10],k=6)
print(random_selection)
[1,3,3,7,6,4]
```

Note that the 'choices' function samples *with replacement* (so we can see the same element more than once in the returned list). To sample *without replacement* in python, we can use 'random.sample' instead.

Remember that for a sequence of N random numbers, the sample mean is given by,

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

We can estimate the variance (or standard deviation) of $\bar{X}$ by using bootstrap samples. The way this works is as follows;

- Select N observations (replacing each time) from the original sequence of random numbers. This is one bootstrap sample.
- Calculate the mean from the bootstrap sample and keep track of it.
- Repeat the first two steps many times (say 100 times) and calculate the standard deviation of the 100 bootstrap sample means. This value is your estimate of the standard deviation of the sample mean.

The snippet below is an example of estimating the standard deviation of the sample mean in random numbers generated from a Gaussian distribution with $\mu = 10$ and $\sigma = 1$.

```
import numpy
observations = numpy.random.normal(10,1,100)
def sample_mean(data):
return numpy.mean(data)

nsamples = 1000
means = []

for s in range(nsamples) :
  bootstrap = random.choices(observations,k=100)
  means.append(sample_mean(bootstrap))

print numpy.std(means)
```

The result of this is around 0.1. This is much smaller than the standard deviation (σ) of the original Gaussian. This is expected because we are not estimating the σ parameter of the Gaussian, but rather we are estimating the standard deviation of the sample mean. You should be careful when dealing with small lists of observations since theres a limit to the number of unique bootstrap samples that can be obtained which is given by $\frac{(2N-1)!}{N!\times(N-1)!}$. In our example this number is $\frac{199!}{(100!)(99!)}$ so you don't need to worry.

Take a look at the notebook **Bootstrapping.ipynb**. You will find a demonstration of the bootstrap method being used in the method of moments estimators for the μ and σ parameters of a Gaussian distribution. As one expects, the variance (and hence standard deviation) estimated with the bootstrap method decreases as the sample size increases.

Week 4 Likelihoods

Perhaps the most common estimator on the market is the *maximum likelihood estimator*. This estimator can be computationally very expensive but the result is a very useful and general tool as an estimator for parameters of probability density functions - or for us, parameters of models for our data! We need to define what we mean first by a *likelihood*.

The likelihood function is a common tool in statistics, and as a particle physicist will be your go-to when performing statistical inference.

The likelihood function L is defined by,

$$L(\theta) \propto P(X|\theta). \tag{105}$$

The likelihood is defined for a fixed observation X and is proportional to the probability to observe X, given a value of some parameter set θ. The likelihood function is a *function of the parameters* θ. Note that we write a proportionality rather than an equality because the likelihood function is *not* a probability density. In practice, we often only care about ratios of likelihoods for different values of the parameter(s) and hence we almost never care about this constant of proportionality.

For N independent observations $\mathbf{X} = \{X_1, X_2, ..., X_N\}$, the likelihood function is,

$$L(\theta) := \prod_{i=1}^{N} f_i(X_i; \theta), \tag{106}$$

where f_i are the p.d.f for each observation.

We often make the distinction between *binned* likelihood and *un-binned* likelihoods.

A binned likelihood is simply a product of Poisson likelihoods since the count in each bin is distributed as a Poisson random variable. Clearly if you have some probability density function f, then the Poisson parameter λ for the bin defined as the range $[x_a, x_b]$ is defined as,

$$\lambda = N \int_{x_a}^{x_b} f(x)dx \tag{107}$$

where N is the number of observations in the dataset. This is needed to turn the density into a rate for the Poisson distribution.

To get the most out of the data, the un-binned likelihood is always better, however there may be restrictions that mean the binned likelihood is more appropriate. For example, if using MC simulation to estimate the distribution of some process, you may only be able to bin the simulation and hence you'll need to use a binned likelihood. Also, you might find that for computation, the binned likelihood is faster - this is usually the case if you have many events in your dataset.

4.1 Maximum Likelihood Estimators

Now we are ready to talk about *maximum likelihood estimators* or MLEs. The use of maximisation/minimisation is common when *estimating* one or more parameters of interest $\boldsymbol{\theta}$. Suppose we have a likelihood defined for our data $L(\boldsymbol{\theta})$. The maximum likelihood estimators $\hat{\boldsymbol{\theta}}$ are defined by,

$$\nabla L(\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = 0 \tag{108}$$

where $\nabla L(\boldsymbol{\theta}) = \left(\frac{\partial L}{\partial \theta_1}, \frac{\partial L}{\partial \theta_2} ... \frac{\partial L}{\partial \theta_n}\right)$. These are the values that maximise the likelihood function.

Typically it is often easier to minimize the negative log of the likelihood function numerically so we typically minimise $q(\boldsymbol{\theta}) = -\ln L(\boldsymbol{\theta})$ as this makes the numbers (and often the algebra) easier to deal with. Let's take a look at a simple example from my own field of particle physics.

Example: We'll take a look at an example of this method where we have a unstable particle (let's say a muon) that decays after being produced in a particle collision. In these processes, the probability that the muon decays after a certain time depends on its lifetime (τ). Suppose we had an experimental setup, capable of detecting when such decays were occuring and we marked down the times, after some t_0, that each decay occurs. The time of each decay will be a random variable and will be distributed according to an exponential decay function,

$$p(t) = \frac{1}{\tau} e^{-t/\tau} \tag{109}$$

The likelihood is just the product over all of the muon decay times,

$$L(\tau) = \prod_i \frac{1}{\tau} e^{-t_i/\tau} \tag{110}$$

and the negative log likelihood is,

$$q = -\ln(L) = N \ln(\tau) + \sum_i^N \frac{t_i}{\tau} \tag{111}$$

and its derivative is,

$$q' = \frac{N}{\tau} - \frac{1}{\tau^2} \sum_i^N t_i \tag{112}$$

Setting equation 112 to zero can be solved to yield the maximum likelihood estimate for τ,

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N t_i \tag{113}$$

4.1.1 Properties of maximum likelihood estimators

In the next section, we'll take a look at some properties of the maximum likelihood estimator and see why it turns out to be very useful as an estimator.

Invariance under change of observables Transformations of the data do not modify the maximum likelihood estimators, provided that the transformation is a bijection (one-to-one mapping) and that the transformation doesn't depend on the parameters being estimated. The reason is that (as we've seen before) under a change of variables $X \to Y = g(X)$,

$$f_Y(Y)dY = f_X(X)dX \implies f_Y(Y) = \frac{f_X(X)}{|g'(X)|} \tag{114}$$

where $g'(X) = \frac{dY}{dX}$. Since the likelihood is equal to the probability density for a random variable, the maximum of it (or minimum of its log) will not change when multiplying (adding) by a constant!

Preservation of relationships of MLEs under change of parameterisation Similarly, although it might seem rather trivial, but an important feature of a maximum likelihood estimate is that changing parameters does not modify the estimator and as a result any relationship between true values is equally valid for their maximum likelihood estimates. Suppose $\hat{\theta}$ is the maximum likelihood estimate for θ and $\alpha = g(\theta)$ is some function of θ, then the maximum likelihood estimator of α will be,

$\hat{\alpha} = g(\hat{\theta})$.

This is obvious since at the point $\theta = \hat{\theta}$,

$$0 = \frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial \alpha}\frac{dg}{d\theta} = \frac{\partial L}{\partial \alpha}g'(\theta) \implies \frac{\partial L}{\partial \alpha} = 0 \tag{115}$$

This is not always true for *any* estimator, but is always true for MLEs.

Example: In this example, we will take a look at a *counting experiment*. This simply means that the observation (data) is just a single number n – the observed event count, and all of the model parameters are included inside the Poisson parameter λ. This is the building block of any binned likelihood and so understanding these simple experiments will help build up to more complicated ones later. For example, in a typical particle physics experiment, we want to study the rate of some particle χ, produced in proton-proton collisions $pp \to \chi \to ff$.

$$\lambda(\theta) = \theta A + B, \tag{116}$$

In the experiment, the aim is to measure the cross-section θ from a measured number of events n. The rate of background events that we expect is B, and A just relates the cross-section to a number of expected signal events.
The likelihood in this case will be based on the Poisson probability since n is an integer count.

$$L(\theta) = \frac{\lambda(\theta)^n e^{-\lambda(\theta)}}{n!} \tag{117}$$

or we can take the negative log instead,

$$q(\theta) = -\ln(L(\theta)) = -n\ln(\lambda) + \lambda + \ln(n!) \tag{118}$$

Remember, the maximum likelihood estimator for θ ($\hat{\theta}$) is given by the value of θ for which $L(\theta)$ is maximised or equivalently when $q(\theta)$ is minimised.

$$\frac{dq}{d\theta} = -n\frac{d(\ln(\lambda))}{d\lambda}\frac{d\lambda}{d\theta} + \frac{d\lambda}{d\theta} = -n\frac{1}{\lambda(\theta)}A + A \tag{119}$$

since $\dfrac{d\lambda}{d\theta} = A$. To find $\hat{\theta}$, we set $\dfrac{dq}{d\theta} = 0$, and this is easily solved to give,

$$\hat{\theta} = \frac{n - B}{A} \tag{120}$$

We can re-parameterise the cross-section in terms of a 'coupling' c, which tells us how strong the interaction is. The cross-section is related to the coupling by $\theta = Gc^2$, where G is a constant. Since we know that MLEs are invariant under change of parameterisation, we can write,

$$\hat{c} = \sqrt{\frac{\hat{\theta}}{G}} = \sqrt{\frac{n - B}{AG}} \tag{121}$$

Suppose we instead find the MLE for c directly. We could instead write the condition for the MLE as,

$$\frac{dq}{dc} = -n\frac{d(\ln(\lambda))}{d\lambda}\frac{d\lambda}{dc} + \frac{d\lambda}{dc} = -n\frac{1}{\lambda(c)}2cAG + 2cAG, \tag{122}$$

since $\lambda(c) = AGc^2 + B$ and $\dfrac{d\lambda}{dc} = 2AGc$. Setting $\dfrac{dq}{dc}$ to zero, again it's easy to see that the solution is the same as before,

$$\hat{c} = \sqrt{\frac{n - B}{AG}}. \tag{123}$$

This shows that the maximum likelihood estimators for the parameters obeys the same functional dependence as the parameters themselves - the MLE is invariant under re-parameterisation!

Consistency in maximum likelihood estimators One of the properties that makes the *maximum likelihood estimator* a good estimator is the fact that it is always *consistent*. Remember that this means as the number of observations increases, the maximum likelihood estimate of a parameter converges to the real value of the parameter. We'll now show that the maximum likelihood estimate for one parameter $\theta \in \Omega$ is a consistent estimator. Note that this extends to an number of parameters, provided the likelihood function meets certain conditions. First, recall that for multiple observations $\mathbf{X} = X_1, X_2, X_3, ..., X_N$ of some observation, the likelihood function is defined by

Eqn. 106. We define the quantity $l_N(\theta)$, by

$$l_N(\theta) = -\frac{1}{N} \ln L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \ln f(X_i; \theta). \tag{124}$$

Notice that minimising $l_N(\theta)$ is the same as maximising $L(\theta)$ since the constants have no effect on where the extreme value lies.

By the law of large numbers, for all θ, $l_N(\theta) \to l(\theta)$ as $N \to \infty$, where,

$$l(\theta) = E\left[-\ln(f(X;\theta))\right]_{\theta=\theta_0} = \int -\ln(f(X;\theta)) f(X;\theta_0) dX, \tag{125}$$

and θ_0 is the true value of θ. The value θ_0 is the value of θ that minimises $l(\theta)$ as can be seen from,

$$\begin{aligned} l(\theta) - l(\theta_0) &= \int -\ln(f(X;\theta)) f(X;\theta_0) dX - \int -\ln(f(X;\theta_0)) f(X;\theta_0) dX & (126)\\ &= \int -\ln\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right) f(X;\theta_0) dX & (127)\\ &= E\left[-\ln\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right)\right]_{\theta=\theta_0} \geq -\ln\left(E\left[\frac{f(X;\theta)}{f(X;\theta_0)}\right]_{\theta=\theta_0}\right), & (128) \end{aligned}$$

where in the last step, we've used Jensen's inequality for the expectation of strictly convex functions, knowing that $-\ln(\cdot)$ is a convex function. Now we notice that,

$$\ln\left(E\left[\frac{f(X;\theta)}{f(X;\theta_0)}\right]_{\theta=\theta_0}\right) = \ln\left(\int \frac{f(X;\theta)}{f(X;\theta_0)} f(X;\theta_0) dX\right) = \ln\left(\int f(X;\theta) dX\right) = \ln(1) = 0, \tag{129}$$

so that,

$$l(\theta) - l(\theta_0) \geq 0 \tag{130}$$

for all θ – meaning θ_0 is the value of θ that minimises $l(\theta)$. We have that for each N $\hat{\theta}_N$ is the value of θ that minimises $l_N(\theta)$ and we know that $l_N(\theta) \to l(\theta)$ as $N \to \infty$. If the maximum likelihood estimates $\hat{\theta}$ are unique, it can be shown that $\hat{\theta}_N \to \theta_0$ as $N \to \infty$, which would prove that the maximum likelihood estimate (or minimum log-likelihood estimate) is consistent. While we won't explicitly prove this last part, it is intuitive since the sequence of likelihood functions l_N will have minimum values that get arbitrarily close to the minimum value of l. Since only one value of θ minimises these functions, it makes sense that these values also get arbitrarily close to θ_0.

Let's take a look at a simple example to demonstrate the maximum likelihood estimator as a consistent estimator.

Example: Suppose we want to estimate the two parameters $\vec{\theta} = (\mu, \sigma)$ of a normal distribution, $\phi(X; \mu, \sigma)$ given a set of observations $X_1, X_2, ..., X_N$. The likelihood function is,

$$q(\mu, \theta) = -\sum_{i=1}^{N} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X_i-\mu}{\sigma}\right)^2}\right) = N\ln(\sigma\sqrt{2\pi}) + \frac{1}{2}\sum_{i=1}^{N}\left(\frac{X_i-\mu}{\sigma}\right)^2. \tag{131}$$

Suppose the true values are μ_0 and σ_0. The maximum likelihood estimator for the μ parameter is $\hat{\mu}$ and we find it by solving $\frac{\partial q}{\partial \mu} = 0$. From the definition of q, we have,

$$0 = \frac{\partial q}{\partial \mu}|_{\mu=\hat{\mu}} = -\frac{1}{\sigma}\sum_{i=1}^{N}(X_i - \hat{\mu}) \tag{132}$$

which we can re-arrange to,

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X}. \tag{133}$$

We already showed that this estimator converges to μ_0 as $N \to \infty$ since its the same as the method of moments estimator. Can you show the same is true for the $\hat{\sigma}$ estimator? Take a look at the **MLEConsistency.ipynb** notebook where we'll use the `minimize` function from the `scipy.optimize` package to see how both estimators behave in the large N limit. The results are shown in Figure 12. Clearly as N increases, the maximum likelihood estimates for μ and σ get closer to the true values.

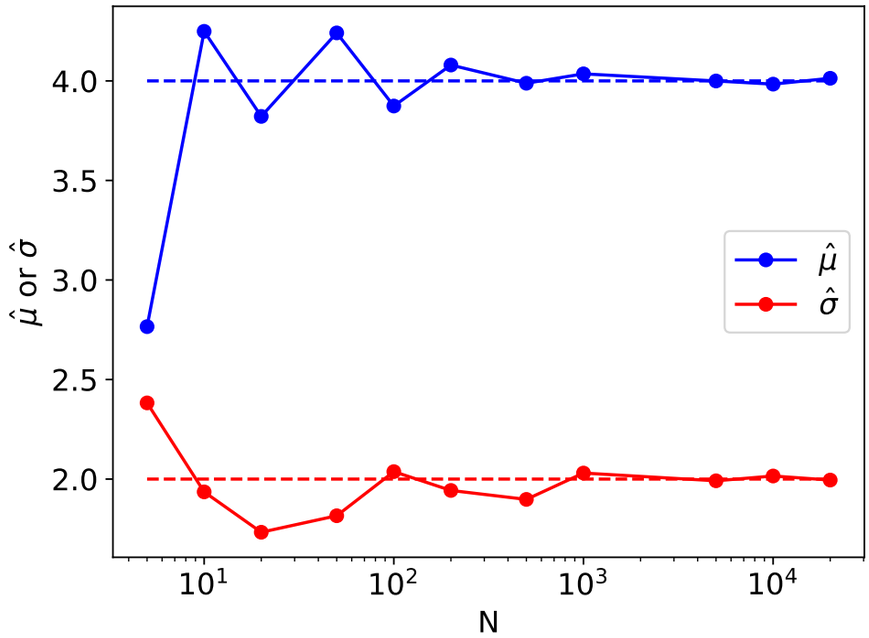

Figure 12: Maximum likelihood estimates for the two parameters of a normal distribution. The likelihood is calculated from N observations and the maximum likelihood estimates are determined for increasing values of N.

Efficiency (or minimum variance) Another important feature of the maxmimum likelihood ratio is that the variance of an estimator achieves the theoretical minimum for any unbiased estimator. There is a result by Cramer and Rao that states that the variance of an estimator of one parameter θ, $V(\hat{\theta})$ is bounded below by,

$$V(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2] \geq \frac{1}{I(\theta)} \tag{134}$$

where $I(\theta)$ is known as the Fisher information,

$$I(\theta) = E\left[\left(\frac{d(\ln f(X;\theta))}{d\theta}\right)^2\right]_{f(X;\theta)} = E\left[-\frac{d^2(\ln f(X;\theta))}{d\theta^2}\right]_{f(X;\theta)} \quad [1] \tag{135}$$

For the maximum likelihood estimator, the inequality in 134 becomes an equality as $N \to \infty$ or as the number of data goes to infinity. This means that the maximum likelihood is the unbiased estimator with the least variance for the large data limit.

It's simple to extend this to a multidimensional parameter case by exchanging the variance for the covariance and the Fisher information becomes the Fisher information matrix,

$$\text{covariance}(\hat{\theta}_i, \hat{\theta}_j) \geq \frac{1}{I_{ij}(\vec{\theta})} \tag{136}$$

and then,

$$I_{ij}(\vec{\theta}) = E\left[-\frac{\partial^2(\ln f(X;\vec{\theta}))}{\partial\theta_i \partial\theta_j}\right]_{f(X;\vec{\theta})} \tag{137}$$

4.1.2 Least squares from MLEs

Linear regression can be used as a very simple way to estimate parameters for a model to describe relationships between random variables (and hence observables in datasets).

Suppose we want to estimate parameters of a linear relationship.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip}, \;\; i = 1 \ldots n, \tag{138}$$

meaning we have p unknowns and n data points to obtain them.

Suppose that each of the y_i are random variables with a normal distribution centered on $\sum_{k=0}^{p} \beta_k x_{ik}$ with a variance of σ^2 - i.e $y_i \sim \phi(\sum_{k=0}^{p} \beta_k x_{ik}, \sigma)$.

We can write this all in matrix form as,

$$\boldsymbol{y} = X\boldsymbol{\beta}, \tag{139}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \tag{140}$$

and

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots x_{1p} \\ 1 & x_{21} & x_{22} & \ldots x_{2p} \\ \vdots & & & \\ 1 & x_{n1} & x_{n2} & \ldots x_{np} \end{pmatrix} \tag{141}$$

such that the joint probability distribution of $\boldsymbol{y}$ is given by a multivariate Gaussian distribution. The likelihood is written as a function of the parameters $\boldsymbol{\beta}$ via,

$$L(\boldsymbol{\beta}) = p(\boldsymbol{y}|X\boldsymbol{\beta}) = \frac{1}{\sqrt{(2\pi)^n \det(\sigma^2 I)}} e^{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})^T(\sigma^2 I)^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})}, \tag{142}$$

where I is the identity matrix. Taking the negative log, we find that,

$$q(\boldsymbol{\beta}) = -\ln L(\boldsymbol{\beta}) = \frac{n}{2}\ln(2\pi\sigma^2) + ||\boldsymbol{y} - X\boldsymbol{\beta}||^2 \frac{1}{2\sigma^2}, \tag{143}$$

since $\det \sigma^2 I = \sigma^{2n}$ and as before $||\boldsymbol{k}||^2 = \sum_i k_i^2$. It should be obvious that since the first term in q is a constant (since σ is a constant), then minimising q is the same thing as minimising $||\boldsymbol{y} - X\boldsymbol{\beta}||^2$

[1] Here we've made some assumptions about the Fisher information but for us they are always going to be true.

– i.e the ordinary least squares solution will yield the same as the maximum likelihood estimator for β when the data are assumed to be distributed according to a multivariate Gaussian with known variance.

Lets take the case where we only have one independent observable (X) so we only have X and Y. Then we only have two parameters in the vector $\boldsymbol{\theta}$ - let's call them c and m - and we can write the quantity that we need to minimise as

$$s = \sum_{i=1}^{N} (mX_i + c - Y_i)^2. \tag{144}$$

and we have two equations to solve by minimising s.

$$\frac{\partial s}{\partial m} = 2\left(m\sum_{i=1}^{N} X_i^2 + c\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} X_i Y_i \right) = 0 \tag{145}$$

and

$$\frac{\partial s}{\partial c} = 2\left(m\sum_{i=1}^{N} X_i + c\sum_{i=1}^{N} 1 - \sum_{i=1}^{N} Y_i \right) = 0 \tag{146}$$

From equation 146, we can immediately find,

$$c = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i - m\sum_{i=1}^{N} X_i \right) \tag{147}$$

and substituting this into equation 145 gives us,

$$m\left[N\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i \right)^2 \right] = N\sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i \tag{148}$$

or, re-arranging,

$$m = \frac{N\sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i \right)^2} \tag{149}$$

You'll see this often referred to as the ordinary least-squares solution to linear regression and its a very simple way to estimate simple linear relationships between pairs of variables. Let's take a look at this method in action in an example.

Example: In the notebook **HubbleConstant.ipynb**, we've taken data from Hubble's paper "A relation between distance and radial velocity among extra-galactic nebulae" (1929) with the names, distances (d in Mpc) and radial velocities (v in km/s) of extra-galactic nebulae. The two quantities are related through Hubble's law as,

$$v = H_0 d, \tag{150}$$

where H_0 is the Hubble constant. We can use our simple linear regression to determine H_0 from the data. To estimate H_0, we optimise the least-squares,

$$s = \sum_{i=1}^{n} (v_i - H_0 d_i)^2, \tag{151}$$

where we have assumes the intercept at $d = 0$ is $v = 0$, the value of H_0 that minimises s (i.e such that $\frac{ds}{dH_0}$=0) is given by,

$$H_0 = \frac{\sum_i d_i v_i}{\sum_i d_i^2}. \tag{152}$$

The data and the result of the regressed line is plotted in Figure 13. We find a value of $H_0 \approx 424$ (km/s)/(Mpc), which is about 7× larger than the measured value today. This is due to the fact that the distances measured were around 7× smaller than they should be!

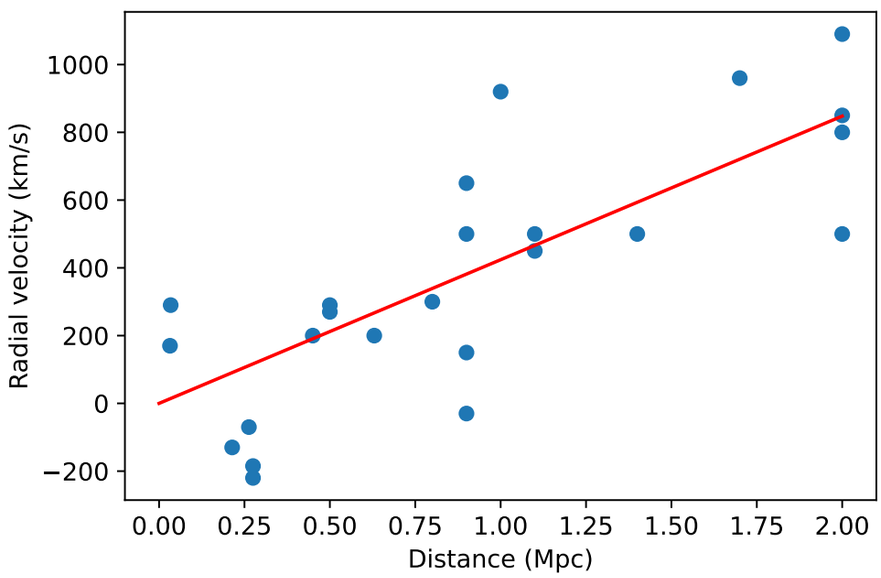

Figure 13: Measured values of distance to vs radial velocity of 24 extra-galactic nebulae from Hubble's 1929 paper. The red line indicates the straight line with a gradient equal to the H_0 value determined using the linear regression via least squares approach on the data.

In the example, we measured the Hubble constant to be $H_0 = 423.937$.

There are a number of great computer algorithms to numerically find the maximum likelihood estimates $\hat{\theta}$, including numerically solving the least squares, and you will have seen some already in your practical course. There's a whole field of research around this problem - optimisation - especially in machine learning applications where the number of parameters can be large. Lot's of these methods make use of the *gradient* of the function being optimised to increase the rate of convergence. We'll take a look at two such methods in the next section.

4.2 Gradient based optimisation

In Machine learning, often you will be optimising algorithms to improve the performance of the algorithm. In architechtures such as neural-networks, the weights of the network are *parameters* that can be varied in order to minimise some *loss* function.

You will see many examples of optimisation and these days there are some extremely sophisticated methods for optimising complicated loss functions. In statistics, the *likelihood* (or negative log-likelihood) function is our loss function and we often need to maximise (minimise) it with respect to one or more of its parameters - as is the case for the maximum likelihood estimator.

In general, this means we're looking for the set of parameter values ($\hat{\theta}$) for which,

$$\frac{\partial q}{\partial \theta}|_{\theta=\hat{\theta}} = 0 \tag{153}$$

where $q = -\log(L)$

There are two common algorithms that use the gradient of the likelihood to find the maximum point. Have a look at the **GradientOptimisation.ipynb** notebook where you'll find examples of the two methods in use.

4.2.1 Newton Method

You are probably very familar with the Newton method (from school even?), which relies on the gradient of the function to find roots (or zeros) of the function. In our setup, since we are trying to find the zeros of the gradient function, we'll also need to know the 2nd derivative.

This method is very reliable for 1-dimensional problems (one parameter that is being maximised). The way this works is that from some initial starting point θ_0, the algorithm proposes the next step as,

$$\theta_1 = \theta_0 - \frac{q'(\theta_0)}{q''(\theta_0)} \tag{154}$$

where $q' = \frac{dq}{d\theta}$, and $q'' = \frac{d^2q}{d\theta^2}$.

The algorithm continues until some tolerance is reached, $|q'| < \delta$ where δ is some small number greater than 0.

4.2.2 Gradient descent

For functions with more than one variable, we can use a gradient descent algorithm to find minima and maxima. These methods only rely on the first derivative (the gradient) to find the best direction to step in to find the minimum (or maximum). They don't use the second derivative because for a large number of parameters n, the number of terms in the second derivative grows as n^2 so this can get quite costly.

There are a number of subtle different algorithms which rely on gradient descent but we'll go over a simple example. The way this works is as follows;

First, we initialise the algorithm at some set of values for the parameters $\boldsymbol{\theta}_{init}$. Then we calcualte the gradient of the negative log-likelihood - $\nabla(q)|_{\boldsymbol{\theta}_{init}}$ - at that initial point. The gradient defines a direction in which the negative log-likelihood changes most rapidly - this makes this direction a good direction to search for a smaller value of q.

$$\nabla(q)|_{\boldsymbol{\theta}_{init}} = \begin{bmatrix} \frac{\partial q}{\partial \theta_1} \\ \frac{\partial q}{\partial \theta_2} \\ \vdots \\ \frac{\partial q}{\partial \theta_n} \end{bmatrix}_{\boldsymbol{\theta}=\boldsymbol{\theta}_{init}} \tag{155}$$

Since we want to *decrease* q, we need to move against this gradient. We step along the negative gradient direction until we find a new minimum - i.e we pick a step size h and keep adding it to the initial point until the value of q stops decreasing,

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{\text{init}} - k \times h \times \nabla(q)|_{\boldsymbol{\theta}_{init}}, \tag{156}$$

where k takes the value $1, 2, 3, \ldots$ etc. At the value of k such that we're no longer decreasing q (lets call it k^*), we recalculate the gradient and find a new direction to go in.

$$\nabla(q)|_{\boldsymbol{\theta}_{k^*}} \tag{157}$$

We keep repeating this process until the modulus of the gradient gets close enough to zero (say is smaller than some tolerance ϵ, that means we keep iterating until,

$$|\nabla(q)|_{\boldsymbol{\theta}}| = \sqrt{\left(\frac{\partial q}{\partial \theta_1}\right)^2 + \left(\frac{\partial q}{\partial \theta_2}\right)^2 + \ldots + \left(\frac{\partial q}{\partial \theta_n}\right)^2} < \epsilon \tag{158}$$

4.2.3 A neutrino oscillation experiment

Let's take a look at an maximum likelihood estimate example from neutrino physics. Neutrinos are very very light particles that interact very rarely with matter. We know from experiments (such as the T2K experiment in Japan) that neutrinos have a mass due to the fact they oscillate between the different types.

In the T2K experiment, neutrinos are produced at the J-PARC accelerator in Tokai (in Japan) and travel 295km to the Super Kamiokande detector in Kamioka.

Physicists can detect neutrinos interacting in the detector through the light they produce. We can predict how many neutrinos should be present a Kamioka as a function of their energy - know as the neutrino flux. You can follow the **NeutrinoOscillation.ipynb** notebook for this example.

The notebook uses the neutrino energy spectrum in data to extract neutrino oscillation parameters from the maximum likelihood estimate method. You should get the figure shown in Fig 14.

Week 5 Hypothesis testing

So far, we have only discussed the concept of *probability* and probability distributions. While this is a very important topic, we haven't yet touched on the main tool-set for learning from and drawing concrete conclusions from experimental data. The concept of *testing hypotheses* is of course at the heart of experimental science – confronting a scientific hypothesis with data is the way we progress science after all. Hypothesis testing is part of a whole branch in statistics called *decision theory*. In these lectures, we won't have time to go into details about decision theory – You can refer to James' book for a good discussion on decision theory, including the difference in approaches between Bayesians and frequentists. Instead it's enough to know that decision theory is concerned with the concept of reaching a decision, on the basis of experimental data, which will minimise the potential loss resulting from making the wrong decision – should we adjust the trigger settings to improve our selection efficiency? Should we build another particle collider? Should we go outside without that umbrella? Our focus here will be on the tools used to make these decisions (rather than the decision making itself). This is where the concept of hypothesis testing enters.

In research data we often deal with the case that one or more of our hypotheses involve some unknown parameter. We refer to these as *composite hypotheses*. If there are no unknowns however, the hypothesis is completely specified and we call this a *simple hypothesis*. A composite hypothesis can obviously be thought of as an ensemble of simple hypothesis. For now, lets stick to simple hypotheses.

Suppose that we have to choose between two hypotheses labelled H_0 and H_1, based on some experimental observations. We typically distinguish the two as $H_0 :=$ *the null hypothesis*, and $H_1 :=$ *the alternate hypothesis*. This is just a convention and we'll see that depending on the test,

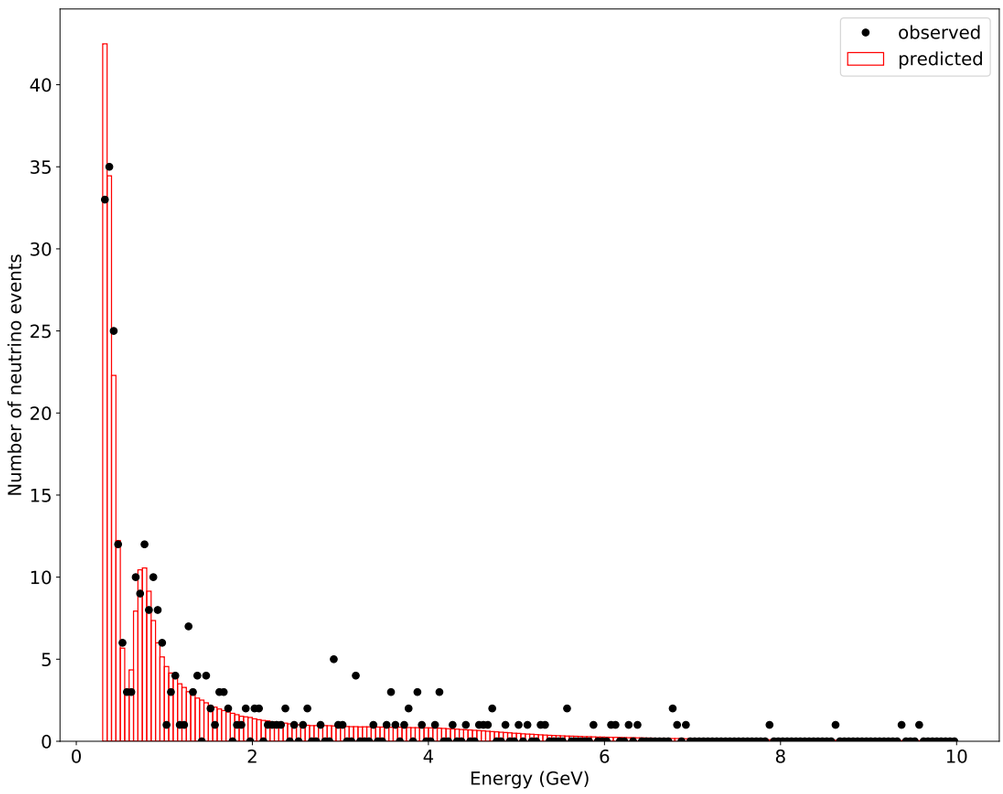

Figure 14: Neutrino energy spectrum predicted (red graph) at the neutrino oscillation experiment T2K at the maximum likelihood estimates using the observed data (black points).

the null and alternate can switch places. Let X be a function of the experimental observations which is supposed to summarize the observations – this is known as a *test statistic*. Choosing the best test statistic to summarize your experiment can be difficult, especially if the setup is complicated, but we'll see that there is a key result to guide us in this choice. First we need to establish a concept which you may be familiar and that is *errors* of type-I and type-II.

5.1 Type-I and type-II errors

Again, let's suppose we have a null hypothesis (H_0) and an alternate (H_1) Suppose then that we have our chosen test statistic $X \in \mathcal{W}$. We divide this region $\mathcal{W}$ into a *critical region* w and a *region of acceptance* $\mathcal{W} - w$). Observations of X falling into w would lead us to believe that our null hypothesis is not true. Defining a *test* of H_0 , given we've decided on our test statistic, then becomes choosing a critical region w.

Type-1 Error In practise, we often tune the critical region so as to obtain a particular probability (known as the *size* of the test) α that X falls into the critical region when H_0 is true (we usually say "under H_0"),

$$P(X \in w | H_0) = \alpha. \tag{159}$$

This is also referred to as the *size* of the test. You can see then that α is exactly the probability to *reject* the null hypothesis if the null hypothesis is *true* - we call this a *type-I error*. Thus, when defining a test, we have to accept that sometimes we will make this error. You'll often find that the level at which we set α strongly depends on what H_0 is. If for example, H_0 is a SUSY model with a particular mass scale, we might accept $\alpha = 0.05$ as a reasonable error. However, if H_0 is the standard model of particle physics, we'd certainly want to choose a much smaller number.

Type-II Error Of course, we also want to know how useful a test is at discriminating against the alternate hypothesis. This is known as the *power of the test*, and is defined as the probability of X falling into the critical region if H_1 is true (under H_1),

$$\text{power} = P(X \in w|H_1) = 1 - \beta. \tag{160}$$

Clearly this is related to the probability that X falls into the acceptance region via,

$$P(X \in \mathcal{W} - w|H_1) = 1 - P(X \in w|H_1) = \beta. \tag{161}$$

We often say that a hypothesis is more 'powerful' than another if its power is a greater value.

This is then the probability that we would *accept* the null hypothesis when the alternative is true – this is known as a *type-II* error. Of course, we want to make sure this error is equally unlikely, hence the choice of a test will amount to maximising the power of the test $(1-\beta)$ for a fixed size of the test. Figure 15 summarizes these two types of errors.

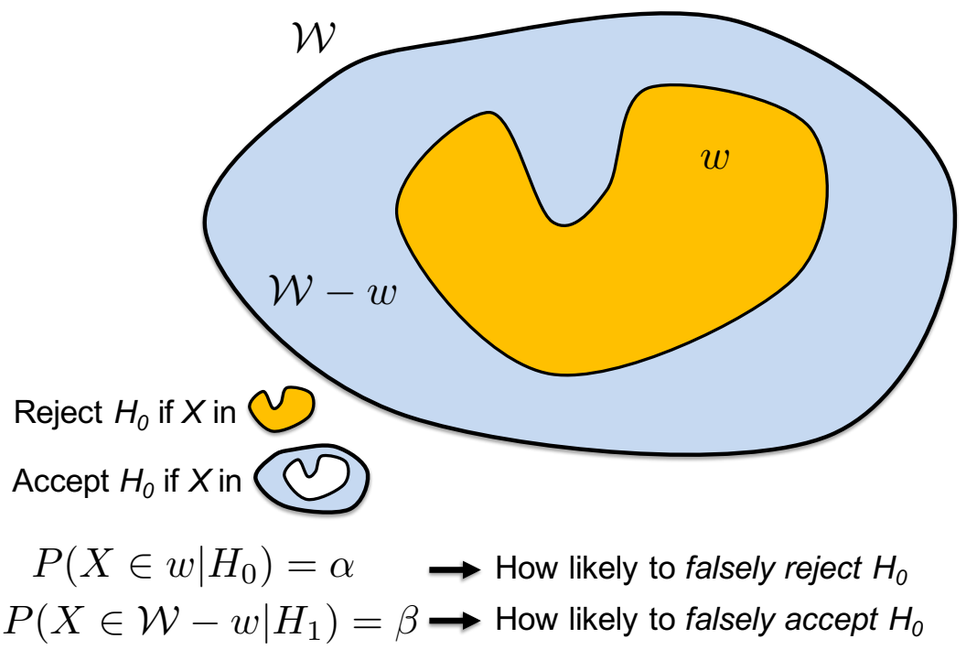

Figure 15: Division of possibilities for the test statistic (X) into the yellow critical region w, and the blue region of acceptance ($\mathcal{W}-w$). The probabilities that X falls into the critical region, under the null hypothesis and alternate hypothesis are α and $(1-\beta)$, respectively.

Some caution should be taken here. Remember we are dealing with two distinct simple hypotheses that cannot both simultaneously be true. Hence accepting one hypothesis means we are forced to reject the other (and vice versa). Physicists however don't usually talk about accepting hypotheses but rather rejecting them or not. This is because we rarely ever have a single test to establish that something is true, but we are often happy to use a single test to reject it - it's actually nearly always the same way with science, we tend to disprove theories rather than prove them outright. Next, we'll take a look at a very common test that you will come across when using machine learning methods.

The procedure for performing a Hypothesis test can be summarized in the following steps;

1. Define a test statistic X that summarizes the experimental data and has some ability to separate between H_0 and H_1.
2. Define a critical region w such that, $\int_w f(X|H_0) = \alpha$ for a specified Type-I error $0 < \alpha < 1$.
3. Determine the value of X in the observed data - X_{obs}.
4. Reject H_0 if $X_{obs} \in w$.

In the next few sections, we'll take a look at a few very common hypothesis tests.

5.2 Kolmogorov-Smirnov test

The Komogorov-Smirnov (KS) test is desined to detect differences in observed distributions from a model. This test is typically used when you have lots of data available (which is of course the typical scenario you'll be working in). It's not always the most powerful test because its designed to catch all kinds of deviations from your model.

For a random variable X (one dimensional) we can test the hypothesis that some probability distribution $f(X;\theta)$ is the underlying model for the data using the 1-sample KS test.

The KS test is based on the *cumulative distribution function* of the model vs the *empirical cumulative distribution* of the data. The empirical distribution $F_{data}(x)$ is given by,

$$F_{data}(x) = \frac{\text{number of data values } < x}{n} \tag{162}$$

where n is the number of entries in a data set (or collection of outcomes of X). Something like the snippet below can be used to calculate the CDF ($F(x)$) and empirical CDF in the case that $f(x)$ is a Gaussian with parameters μ, σ.

```
def cdf(x):
    return norm.cdf(x,mu,sigma)

def emp_cdf(data,x):
    return float(len(data[data < x]))/len(data)
```

The 1-sample KS test looks for deviations in these two distributions. First, we define the KS test-statistic,

$$D = \sup_x |F_{data}(x) - F(x)| \tag{163}$$

What this measures is the largest value over x of the difference between the two CDFs. Larger values of D would indicate worse agreement between the CDFs, therefore our critical region should be constructed from large values of D. To determine a critical region, we need to use toy data to determine the distribution of D under the null hypothesis.

5.3 Wald-Wolfowitz Runs test

The Wald-Wolfowitz (WW) runs test is an example of a test that is based on the concept of counting 'runs' in some dataset.

For a sequence of independent binary random values, we define a run as a maximal segment of the sequence with the same outcome. For example, the sequence 00011100000111 has a 4 runs, 2 of which are 0's and 2 of which are 1's.

In testing whether an observed sample is consistent with a specified hypothesis represented by a probability density $f(X)$, we can bin our data in a histogram and assign a $+$ to bins for which the histogram density is greater than the $f(X)$ and a $-$ to when the density is smaller than $f(X)$. For example, if figure 16, $f(X)$ is a Gaussian distribution.

The test statistic is simply the number of runs N. For this test-statistic, both small and large values of N, relative to $E[N]$, indicate disagreement of the data with H_0.

5.4 Student's t-test

The Student's t-test is a simple hypothesis test that is very easy to implement and applicable to a wide range of datasets. This time, we only need to specify the first algebraic moment - μ_1 - of

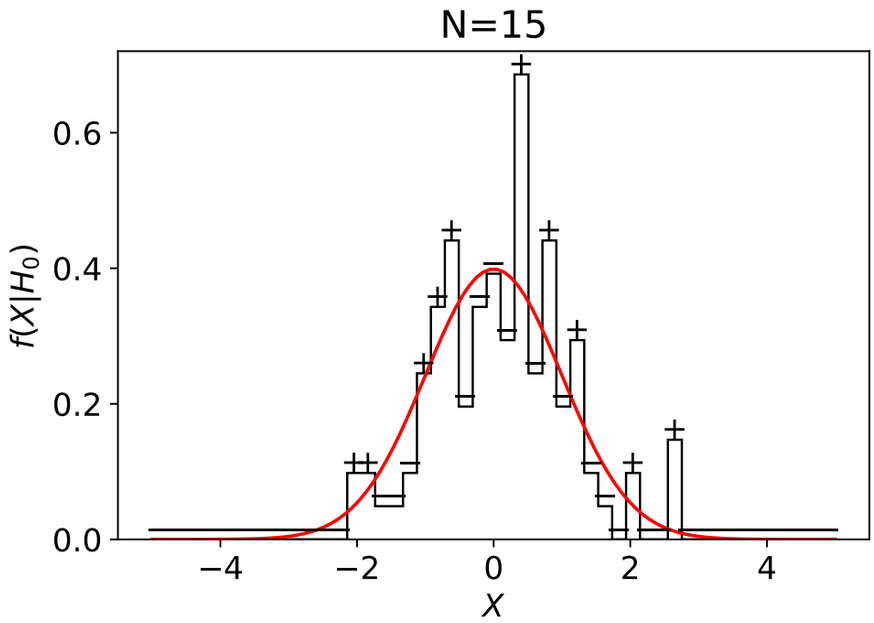

Figure 16: Gaussian probability density (red) and histogram data (black). Each bin is assigned a + or − depending on whether the density in the data is greater or less than the Gaussian density.

the probability distribution under H_0.

For a dataset $\mathbf{X} = \{X_1, X_2, ..., X_N\}$ The test statistic is defined by,

$$t = \frac{\sqrt{N}}{\sqrt{\bar{V}}} \left(\bar{X} - \mu_1\right)$$

where N is the sample size, and $\bar{X}$ and $\bar{V}$ are the sample means and sample variance of the dataset, respectively.

5.5 Two-sample tests

Two-sample tests are a little different from the other tests we've discussed. There are multiple use cases but an important one for machine learning (ML) applications is a test for *over-training* of a model.

When you come to work with classifiers (such as BDTs or neural-networks), you'll find that you often want to use some data to train the model, and some to test its performance. We want to know if the two distributions of data come from the same underlying model or not. In this case our H_0 is the hypothesis that they originate from the *same* underlying distribution (the alternate hypothesis being that they do not). If we can reject H_0, then this is evidence that the ML model is over-trained and therefore not generalisable to other data sets.

Of course, in order to perform our hypothesis test, we need a choice of H_0 and a way to estimate the distribution of whichever test-statistic we use under H_0. A good choice is to use H_0 as the sum the two distributions, randomly split into two distributions. By repeating this random split, we can create samples of data under H_0 - this is another form of bootstrapping!

5.5.1 Two-sample KS test

In the 2-sample KS test, our test-statistic is based on the two empirical CDFs, one for each test/train dataset.

$$D_2 = \sup_x \left|F_{testdata}(x) - F_{traindata}(x)\right|, \tag{164}$$

where

$$F_{data}(x) = \frac{\text{number of data values } < x}{n}$$

and n is the number of entries in the data set for which F_{data} is being determined.

Larger values of D_2 indicate disagreement with the hypothesis that the two distributions are the same. Since we don't have a probability density in this case to represent H_0, you can use bootstrapping to calculate the distribution of D_2 and select a critical region.

5.5.2 Wasserstein p=1 Test

The Wasserstein test (see "Markov Processes over Denumerable Products of Spaces, Describing Large Systems of Automata" - Wasserstein L, N (1969).), uses the difference between *quantiles* of the two distributions to test the H_0 hypothesis. If we have two datasets, $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ and Y $= \{Y_1, Y_2, ..., Y_m\}$ (i.e $\mathbf{X}$ could be the training data and Y could be our test data), then our test-statistic W_p is,

$$W_p = \frac{1}{K}\sum_{i=1}^{K}|Q_i(\mathbf{X}) - Q_i(\mathbf{Y})|, \tag{165}$$

where $K \leq \min(n, m)$ and $Q_i(\mathbf{D})$ is the function that returns i-th quantile of the dataset $\mathbf{D}$. The quantiles are calculated as the value within the dataset for which some fraction of the dataset lies to the left. For example, if the first quantile is 0.05, then the function Q_1 would need find the value p such that 5% of the events in D are smaller than the p-th value of dataset D_p and $Q_1(\mathbf{D}) = D_p$. Fortunately, there is a useful function in `numpy` to do that for us if we sort our dataset from smallest to largest,

```
data.sort()
quantiles = range(1,100,1) # the  quantiles we define
values = numpy.quantile(data,quantiles)
```

In the simple case where $n = m$ we can simplify the test statistic to,

$$W_p = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i|$$

Again, large values of W_p indicate evidence against H_0.

Example: In CMS, we use a Boosted Decision Tree (BDT) to distinguish between two different types of Higgs boson production. We use a sample of simulated events to train the BDT and another sample (which is not used to train) to test it. We want to know if the two samples are consistent with originating from the same underlying distribution (H_0) or not. If we can reject H_0, this would lead us to believe the BDT model is *overtrained*! Take a look at the **TwoSampleTests.ipynb** notebook. The notebook shows hypothesis tests using either D_2 and W_p as the test-statistic to check for over-training.
The two data distributions being tested and the distributions of D_2 and W_p are shown in figure 17 as obtained using a bootstrap method. The observed values and values defining the critical region of the test-statistics are also shown. In both cases, We can see that (thankfully) this model is not overtrained!

5.6 Permutation tests (Non-examinable)

Sometimes a test may not involve a parametric model or distribution of any kind but simply represent a statement about random observables in your dataset. A very common question to ask of two

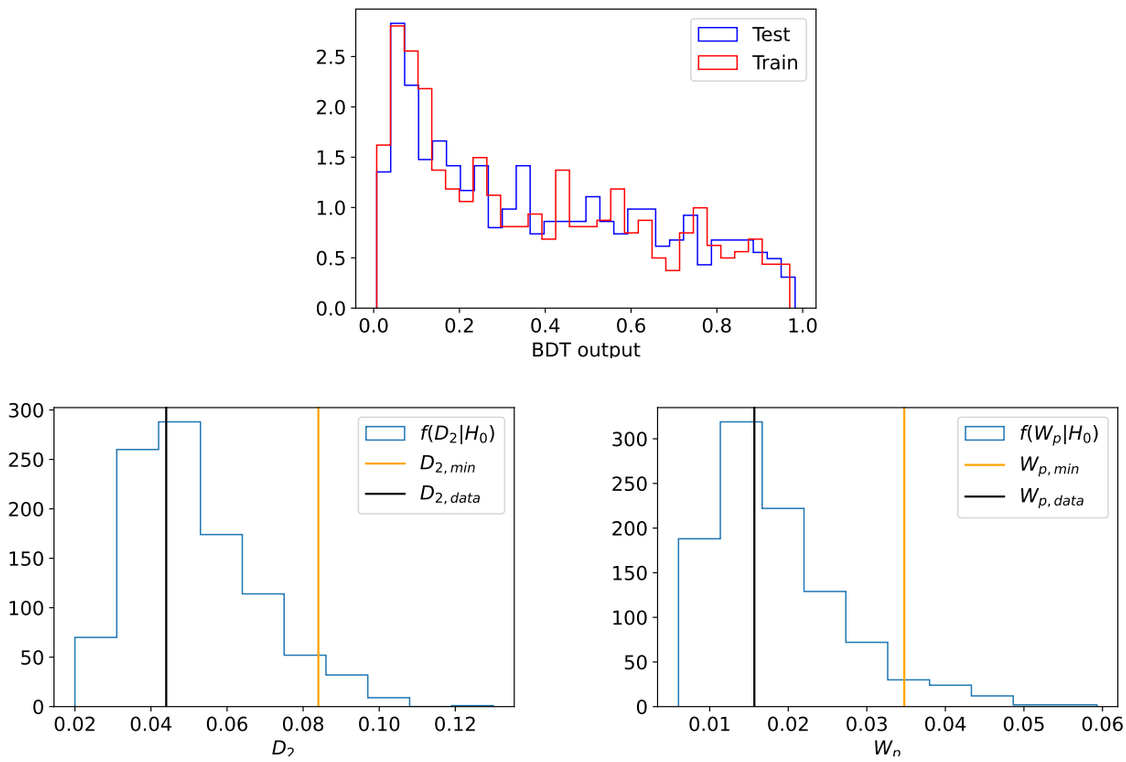

Figure 17: Top: Distribution of datasets used to train the BDT model and to test the model. Bottom-Left: distribution of the D_2 test statistic. Bottom-Right: distribution of the W_p test statistic.

observables is 'are these observables *independent*'. We have already shown that two variables which are uncorrelated (have a correlation value of zero) are not necessarily independent. There are tests for independence that rely on randomly permuting elements of the dataset and comparing the resulting distritubion with what you would expect if the two variables were independent. The notebook **UStatisticPermutationTest.ipynb** shows an example of such a hypothesis using a *U*-statistic based permutation.

5.7 Neyman-Pearson Lemma

Let's go back to the problem of finding the most powerful test of our null hypothesis H_0 against an alternate hypothesis H_1.

As we've seen, this problem involves 3 steps - specifying the size of the test (α), choosing a suitable test-statistic, and choosing a critical region w. If we want to maximise the power, this is the same as choosing the *best critical region*.

Let $\mathbf{D}$ represent our observations or 'experimental data' and suppose it has a p.d.f $f(\mathbf{D}|H(\theta))$, where θ represents a specific hypothesis. We null hypothesis H_0 and $\theta = 1$ represents the alternate hypothesis H_1. Note that I use the notation $f(D|H(\theta))$ to mean that H is the hypothesis for which $f(D|H)$ yields the distribution under which D is distributed – i.e we are assuming H to be true. In other words, this is the same as setting some specific parameter value for θ so that $f(D|H(\theta)) = f(D;\theta)$ but i'll try not to confuse these notations. For our purposes, we can identify two such values of theta with $H_0 \iff \theta_0$ and $H_1 \iff \theta_1$ so that when you see $f(\mathbf{D}|H_0)$, you should think of this as being the same as $f(\mathbf{D};\theta_0)$.

We can think of the test-statistic X as being a function of our data $X(\mathbf{D})$. For a specific critical region w, there will be a region $\mathbf{D} \in w'$ that this function maps to w or $X(w') \to w$ (see figure 18). Since probability is conserved, we must have that $P(\mathbf{D} \in w'|H_0) = P(X \in w|H_0) = \alpha$ and likewise $P(\mathbf{D} \in w'|H_1) = P(X \in w|H_1) = 1 - \beta$.

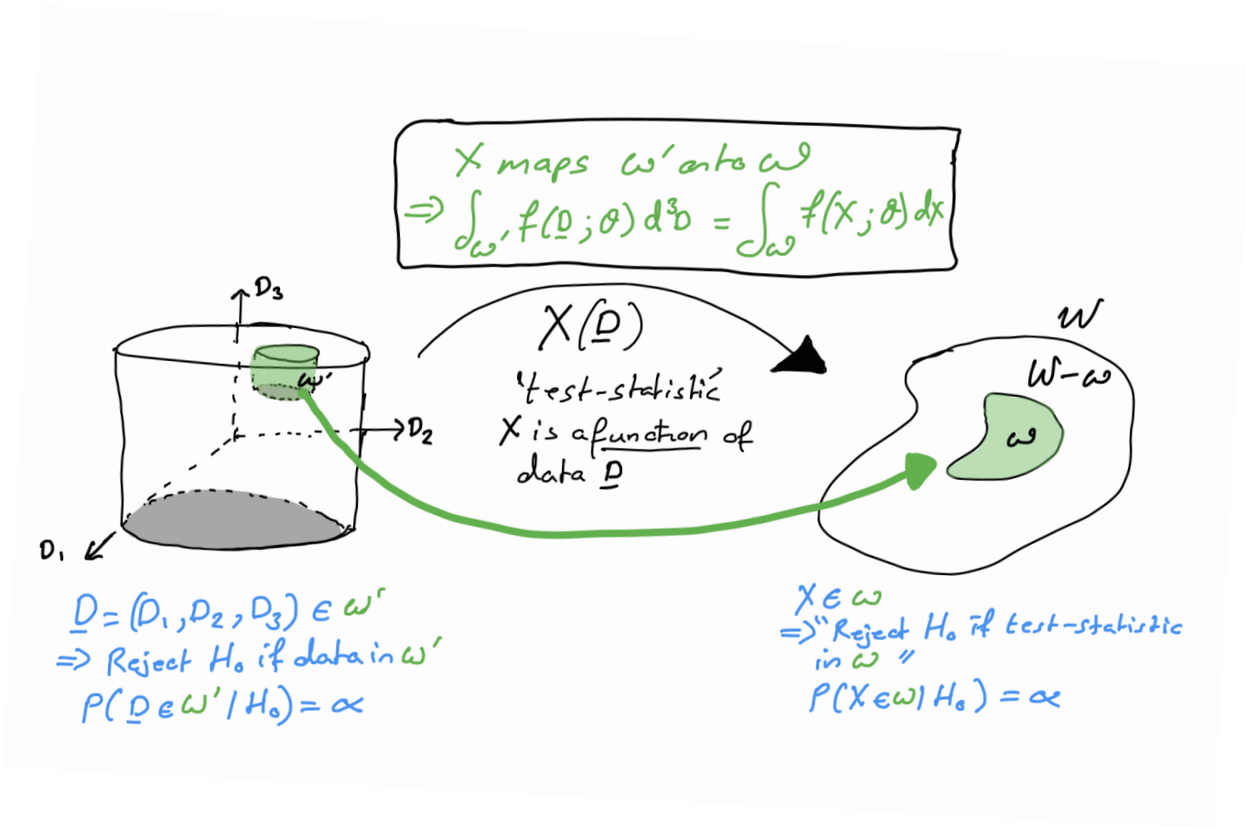

Figure 18: Mapping between the 'data-space', where we observe the dataset $\underline{D}$ and the 'test-statistic space' where the data is summarized by the function $X(\underline{D})$. The region w' in the data-space is mapped onto w by X so all outcomes $\underline{D} \in w'$ are the same outcomes as $X \in w$. Note that here $f(\mathbf{D};\theta)$ is the same as $f(\mathbf{D}|H(\theta))$.

For a specific α then, we choose w' such that,

$$\int_{w'} f(\mathbf{D}|H_0)d\mathbf{D} = \int_{w} f(X|H_0)dX = \alpha, \tag{166}$$

where $f(X|H_0)$ is the probability distribution of our test statistic X under H_0.

We want to find the region w' which maximises $1-\beta$. We can write Eqn. 160 as,

$$\begin{aligned} 1-\beta &= \int_{w'} f(\mathbf{D}|H_1)d\mathbf{D} && (167)\\ &= \int_{w'} \frac{f(\mathbf{D}|H_1)}{f(\mathbf{D}|H_0)} f(\mathbf{D}|H_0)d\mathbf{D} && (168)\\ &= E\left[\frac{f(\mathbf{D}|H_1)}{f(\mathbf{D}|H_0)}\right]_{w'}, && (169) \end{aligned}$$

where the last line denotes the expectation value inside w', under the null hypothesis of the quantity,

$$\Lambda = \frac{f(\mathbf{D}|H_1)}{f(\mathbf{D}|H_0)} = \frac{L(H_1)}{L(H_0)}. \tag{170}$$

This quantity is the ratio of the likelihood function, evaluated under the two hypotheses. The expectation in Eqn. 167 will be maximal when w' is chosen to contain the largest values of Λ.

We could instead choose Λ directly as our test statistic so that $\int_{w'} f(\mathbf{D}|H_1)d\mathbf{D} = \int_{w} f(\Lambda|H_1)d\Lambda$! Then, the best critical region w is the set of points for which $\Lambda \geq c_\alpha \in \mathbb{R}$, where c_α satisfies Eqn. 166. If $\Lambda > c_\alpha$, we would choose H_1, while $\Lambda \leq c_\alpha$ leads us to choose H_0.

Often, given that this test is the most powerful, you will find that in many applications we use Λ as *the* test statistic, even if the hypotheses that we're testing are not simple ones. You should note however that the Neyman-Pearson lemma only applies to simple hypotheses and that Λ is not necessarily the most optimal test statistic for other hypotheses – in practice it turns out to have extensions which are extremely convenient so we still use ratios of likelihood functions ubiquitously in hypothesis testing.

Example: As an example, suppose H_0 is a single Gaussian with $\mu = 0, \sigma = 1$ and H_1 is the sum of two Gaussians with $\mu_1 = \mu + \delta\mu, \mu_2 = \mu - \delta\mu$ and $\sigma_1 = \sigma_2 = \sigma$. For example, if measuring flashes from a light source, H_0 would represent a single source and H_1 two sources separated by $2\delta\mu$, as shown in the figure below.
Take a look at a comparison between the power of the likelihood-ratio based test (LRT) and the 1D KS-test in this example in the **LikelihoodRatioTest.ipynb** notebook. The results, as shown in Figure 20, show that the likelihood ratio based test has a greater power than the KS test in this case (as expected).

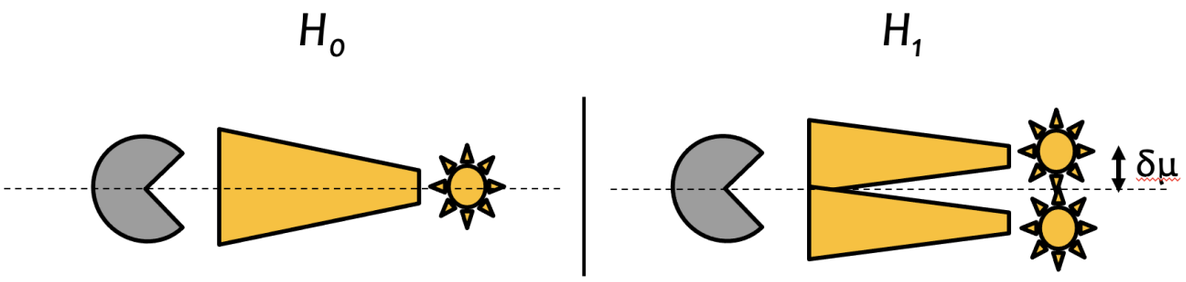

Figure 19: Light pulses detected in a camera (gray pac-mac) can be used to separate the hypothesis of a single light source (H_0) or two light sources H_1 separated by $2\delta\mu$.

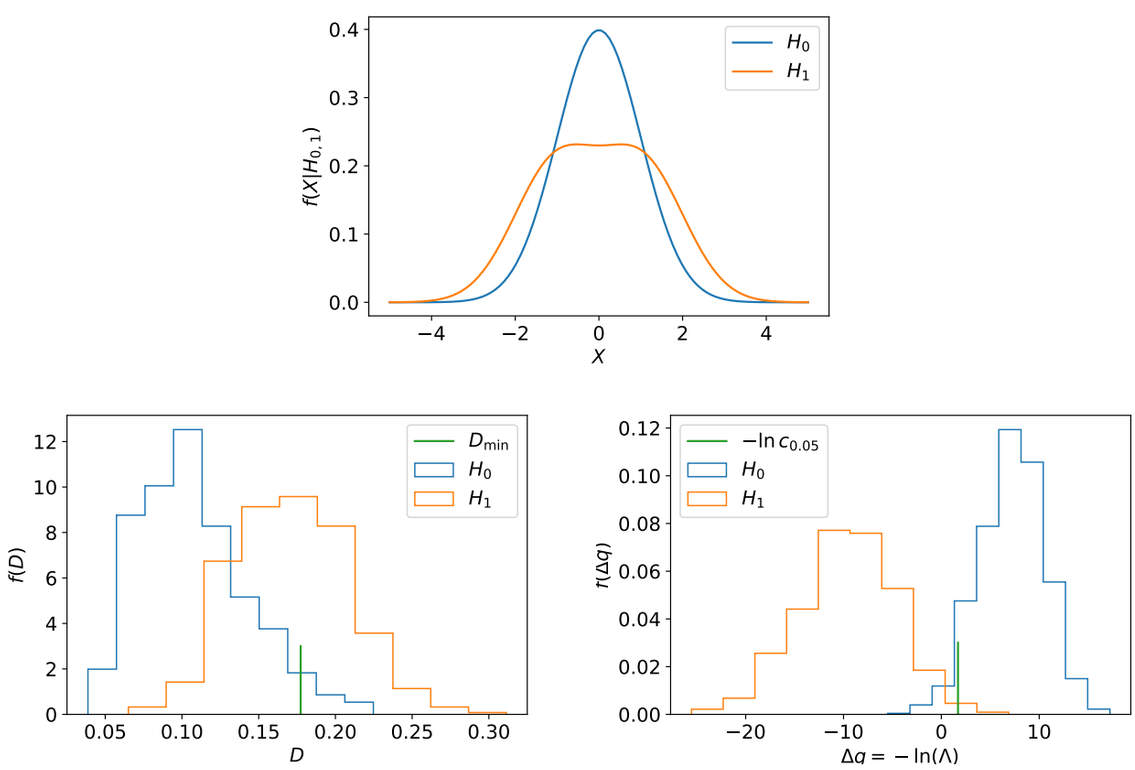

Figure 20: Tests for one Gaussian distribution (H_0 versus two Guassian distributions (H_1) with a difference in μ parameters. Top: probability distributions under the two hypothesis. Bottom-left: distribution of the KS test statistic under each hypothesis and critical value for $\alpha = 0.05$. Bottom-right: Distribution of the LRT test statistic under each hypothesis and critical value for $\alpha = 0.05$.

In the example above, the likelihood-ratio based test was more powerful than the KS-test, as expected by the Neyman-Pearson lemma. We should see the same in general for other hypothesis tests.

Example: Let's take a look at a simple example again where our two hypotheses H_0 and H_1 are represented by Gaussian probability densities $H_0 := \phi(\mu_{H_0}, \sigma_{H_0})$ and $H_1 := \phi(\mu_{H_1}, \sigma_{H_1})$. In the **GaussianHypothesisTests.ipynb** notebook, you will find the power calculated for the KS-test, the WW-runs test and Student's-t test for different values of μ_{H_1} and σ_{H_1} when $\mu_{H_0} = 0$ and $\sigma_{H_0} = 1$. Figure 21 shows the power for each test for different values of μ_{H_1} and σ_{H_1} where the size of the test is always $\alpha = 0.05$. Clearly, the likelihood ratio test is the most powerful in all cases.

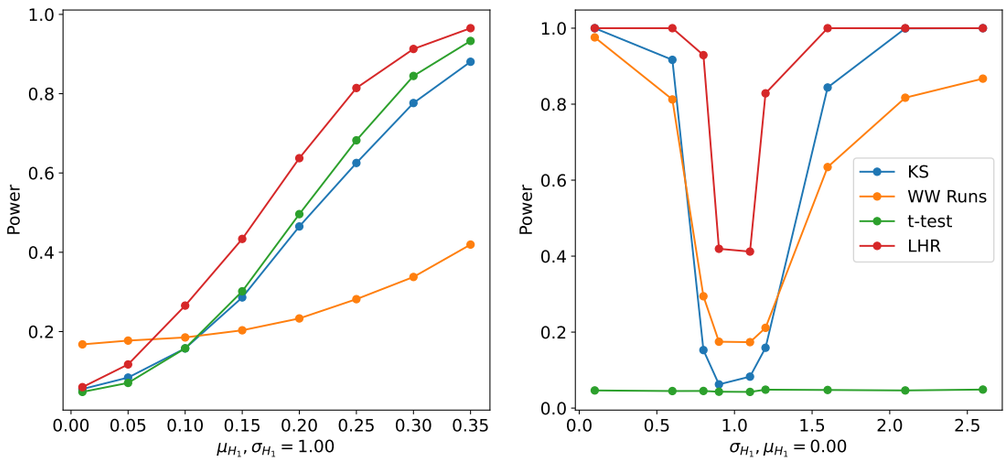

Figure 21: Power calculated for the case of $H_0 := \phi(\mu_{H_0}, \sigma_{H_0})$ and $H_1 := \phi(\mu_{H_1}, \sigma_{H_1})$ for different values of μ_{H_1} (left) and σ_{H_1} (right), and different hypothesis tests. The size of the test is always $\alpha = 0.05$.

5.8 Likelihood-based goodness of fit test

So far, when using the likelihood ratio, we have mostly discussed comparing one hypothesis against another for the purposes of making decisions / choosing between them. However, what if we wanted to just compare the data to a particular hypothesis and make a statement about whether or not the distribution used to describe the data is suitable or not. This problem falls under the class of tests called *goodness of fit tests*. We won't go into detail here as there are many different goodness of fit tests available which are sensitive to different features in the data. Some are good at spotting poor description in the tails of distributions and some are better for giving an overall impression of the agreement. Instead, let's focus on one goodness of fit test which provides another use of ratios of likelihoods.

Suppose we have a histogram of data and a histogram (either derived from some analytic function or from MC simulation) describing what we expect to see under some hypothesis. We can ask how well the hypothesis agrees with that data. Each bin in the histogram can be thought of as an independent Poisson random process. The observed number of events in each bin is labelled o_i and the expected (Poisson mean) values are λ_i. The likelihood function is then the product over Poisson probability distributions,

$$L(H_0) = \prod_i (\lambda_i)^{o_i} e^{-\lambda_i}. \tag{171}$$

Note that we've ignored the $o_i!$ terms, since in the end, this constant will drop out of the ratio of likelihoods.

What if we haven't specified an alternate hypothesis or don't have a particular one in mind. In this case, we can construct an alternate hypothesis H_1 defined by setting $\lambda_i = \hat{\lambda}_i = o_i$ - i.e we use the *maximum likelihood estimators* for the Poisson parameters λ_i to specify H_1. This is often referred to as the *saturated model* as it is the one for which the maximum likelihood possible, given the data observed, is obtained. We then have,

$$\Lambda = \frac{L(H_0)}{L(H_1)} = \frac{\prod_i (\lambda_i)^{o_i} e^{-\lambda_i}}{\prod_i (o_i)^{o_i} e^{-o_i}}. \tag{172}$$

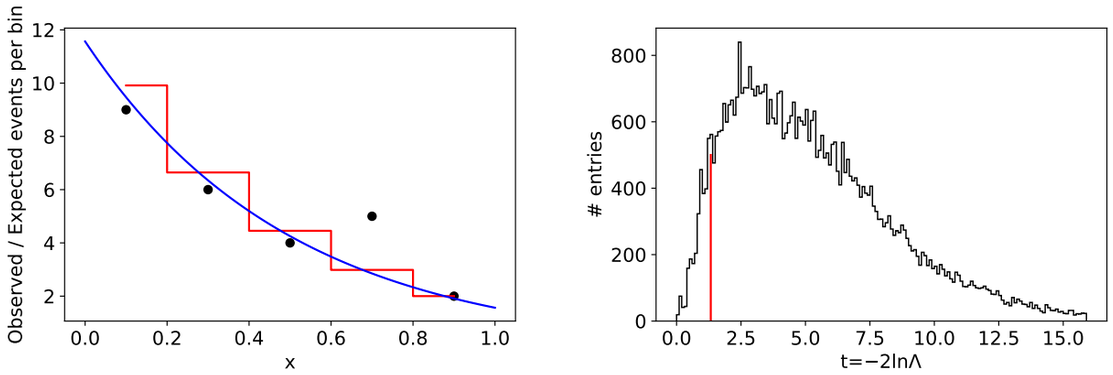

Figure 22: Goodness of fit test for an exponential hypothesis (H_0) using the saturated model as the alternate hypothesis (H_1). The left plot shows the distribution observed in data (black markers) and the expectation under H_0 (red markers), while the right shows the distribution of $-2\ln\Lambda$ under H_0 (black histogram) and the observed value (red line). The cyan histogram shows the region integrated to calculate the p-value.

In our previous example, the values of the likelihood function were $\mathcal{O}(1)$, however if we have many bins, these values will start to get very very small and hence numerically difficult. A common solution in HEP is to instead take the natural log of the likelihood - in fact (for reasons that will become clear later), we usually take $-2\times$ the log of the likelihood. The product becomes a sum, and we have,

$$-2\ln\Lambda = -2\ln\frac{L(H_0)}{L(H_1)} = -2\sum_i \left(o_i \ln(\lambda_i) - \lambda_i - o_i\ln(o_i) + o_i\right) \tag{173}$$

Example: Suppose we measure the energy of outgoing particles that scatter off a target. We know their incoming energy and hence we measure the distribution of the fraction of that energy which is carried away by the scattered particle. Let the energy fraction be $x \in [0,1]$ and our hypothesis for the distribution is an exponential with slope parameter -2,

$$f(x) = Ne^{-2\cdot x},\ N = 2(1-e^{-2})^{-1} \approx 2.313 \tag{174}$$

We bin the observable x into 5 equal width bins in the range $[0,1]$ and count the number of events that fall into each bin. We observe the following counts in the data 9, 6, 4, 5 and 2. Given H_0, the expected counts can be determined as $\lambda_i = 26 \times \int_{x_{a_i}}^{x_{b_i}} f(x)dx$, where $[x_{a_i}, x_{b_i}]$ defines the boundaries of bin i and we've observed 26 events in total. Plugging in our observation, we see a value of $-2\ln\Lambda = 1.32$. We can calculate the distribution of $-2\ln\Lambda$ and hence a p-value of ≈ 0.94.
Take a look at the **GoodnessOfFit.ipynb** notebook where this is done using MC simulation. It's important to remember that in the test statistic, we use the toy data in place of the observation as both the *data* and as the *saturated model*.

5.9 p-values and Significance tests

Remember that the frequentist procedure for hypothesis testing is based on the following steps,

1. Define a test statistic $t \in \mathbb{R}$ that summarizes the observations and has some separation power between H_0 and H_1.
2. Define a critical region w such that, $\int_w f(t|H_0) = \alpha$, where α (the size of the test) is some pre-defined value between 0 and 1.
3. Determine the value of t in the observed dataset - t_{obs}.
4. Reject H_0 if $t_{\text{obs}} \in w$.

You will have heard of the concepts of 'significance' and '$p-$values' but so far these have not been featured in our procedure. $p-$values are often misrepresented in literature and this has led to them

being given a bad reputation. However, with the knowledge of what they are and how to use them, they can be very useful. A $p-$value is defined simply as a tail probability under a specific hypothesis - usually the *null* hypothesis (see cartoon in Figure 23),

$$p = \int_{t_{\rm obs}}^{+\infty} f(t|H_0)dt. \tag{175}$$

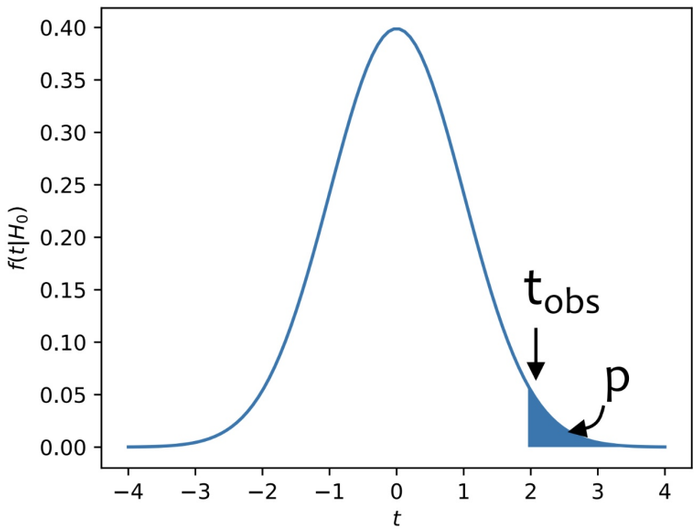

Figure 23: Cartoon of a $p-$value calculated from the observed value of the test-statistic t and its distribution $f(t|H_0)$ under the null hypothesis.

Be very careful to note that the $p-$value is not the probability that H_0 is true, nor is it the probability that you would reject the null hypothesis (that is the size α). Instead, the $p-$value is simply a random variable that *depends on the observed data* and is uniformly distributed between 0 and 1 under the null hypothesis. To see this, lets look at the following. We can think of the $p-$value as a probability,

$$\begin{aligned} p = P(t > t_{\rm obs}|H_0) &= 1 - P(t < t_{\rm obs}|H_0) & (176)\\ &= 1 - F(t), & (177)\end{aligned}$$

where $F(t)$ is the cumulative distribution. Since $F(t)$ is a monotonically increasing function, if $t > t_{\rm obs}$ it must also be the case that $F(t) > F(t_{\rm obs})$. This means that,

$$\begin{aligned} 1 - F(t) = P(t > t_{\rm obs}|H_0) &= P(F(t) > F(t_{\rm obs})|H_0) & (178)\\ &= 1 - P(F(t) < F(t_{\rm obs})|H_0) & (179)\end{aligned}$$

and so,

$$P(F(t) < F(t_{\rm obs})|H_0) = F(t), \tag{180}$$

for any value of $t_{\rm obs}$. This means $F(t)$ is uniform under the null distribution, which also means that $p = 1 - F(t)$ must also be uniform under the null distribution.

We often convert $p-$values into "Z-scores" (or sometimes referred to as the significance) to yield simpler numbers. This is a simple conversion which makes use of the standard normal distribution $\phi(X; 0, 1)$ and Z is determined from,

$$p = \int_{-Z}^{+Z} \phi(X; 0, 1)dX. \tag{181}$$

For example, when $p \approx 0.003, Z = 3$ and when $p \approx 0.046, Z = 2$. This is just a convention to convert the $p-$value into more sensible numbers.

Trials factors Have a look at Figure 24. Each of the points represents a measurement of a Gaussian distributed random variable, each with the same σ. Suppose our null hypothesis is represented by the dashed line at 0 and we want to test whether or not our measurements are consistent with this null hypothesis. Using the LHR test, we would in fact obtain a $p-$value of 0.0001, so we might consider the result circled in red significant. However, would we also equally care about any other measurement being far from zero. The probability to observe any of the points being significantly far from zero is larger than the probability of this one particular measurement being far from zero.

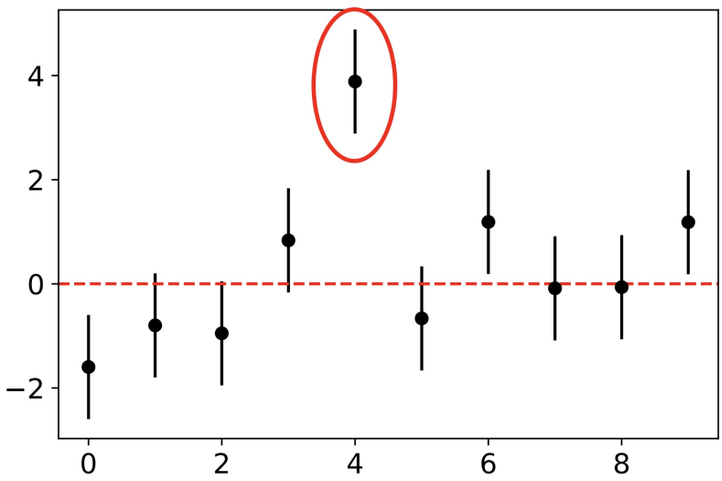

Figure 24: Ten measurements of Gaussian distributed random variables. The null hypothesis is represented by the red dashed line at zero, while the measurement that most significantly deviates from the null hypothesis is circled in red.

We want to calculate the probability to observe any result that is as significant as the one we see which is the same as the probability that the *smallest* $p-$value is smaller than the one we observe. In Figure25, the left panel shows the distribution of p. As we expect, it is a uniform distribution. However, on the right, we see the distribution of the minimum p calculated for any point which is clearly not uniform.

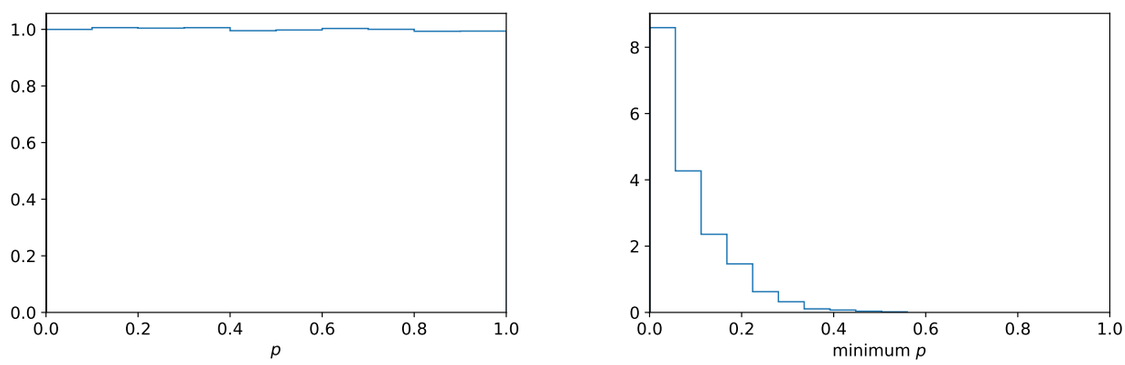

Figure 25: Left: Distribution of $p-$value for the 5th measurement where the largest deviation from zero is observed. Right: Distribution of the minimum $p-$value for any of the measurements.

To calibrate for this, we need to introduce a trials factor T defined as the ratio of probabilities,

$$T = \frac{P(\text{minimum } p < p_{\text{obs}})}{P(p < p_{obs})}. \tag{182}$$

It can be shown that for a set of n independent measurements, the trials factor will be equal to $T = n$. In our case, the trials factor is 10. Generally speaking, the more independent measurements made, the more likely it is to see a significant result and this must be accounted for.

Week 6 Uncertainty Intervals

In the previous section, we introduced the likelihood function for use in hypothesis testing and in particular rejecting hypotheses which fail some criteria based it and for estimating parameter values of probability distributions and models of nature.

We usually report measurements of a single parameter as $X \pm \sigma_X$. There are a number of ways one can obtain X and σ_X here – for example, it could refer to the sample mean $\bar{X}$ and standard deviation $S = \sqrt{\sum_i (X_i - \bar{X})^2}$ or it could refer to the moments of the distribution $f(X)$, μ_1, $\sqrt{\nu_2}$. A very large part of what we do in data analysis, is to *improve on the sensitivity* of some measurement, which often involves designing experiments to measure physical quantities more precisely. Two competing experiments may claim to have measured $X_1 \pm \sigma_{X,1}$ and $X_2 \pm \sigma_{X,2}$ but without knowing what they mean, we can't compare those numbers fairly. A good comparison can be made for measurements for which two properties are well understood, namely the *bias* of the measured value and the *coverage* of the interval. We'll discuss what we mean by confidence intervals but first, we need a quick aside about categorising parameters based on whether we care or not about measuring them.

6.1 Neyman construction and confidence intervals

The frequentist approach to reporting uncertainties on measurements is to consider the uncertainty itself as a random variable. The freqentist will report an *interval* (or region in the case of more than one parameter of interest) of values of X say $X_l \leq X \leq X_u$, at a specified *confidence level* $(1 - \alpha)$. Knowing that X_l and X_u are random variables, the frequentist knows only that in an ensemble of such intervals, the fraction of intervals that contain the true value X_0 will be $(1 - \alpha)$. Note that this says nothing about whether or not a particular interval contains the true value but is only a statement about the ensemble of intervals! When a scientist reports s $X \pm \sigma_X$, the σ_X might actually be written $^{+(\theta_u - X)}_{-(X - \theta_l)}$ and X_u and X_l will usually correspond to the endpoints of the 0.683 (or 68.3%) confidence level interval. Note that this means, if we have a composite hypothesis $H(X)$, the statement of the interval corresponds to excluding all hypotheses in the set $\{H(X) : X \notin [X_l, X_u]\}$ with an error rate of at most α. The *coverage* of a particular method to obtain those intervals is the actual fraction of intervals which contain the true value – eg a reported 95% confidence level interval might only cover with a fraction 0.93, meaning it *under covers*. We'll see later that some methods only *cover* in certain circumstances. There is a method designed to cover in all (nearly all) circumstances, known as the Neyman construction. This is explained by way of example.

Example: Consider the case of estimating the temperature T of the fusion reactor at the centre of the sun by using an estimate of the solar neutrino flux ϕ (the test statistic) from one month's data from a large solar neutrino detector. Take a look at Figure 26. The Neyman construction uses $P(\phi|T)$ at a given value of T to choose a region in ϕ that constitutes $1 - \alpha$ of the probability distribution. Note that this is *a* region, rather than *the* region, as there are many possible regions containing the required fraction of outcomes. Thus the Neyman construction can be used to produce central regions, upper limits, lower limits, etc, simply by changing the "ordering rule" used for accumulating different test statistic values up to the required confidence level. in ϕ for which the fraction of outcomes in that region is a pre-determined level $1 - \alpha$, for example 90%.

This is repeated for all values of T to produce the shaded confidence band, showing the likely[a] values of the data for any value of the parameter of interest. Then we collect one month's data, from which we deduce a flux ϕ_d. The intersection of a vertical line at ϕ_d with the confidence band then gives the frequentist range for the parameter T, $T_l \leq T \leq T_u$, at the chosen confidence level.

[a] We refer to the chosen values of ϕ as "likely values", as they are the ones selected for a particular ordering rule.

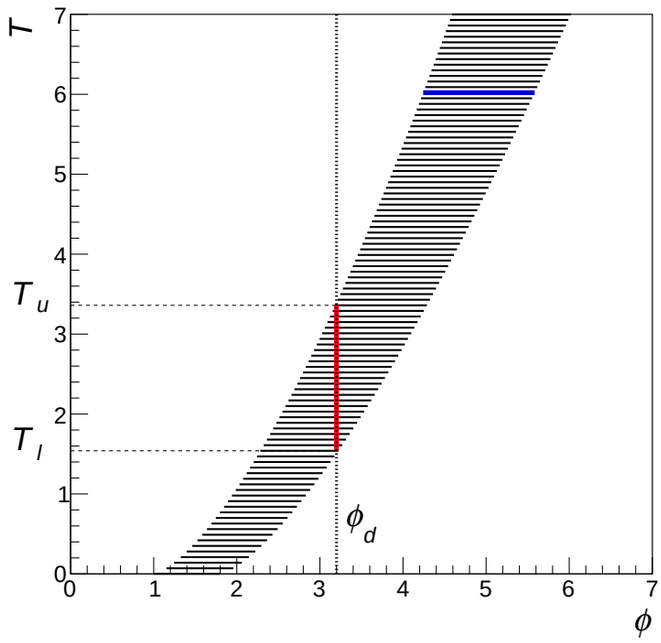

Figure 26: 302 The Neyman construction. The horizontal bands gives likely values (at the 90% level) of the test statistic ϕ (flux) for each value of the theory parameter T (temperature). The observed value of the test statistic ϕ_d is indicated by the vertical dotted line. This uses only the probability of different data, for given values of T; it does not involve probabilities of different values of T. A dotted vertical line at ϕ_d intersects the edges of the horizontal lines at T_l and T_u, and these define the frequentist range for T (indicated by the thick red vertical line). The value of $T = 6$ in this case lies outside of the interval as the 90% likely values of ϕ (indicated by the thick blue horizontal line) do not overlap with the observed value ϕ_d.

6.1.1 Likelihood ratio test-statistic

In the Neyman construction, one must choose an ordering principle to select outcomes to include in the bands. Very often, the likelihood ratio is used to select the "likely values" to be included. In modern applications, this is particularly useful since it allows for a natural way to include nuisance parameters through the use of *profiled* likelihoods (we'll see this later). Moreover, rather than considering a sampled value ϕ as the test statistic and order using profile likelihood ratios, we can just use the profile likelihood ratio as the test statistic and as the ordering principle. This also very naturally extends to more than one parameter of interest. The procedure is as follows;

- Define the test statistic using the profiled likelihood function as,

$$\zeta_\theta = 2(q(\theta) - q(\hat{\theta})) \tag{183}$$

  where as usual θ are the parameters and $\hat{\theta}$ means the value(s) of θ that minimize q.

- For each value of θ (representing the hypothesis $H(\theta)$), calculate $\zeta_\theta^{\text{obs}}$ for the observed data and generate toy data, to calculate the distribution $f(\zeta_\theta|H(\theta))$.

- Choose a confidence level $(1-\alpha)$ and select the values of θ for which,

$$p_\theta = \int_{\zeta_\theta^{\text{obs}}}^{+\infty} f(\zeta_\theta|H(\theta))d\zeta_\theta \geq \alpha. \tag{184}$$

  The union of all of these values forms the $100 \times (1-\alpha)\%$ confidence region (interval for 1 parameter). Note that the ordering principle that we've used to calculate p_θ is the same as the Feldman-Cousins ordering rule (in the case of no nuisance parameters) and Kendall and Stuart (with nuisance parameters).

Example: Let's go back to the example of a radioactive decay, this time we'll determine the 68% confidence interval on the parameter τ. You can follow the **FrequentistIntervals.ipynb** notebook for this. It's a bit slow if you run it since it is throwing lots of toys. Figure 27 shows the distribution $f(\zeta_\tau|H(\tau))$ for fixed values of τ. Also shown is the *observed value* ζ_τ^{obs} – unlike the solar temperature example, this is not a single number but of course depends on τ. Finally, also plotted is the value of ζ_τ (called ζ_τ^{68}) which is larger than 68% of the toys in the distribution. This means that the values of τ to be included in the interval are those for which $\zeta_\tau^{\text{obs}} < \zeta_\tau^{68}$. We need to keep track of which τ values meet the criteria to construct the interval.

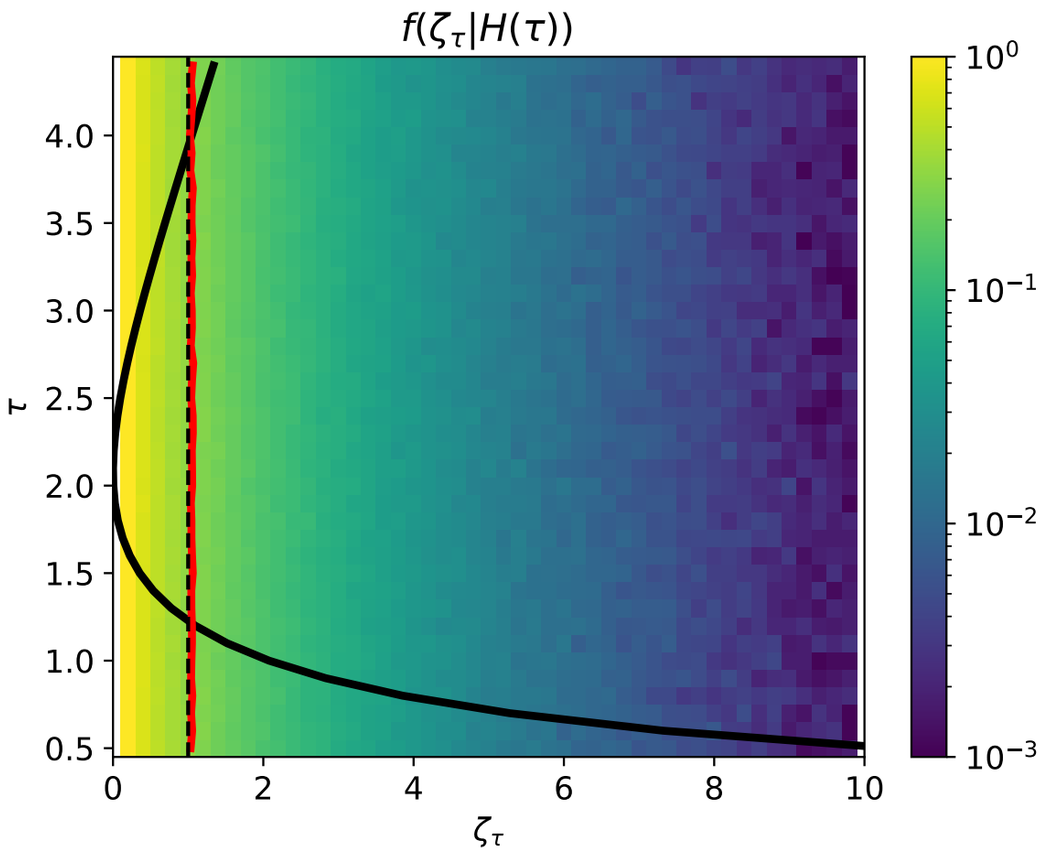

Figure 27: Distributions $f(\zeta_\tau|H(\tau))$ for different values of τ are shown by the color scale. The white regions show where there are no entries, highlighting the discrete nature of the Poisson probability distribution. The black like shows the observed values ζ_τ^{obs} and the red line shows ζ_τ^{68} – the value that 68% of the distribution is smaller than. The values of τ for which $\zeta_\tau^{\text{obs}} < \zeta_\tau^{68}$ form the 68% confidence interval.

We see that there is a particular value for our ζ_τ^{68}, which seems to be *independent* of τ. This turns out to be a common feature and in fact for *any* likelihood function, as we'll seem the central limit theorem tells us how the quantity ζ_τ should be distributed under the hypothesis that τ is the true value in the limit of large numbers. This means there is no need to determine the ζ_τ^{68} values using toys or even to determine the distribution at all with toys – this is a convenient result (theorem) by *Wilks'*.

6.2 Wilks' theorem

Wilks' theorem provides a very powerful result that allows us to calculate the distribution of ζ_θ, under certain conditions, in the limit of a large sample size. The most important of these conditions that we sometimes ignore is that the maximum likelihood estimate must not be at or beyond a boundary of the parameter space. In our counting experiment, this means that introducing the condition $\mu > 0$, will have implications when the maximum likelihood estimate is close to 0. There are other conditions, but these are almost always satisfied so there's no need to cover them.

Let's start with the simplest case where we have a single parameter of interest θ with no nuisance

parameters. As usual, for N observations of a random variable $X \sim f(X;\theta)$, we define,

$$\zeta_{N,\theta} = 2(q_N(\theta) - q_N(\hat{\theta})) = -2\sum_{i=1}^{N} \ln f(X_i;\theta) + 2\sum_{i=1}^{N} \ln f(X_i;\hat{\theta}). \tag{185}$$

Providing that the derivatives of f exist, we can Taylor expand the first derivative of q_N to approximate the value at $\hat{\theta}$,

$$0 = \left.\frac{dq_N}{d\theta}\right|_{\theta=\hat{\theta}} = q'_N(\hat{\theta}) = q'_N(\theta) + q''_N(\theta)(\hat{\theta} - \theta), \tag{186}$$

where the first equality is due to the fact that $\hat{\theta}$ is the value of θ that maximises the likelihood function q_N. We can re-write the equation as,

$$(\hat{\theta} - \theta)\sqrt{q''_N(\theta)} = -\frac{q'_N(\theta)}{\sqrt{q''_N(\theta)}}, \tag{187}$$

which looks like an odd thing to do but it will be useful. Looking at the numerator on the RHS, we have,

$$-q'_N(\theta) = \frac{d}{d\theta}\sum_{i=1}^{N} \ln f(X_i;\theta) = \sum_{i=1}^{N} \frac{d}{d\theta}\left(\ln f(X_i;\theta)\right). \tag{188}$$

Now let's define the random variable $u = \dfrac{d}{d\theta}\ln f(X|\theta)$, then we have that $-q'_N(\theta)$ looks like a sample mean of this random variable. In fact $-q'_N(\theta) = N\bar{u}$, where $\bar{u}$ is the sample mean of u. The expectation of u can be found by considering that for any θ,

$$1 = \int f(X;\theta)dX, \tag{189}$$

$$\tag{190}$$

and therefore, taking derivatives,

$$0 = \frac{d}{d\theta}\int f(X;\theta)dX = \int \frac{d}{d\theta} f(X;\theta)dX = \int \frac{d}{d\theta}\left(\ln f(X;\theta)\right)\cdot f(X;\theta)dX \tag{191}$$

$$= E\left[\frac{d}{d\theta}\ln f(X;\theta)\right]_\theta = E\left[u\right]_\theta, \tag{192}$$

and hence the expectation of u under $f(X;\theta)$ is 0. We can start to form a quantity T_N, given by,

$$T_{N,\theta} = \frac{N\bar{u} - NE[u]_\theta}{\sqrt{N\cdot V(u)_\theta}} \tag{193}$$

We want to see what happens in the limit of $N \to \infty$, to the quantity $\sqrt{q''(\theta)}$. If its proportional to $\sqrt{N\cdot V(u)_\theta}$, then we have, by the central limit theorem, that T_θ (the limit of $T_{N,\theta}$) is distributed as $T_\theta \sim \phi(T_\theta;0,1)$.

Let's take a look at $V(u)_\theta$. We have by definition that,

$$V(u)_\theta = E\left[(u - E[u]_\theta)^2\right]_\theta = E\left[(u)^2\right]_\theta \tag{194}$$

$$= \int \left(\frac{d}{d\theta}\ln f(X;\theta)\right)^2 f(X;\theta)dX. \tag{195}$$

Recall again that,

$$1 = \int f(X;\theta)dX \tag{196}$$

and differentiating twice we have,

$$0 = \int \frac{d}{d\theta}\left(\frac{d}{d\theta}\ln f(x;\theta)\cdot f(X;\theta)\right)dX \tag{197}$$

$$= \int \frac{d^2}{d\theta^2}\ln f(X;\theta)\cdot f(X;\theta)dX + \int \frac{d}{d\theta}\ln f(X;\theta)\cdot\frac{d}{d\theta} f(X;\theta)dX \tag{198}$$

$$= \int \frac{d^2}{d\theta^2}\ln f(X;\theta)\cdot f(X;\theta)dX + \int \frac{d}{d\theta}\ln f(X;\theta)\cdot\frac{d}{d\theta}\ln f(X;\theta)\cdot f(X;\theta)dX \tag{199}$$

$$= \int \frac{d^2}{d\theta^2}\ln f(X;\theta)\cdot f(X;\theta)dX + \int \left(\frac{d}{d\theta}\ln f(X;\theta)\right)^2 f(X;\theta)dX \tag{200}$$

$$= \int \frac{d^2}{d\theta^2}\ln f(X;\theta)\cdot f(X;\theta)dX + V(u)_\theta, \tag{201}$$

so that,

$$V(u)_\theta = -\int \frac{d^2}{d\theta^2} \ln f(X;\theta) \cdot f(X;\theta) dX = -E\left[\frac{d^2}{d\theta^2} \ln f(X;\theta)\right]_\theta . \tag{202}$$

But remember that $q''_N(\theta) = -\sum_{i=1}^{N} \frac{d^2}{d\theta^2} \ln f(X;\theta)$. By the law of large numbers, we must have that $\frac{-q''_N(\theta)}{N} \to E\left[\frac{d^2}{d\theta^2} \ln f(X;\theta)\right]_\theta$ as $N \to \infty$. But this is just $-V(u)_\theta$! So then we have $\frac{q''_N(\theta)}{N} \to V(u)_\theta$. Now (in a rather hand-wavy fashion), we can say that,

$$(\hat{\theta} - \theta)\sqrt{q''_N(\theta)} = -\frac{q'_N(\theta)}{\sqrt{q''_N(\theta)}} \to \frac{q'_N(\theta)}{\sqrt{N \cdot V(u)_\theta}} = \frac{N\bar{u} - NE[u]_\theta}{\sqrt{N \cdot V(u)_\theta}} = T_{N,\theta} \to T_\theta \sim \phi(T_\theta; 0, 1). \tag{203}$$

We've been a bit careless here since the numerator converges in distribution and the denominator converges in probability, and we've performed the two limits separately. However, but there is a theorem (which we won't go into) for ratios of such random variables, numerator converging in distribution and denominator in probability, due to Slutsky which yields the same result so in the end its ok.

Let's go back to the definition of $\zeta_{N,\theta}$. Again, we take a Taylor expansion, this time of the function $q_N(\theta)$,

$$q_N(\theta) = q(\hat{\theta}) + \cancelto{0}{q'_N(\hat{\theta})}(\theta - \hat{\theta}) + \frac{1}{2} q''_N(\hat{\theta})(\theta - \hat{\theta})^2, \tag{204}$$

and so,

$$\zeta_{N,\theta} = 2(q_N(\theta) - q_N(\hat{\theta}))) = q''_N(\hat{\theta})(\theta - \hat{\theta})^2 = \left((\hat{\theta} - \theta)\sqrt{q''_N(\hat{\theta})}\right)^2 . \tag{205}$$

Now the term inside the square on the RHS of Eqn 205 is the same as in the LHS of Eqn. 203, except for the value of θ at which the second derivative is evaluated. However, since the maximum likelihood estimate is *consistent*, $\hat{\theta} \to \theta_0$ as $N \to \infty$. So in the limit $N \to \infty$, we replace θ with θ_0, and study the distribution of ζ_θ for the true value of $\theta = \theta_0$. In this limit we can then equate the LHS of Eqn. 203 with the term inside the square, so that,

$$\zeta_{N,\theta_0} = (T_{N,\theta_0})^2. \tag{206}$$

But we know that T_{N,θ_0} is distributed as a unit normal $\phi(T; 0, 1)$ under the hypothesis $H(\theta_0)$ *for any value of* θ as $N \to \infty$ so we can also figure out the distribution of $(T_{N,\theta_0})^2$. This distribution is known as a χ^2 – a *chi-square* distribution with 1 degree of freedom. It has the probability density function,

$$\chi^2(X; 1) = \frac{1}{\sqrt{2\pi X}} e^{-\frac{X}{2}}. \tag{207}$$

This is the result of Wilks' theorem for one parameter. In general, Wilks' theorem gives us the result for any number of degrees of freedom. The result is that for a log-likelihood difference with n parameters $\theta_1, \theta_2, ..., \theta_n$, the test statistic $\zeta_{\theta_1,...,\theta_n}$ will be distributed under $H(\theta_1, ..., \theta_n)$ (the null hypothesis) as,

$$f(\zeta_{\theta_1,...,\theta_n} | H(\theta_1, ..., \theta_n)) = \chi^2(\zeta_{\theta_1,...,\theta_n}; n), \tag{208}$$

where $\chi^2(\cdot; n)$ is the chi-square distribution with n degrees of freedom, in the limit of large sample sizes. If we know the distribution of ζ, then we can calculate p_θ for any value of ζ_θ by looking up the cumulative distribution function of the χ^2 functions. Moreover, the value of ζ_θ^{68} (and any quantile, not just the 68% one) is independent of θ!. Any one of your favourite statistics programming languages will be able to calculate this for you, for example as we saw way back in the early lectures, in python, we can use the `scipy.stats.chi2` class and use the `chi2.cdf` function.

Calculating intervals or confidence regions for parameters becomes straightforward since for each value, we know the distribution of ζ_θ under the hypothesis $H(\theta)$ simply by knowing the dimension of θ and calculating $\zeta_\theta^{\text{obs}}$. The $(1 - \alpha)$ confidence region in $n-$dimensions is determined by the values of θ for which,

$$\zeta_\theta^{\text{obs}} = q(\theta) - q(\hat{\theta}) = \Delta q(\theta) \leq Q \tag{209}$$

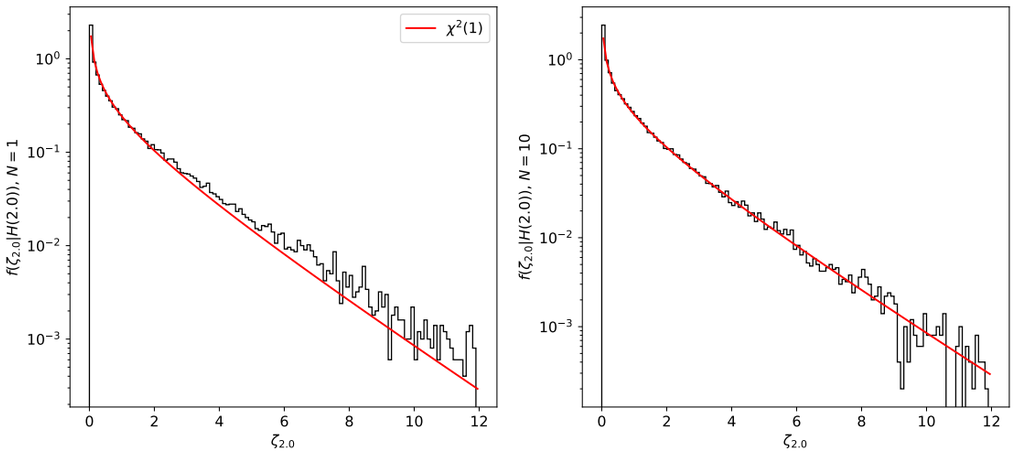

Figure 28: Distribution of ζ_τ for $\tau = 2.0$ with a sample size of $N = 1$ (left) and $N = 10$ (right), for the exponential decay. The black histograms are the real distribution, evaluated using toys, while the red line is the approximate distribution using Wilks' theorem.

where $\int_Q^{+\infty} \chi^2(\zeta_\theta; n) d\zeta_\theta = \alpha$. For example, if $n = 1$ and $(1 - \alpha) = 0.683$, then $Q = 1$, while if $n = 1$ and $(1 - \alpha) = 0.954$, then $Q = 4$. In our goodness of fit test, we could have used this result to calculate the $p-$value and our number of degrees of freedom would be the number of bins.

Going back to our example of an exponential decay, we can look at the distribution of ζ_{tau} and compare to what Wilks' theorem predicts. Figure 28 shows the distribution of ζ_τ assuming $\tau = 2.1$ for two different sample sizes $N = 1$ and $N = 10$. For the larger value, the approximation of a $\chi^2(1)$ is much more accurate.

6.2.1 Coverage in a counting experiment

Let's think about a simple counting experiment where our random variable is an integer n - $n \sim \frac{\mu^n}{n!} e^{-\mu}$ - and we want to measure the Poisson parameter μ. Close to the boundary $\mu = 0$, we have two issues with applying Wilks' theorem. The first is not being in the limit of large N and the second being the boundary itself. We can check what the coverage of the method (say for the 68.3% interval) by determining the *fraction of intervals* in μ, as a function of the true value μ_0, that contain μ_0. It sounds like a rather painful ordeal given that calculating a single interval can take time, however, we do not need to calculate each interval to figure out the coverage. Remember that μ is included in the interval provided $\zeta_\mu^{\text{obs}} \leq \zeta_\mu^{68.3}$. For the Neyman construction, we can use toys to calculate $\zeta_\mu^{68.3}$, while for the `MINOS` method, we assume $\zeta_\mu^{68.3} = 1$. Figure 29 shows the coverage in μ when calculating the intervals using the Neyman construction and using the `MINOS` method for our counting experiment. You can see that the Neyman construction gives very close coverage to the desired 68.3% except at very small μ, where the discrete nature of the Poisson distribution makes it difficult to find the exact $\zeta_\mu^{68.3}$ – in fact this is still seen at larger values, though the effect gets reduced. Instead, the `MINOS` method, jumps between over-coverage and undercoverage, eventually settling down only above for larger values of μ. This is not surprising since the distribution of ζ_μ really doesn't look like a $\chi^2(1)$.

6.2.2 Multiple Dimensions

Let's look at an example of trying to measure two signal rates over a falling background distribution, using a binned likelihood.

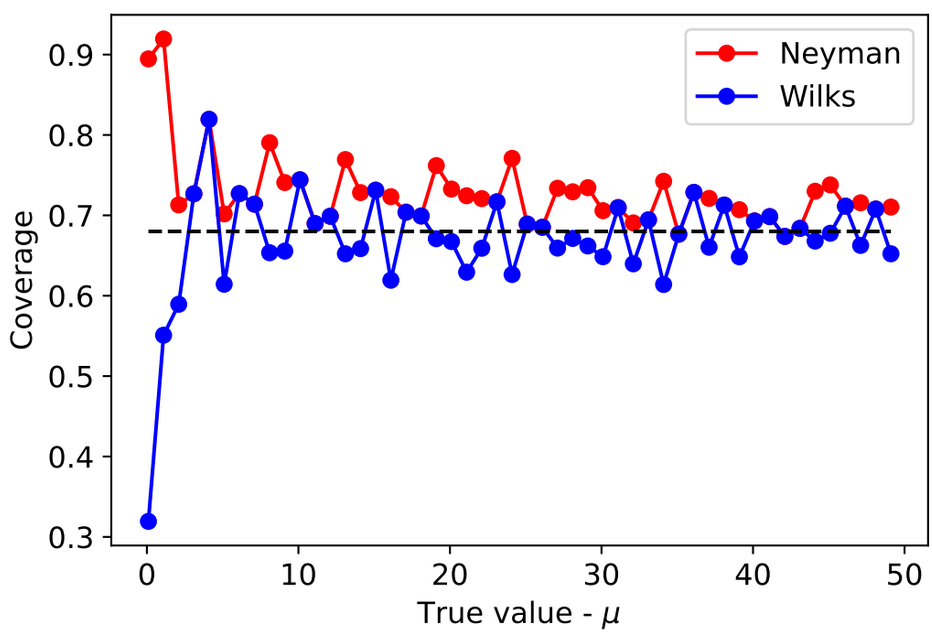

Figure 29: Coverage of intervals in μ for the counting experiment, when calculated using the Neyman construction compared to the `MINOS` method, as a function of the true value μ.

Example: We have an analysis in which the number of events in bins of a distribution is counted. There is a background process and two signal processes with signal strength modifiers μ_1 and μ_2, with no restrictions on their range. Our log-likelihood function for this will be,

$$q(\mu_1, \mu_2) = -\left(\sum_i n \ln \lambda_i - \lambda_i\right), \tag{210}$$

where i runs over the bins, $\lambda_i = \mu_1 s_{i,1} + \mu_2 s_{i,2} + b_i$.
Take a look at the **WilksTheorem2D.ipynb** notebook where the model is defined. Figure 30 shows the background and signal distributions at the global minimum $(\hat{\mu}_1, \hat{\mu}_2)$. The data is shown on top with each data point indicating the standard deviation of a Poisson distribution with the same mean as the observed data in that bin. We scan the value of,

$$2\Delta q(\mu_1, \mu_2) = 2(q(\mu_1, \mu_2) - q(\hat{\mu}_1, \hat{\mu}_2)). \tag{211}$$

Figure 31 shows the value using a color scale. Using Wilks' theorem, the 68.3% and 95.4% confidence regions are the regions for which $2\Delta q(\mu_1, \mu_2) < 2.3$ and $2\Delta q(\mu_1, \mu_2) < 5.99$ respectively, as indicated by the contours. If we are interested in only the first parameter, μ_1, then we can use the function,

$$2\Delta q(\mu_1) = 2(q(\mu_1, \hat{\mu}_{2,mu_1}) - q(\hat{\mu}_1, \hat{\mu}_2)). \tag{212}$$

which is sometimes called a *profile likelihood*, because we have *profiled* over one of the parameters (more on this later).
The 68.3% and 95.4% intervals are found as the region for which $2\Delta q(\mu_1) < 1$ and $2\Delta q(\mu_1) < 4$, respectively. We can find these intersections using some code similar to that below,

```
def findIntervals(x,y,conts=[1,4]):
  xx0,yy0 = x[0],y[0]
  crossing_x = []
  for xx,yy in zip(x[1:],y[1:]):
    for K in conts:
      if (yy < K and yy0 > K) or (yy > K and yy0 < K):
        crossing_x.append(return_crossing(xx0,yy0,xx,yy,K))
    xx0=xx
    yy0=yy
  return crossing_x
```

The function $\Delta q(\mu_1)$ and intervals are shown in Figure 31. If we didn't profile μ_2, the intervals would be smaller, which shows the effect of correlations between parameters of interest.

Extracting the intervals/regions this way (using the profiled log-likelihood) is often referred to as the `MINOS` method (as this is the method used in the `MINOS` code to determine uncertainties).

Remember that in our definition of ζ_μ for the simple counting experiment, we had two clauses depending on whether or not the signal strength (μ) was greater than 0 or not. The introduction of this boundary poses no problem when constructing the frequentist intervals, however, it will have an effect when appealing to Wilks' theorem, as the assumption that ζ_μ is distributed as a $\chi^2(2)$ will breakdown when $\mu \sim 0$. Figure 32 shows an example of this happening in a CMS analysis of Higgs boson decays to $\gamma\gamma$ with the Run-1 dataset. The parameters $\mu_{\text{ggH+ttH}}$ and $\mu_{\text{VH+qqH}}$ are signal strengths for Higgs boson production modes involving fermion and vector-boson couplings, respectively. In this model, these parameters are bounded to be > 0. The contours indicate the 68% and 95% confidence regions determined from the observed data using a Neyman construction with likelihood ratio ordering (labelled "Feldman-Cousins") and appealing to Wilks' theorem (labelled "Likelihood Scan"). Away from the boundaries, the contours agree rather well – meaning that the conditions for Wilks' theorem to apply hold. However, close to the boundaries, the contours start to disagree. This is not due to there being not enough events in the sample, but really a consequence of the boundary. It is therefore important to check the coverage of the intervals/regions reported when using the `MINOS` method, when there are boundaries in the problem.

Note that we can even go one step further in approximation. Let's go back to equation 204.

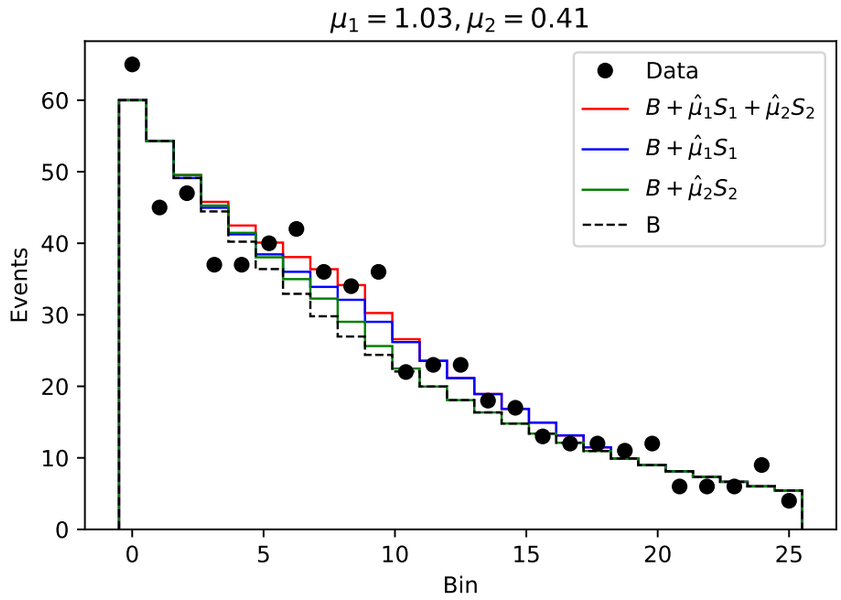

Figure 30: Background (B) and signal (S_1 and S_2) distributions at the global minimum of the likelihood parameters, stacked on top of one another. The data is shown too.

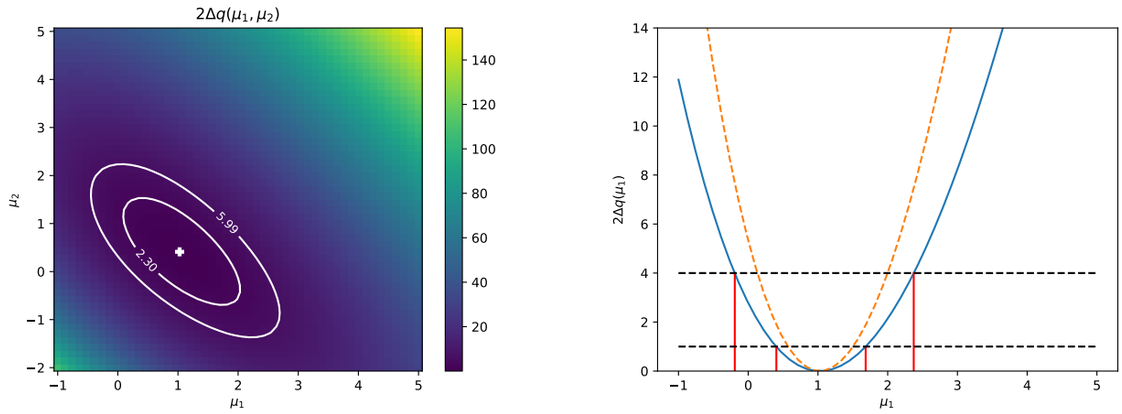

Figure 31: Left: $2\Delta q(\mu_1, \mu_2)$ in color scale and contours corresponding to the boundaries of the 68.3% and 95.4% confidence regions. Right: $2\Delta q(\mu_1)$ both profiling (blue line) and fixing μ_2 (orange dashed line) the 68.3% and 95.4% confidence intervals are indicated by the red lines.

We can see that in 1-dimension, this is just a parabolic function. Re-arranging gives us the usual approximation,

$$q_N(\theta) - q(\hat{\theta}) = \frac{1}{2} q_N''(\hat{\theta})(\theta - \hat{\theta})^2 \tag{213}$$

i.e the difference in the negative log-likelihood to the value at the minimum looks like a parabola, close to the minimum, that is centered at $\hat{\theta}$. If we had a random variable that was distributed as a Gaussian with mean θ and standard deviation σ, and we observed the value $\hat{\theta}$ as our MLE, then we would find that that our function,

$$q_N(\theta) - q(\hat{\theta}) = -\left(\ln(\phi(\theta, \sigma)) - \ln(\phi(\hat{\theta}, \sigma))\right) = -\left(-\frac{1}{2}\left(\frac{\theta - \hat{\theta}}{\sigma}\right)^2 + \frac{1}{2}\left(\frac{\hat{\theta} - \hat{\theta}}{\sigma}\right)^2\right) = \frac{1}{2}\left(\frac{\theta - \hat{\theta}}{\sigma}\right)^2, \tag{214}$$

This means if we match with equation 213, we can clearly see that,

$$\frac{1}{\sigma^2} = q_N''(\hat{\theta}). \tag{215}$$

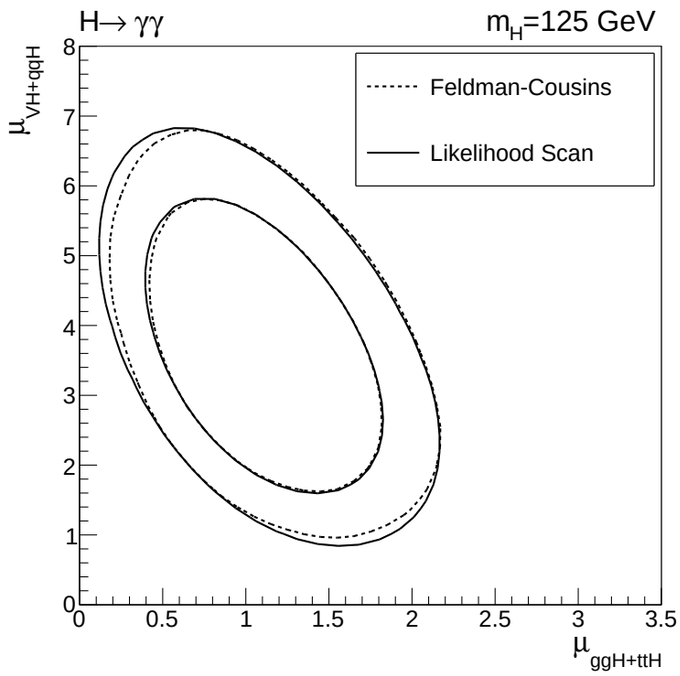

Figure 32: CMS analysis of Higgs boson decays to $\gamma\gamma$ with the Run-1 dataset. The two parameters are bounded to be > 0. The contours indicate the 68% and 95% confidence regions determined from the observed data using a Neyman construction with likelihood ratio ordering (labelled "Feldman-Cousins") and appealing to Wilks' theorem (labelled "Likelihood Scan").

So in this approximation, since for a Gaussian distribution, the 68.3% interval is given by the standard deviation, our 68.3% is given by the 2nd derivative of twice the difference in the negative log-likelihood function, evaluated at the minimum. This result extends to more than one dimension simply replacing q''_N with the Hessian yields an approximation of the co-variance matrix in multi-dimensional fits. In general we can approximate

$$\nu^{ij}_{1,1} = \left(\left[\frac{\partial^2 q}{\partial\theta_k \partial\theta_l} \right] \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right)^{-1}_{ij} \tag{216}$$

This means if we can calculate the matrix of second derivatives $\left[\frac{\partial^2 q}{\partial\theta_k\partial\theta_l}\right]$ evaluated at the maximum likelihood estimators $\theta = \hat{\theta}$, which is known as the *Hessian* matrix, then we can derive a covariance matrix using Eqn 216.

Taking the example from the two signals on top of a background, we find that,

$$(\nu^{\mu_i,\mu_j}_{1,1})^{-1} = \begin{bmatrix} \frac{\partial^2 q}{\partial\mu_1^2} & \frac{\partial^2 q}{\partial\mu_1\partial\mu_2} \\ \frac{\partial^2 q}{\partial\mu_2\partial\mu_1} & \frac{\partial^2 q}{\partial\mu_2^2} \end{bmatrix}_{(\hat{\mu}_1,\hat{\mu}_2)} \tag{217}$$

And in our case, we can find (I leave it for you to show it!)

$$\frac{\partial^2 q}{\partial\mu_1^2} = \sum_i n_i \frac{1}{\lambda_i^2} \left(\frac{\partial\lambda_i}{\partial\mu_1} \right)^2 \tag{218}$$

and

$$\frac{\partial^2 q}{\partial\mu_2^2} = \sum_i n_i \frac{1}{\lambda_i^2} \left(\frac{\partial\lambda_i}{\partial\mu_2} \right)^2 \tag{219}$$

and

$$\frac{\partial^2 q}{\partial\mu_1\partial\mu_2} = \frac{\partial^2 q}{\partial\mu_2\partial\mu_1} = \sum_i n_i \frac{1}{\lambda_i^2} \frac{\partial\lambda_i}{\partial\mu_1} \frac{\partial\lambda_i}{\partial\mu_2} \tag{220}$$

You will often find that minimisation packages have methods to calculate this for you. In scipy, the fit result holds onto the *inverse of the Hessian* so in our two signal example we can get the covariance from,

```
init_params = [0.1,0.1]
bounds = [(-10,10),(-10,10)]

fit_result = minimize(q_unconstrained,
      init_params,args=data,bounds=bounds)

print(2*fit_result.hess_inv.todense())
```

Where the factor of 2 comes from the fact that we need the inverse of *half* of the Hessian matrix according to Eqn 216. Have a look at **WilksTheorem2D.ipynb** to see a comparison of the analytic and numeric derived covariance matrices.

6.2.3 Gaussian models

Note, often its better to use the method of scanning the likelihood function previously described finding the intervals via the crossings at set levels to determine confidence intervals. However this approximation allows us to make a very helpful summary of measured quantities in that we can describe the measurements $\boldsymbol{\theta}$ as random variables that follow a multi-variate Gaussian,distribution with a mean vector of $\hat{\boldsymbol{\theta}} = (\theta_1, \theta_2, ..., \theta_N)$ and a covariance matrix $\nu_{1,1}$ or $\boldsymbol{\theta} \sim \Phi(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \nu_{1,1})$ where,

$$\Phi(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \nu_{1,1}) = \frac{1}{\sqrt{(2\pi)^N \det(\nu_{1,1})}} e^{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T (\nu_{1,1})^{-1} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})} \tag{221}$$

This says that given a set of random variables (i.e measurements) with known covariance and maximum likelihood estimators, through Wilks' theorem, we can approximate their joint probability distribution as a multivariate Gaussian. This is a remarkable result that a set of complicated measurements can be summarised with a set of maximum likelihood estimators and a covariance matrix! With this result physicists make use of a whole number of Gaussian based methods to model their measurements and present their data. These methods are beyond the scope of this lecture course, but you may come across them in your practical module.

6.3 Nuisance parameters (Non-examinable)

As we've seen Wilks' theorem gives us a powerful shortcut to determining confidence intervals. In particular, this is extremely useful when we want to introduce additional parameters in the model that we don't necessarily want to measure but we have to account for – these are known as *nuisance parameters*.

Previously we've talked about testing *simple* hypotheses. As with most things in real life, the hypotheses that we want to test are anything but simple. More often than not, we test *composite* hypotheses against one another. Composite hypotheses can be thought of as a set of hypotheses that are parameterised by one (or more) parameters, often related to parameters of a physical model we are interested in. In experimental data analysis, we also encounter parameters related to our experimental setup. These might be calibration coefficeints of the calorimeter, efficiencies on our event selection or the flux of a neutrino beam. Usually we are not so interested in defining the hypothesis with regards to all of these later parameters. We split the parameters of the model into two sets, $\theta = (\mu, \eta)$. The parameters μ are the *parameters of interest* and typically correspond to the parameters that represent the set of hypotheses we are interested in testing. Instead, the parameters η represent those parameters which we don't care about – these are the *nuisance parameters*. Typically, nuisance parameters are *measured*, and we can think of those measured values as being random variables in their own right.

Our composite hypotheses will be labelled as $H(\mu) = H(\mu, \eta(\mu))$ meaning in order to specify the

hypotheses corresponding to different values of μ, we must make a choice of what to do with η, which can depend on the value of μ. For dealing with experimental data, we will be using the likelihood function and the choice boils down to one of two procedures, *profiling* and *marginalisation*. Typically, you will encounter profiling in frequentist procedures, and marginalisation in Bayesian procedures.

Profiled likelihoods Profiling the likelihood can sometimes be thought of selecting the "best" value of the nuisance parameters for any value of the parameters of interest. The definition of "best" here means that which maximises the likelihood for a particular value(s) of μ. We'll see later that this choice will have useful implications, but for now, know that this is *the choice* for frequentists. We start by taking the negative of the log-likelihood function,

$$q(\mu, \eta) = -\ln L(\mu, \eta). \tag{222}$$

We then choose $\eta(\mu)$ as $\hat{\eta}_\mu$, which represents the values of η for which q minimized when evaluated at μ. Specifically,

$$\hat{\eta}_\mu = \text{argmin}_\eta \left\{ q(\mu, \eta) \Big|_\mu \right\}, \tag{223}$$

meaning we first fix μ and then find the values of η which minimize q. We then write the *profiled log-likelihood* as,

$$q(\mu) = q(\mu, \hat{\eta}_\mu), \tag{224}$$

to remove the η dependence from the likelihood. Note that we've not specified the likelihood here, and we could have separated it into two components; $L(\text{data}|\mu, \eta) \cdot L(\eta_{\text{measured}}|\eta)$. The term $L(\eta_{\text{measured}}|\eta)$ is the constraint term, which typically arises from some previous measurement of η, η_{measured}.

Going back to our simple counting experiment, suppose that we didn't know the value of the luminosity measurement perfectly but only within some measured value. We could modify the rate (λ) of our signal as,

$$\lambda(\mu) \to \lambda(\mu, \eta) = \mu A \epsilon l_0 (1 + k)^\eta + B \tag{225}$$

where k tells us the uncertainty on the nominal luminosity measurement l_0.

Have a look at Figure 33. On the left hand side, the colours show the value of $q(\mu, \eta)$ for our simple counting experiment, where we've used the values $n = 2$, $B = 0.5$, $l_0 = 1$, $A = 0.5$, $\epsilon = 0.9$, and $k = 0.3$.

The red line shows the value of η which minimize q at every valye of μ. These are the *profiled* values of η. The right hand side shows the resulting *profiled likelihood* curve in red from using these values of η. If instead we just assumed $\eta = 0$ (the orange line) we would get a more narrow curve. This tells us that the inclusion of a nuisance parameter has the effect of increasing the uncertainty (our interval gets larger) when profiling over the nuisance parameter.

The concept of profiling nuisance parameters is very important when we want to account for systematic uncertainties in our measurements. As we saw, we can introduce systematic uncertainties without having to construct intervals for them by profiling over them in the test statistic. When we have lots of these nuisance parameters, this saves a lot of computation power and allows us to only report uncertainties in parameters of interest while still accounting for the effect of systematic uncertainties.

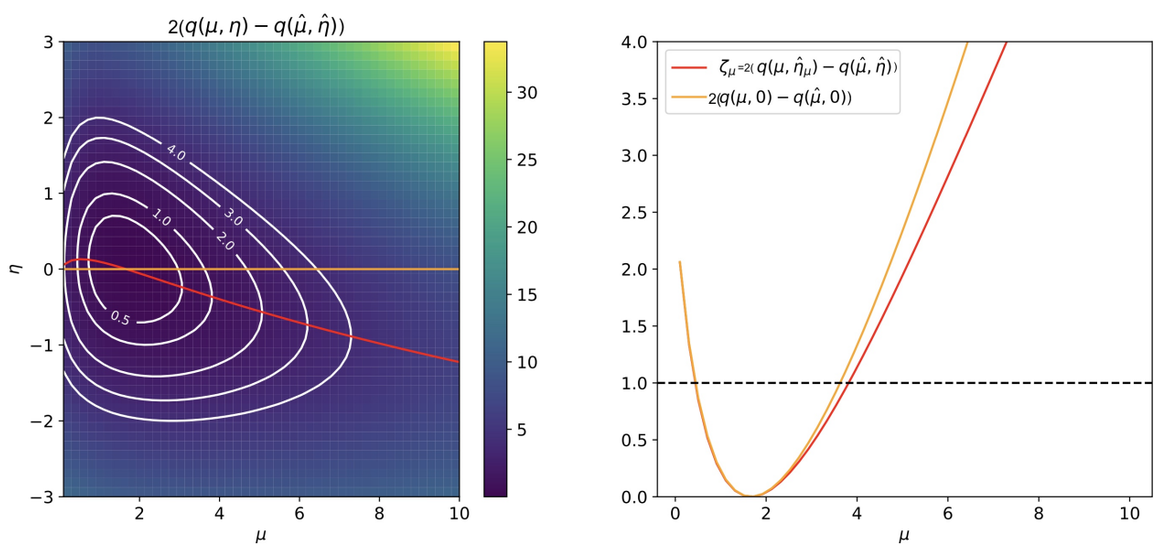

Figure 33: Left: $2q(\mu, \eta)$ for the simple counting experiment. The red line shows the profiled values of η for each value of μ. Right: profiled log-likelihood curves as a function of μ for the case when η is profiled (red) or fixed to zero (orange).

Summary

We're at the end of the first half of this course on statistics. There are many other statistical concepts which we didn't have time to cover, and you may come across in your projects (and of course the 2nd half of the lecture course)

- Information theory – We touched on this when discussing Wilks' theorem (its related to the second derivative of the log-likelihood), but this is a whole topic unto itself.
- Asymptotic theories – Wilks' theorem is extremely powerful for determining intervals based on likelihood ratio test statistics, however, there are many other test statistics out there with well understood asymptotic behaviour. These may become more common as our computing gets more efficient.
- Systematic uncertainties – We only covered how to deal with systematic uncertainties as nuisance parameters, but estimating and dealing with systematic uncertainties in hypothesis testing is a huge field. Furthermore, we didn't cover the case where a parameter in the model isn't specified under one or more hypotheses. This can lead to something known as the *Look elsewhere effect* and there are many interesting methods for dealing with this.
- Model selection – How do we choose how many free parameters to allow in a model of data and what form it takes? There's a few different methods for this – Akaike information criterion, the Fisher test and incorperating model choice as an uncertainty The discrete profiling method. You will find out more about Bayesian solutions to this in the second half of these lectures.

The following is a list of recommended further reading to find out about some of these subjects and other issues in statistics pertinent for experimantal data (at least ones that I found very useful when I was learning this stuff!)

- F. James, "*Statistical Methods in Experimental Physics*", ISBN: 978-9-812-70527-3 (2006).
- G. Cowan, "*Statistical Data Analysis*", ISBN: 978-0-198-50155-8 (1998).
- G. Cowan, "Statistics" (section 39) in "*Review of particle physics*", Chin. Phys. C 40, 100001 (2016).
- L. Lista, "*Statistical Methods for Data Analysis in Particle Physics*", ISBN 978-3-319-20176-4, (2015).
- A. Stuart, K. Ord, S. Arnold, "*Kendall's Advanced theory of Statistics*", Vol 2A: Classical inference and the linear model, ISBN: 978-0-470-68924-0 (2010).
- L. Lyons, N. Wardle, "*Statistical issues in searches for new phenomena in High Energy Physics*", Journal of Physics G: Nuclear and Particle Physics, Volume 45, Number 3.
- O. Behnke, K. Kroninger, G. Schott, T. Schorner-Sadenius, "*Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods*", ISBN: 978-3-527-41058-3 (2013).
- K. Cramner, "*Practical Statistics for the LHC*", Proceedings, 2011 European School of High-Energy Physics, (2011).