ASTRONOMICAL STATISTICS

John Peacock
Royal Observatory Edinburgh
jap@roe.ac.uk, http://www.roe.ac.uk/∼jap/teaching/astrostats/

May 31, 2021

"Probability theory is nothing but common sense reduced to calculation"$_{\text{Laplace}}$

Introductory Notes

This course introduces essential statistical concepts within the context of their use in Astronomical data analysis. It covers the basics of probability theory, through to advanced likelihood analysis for large-scale data challenges.

Schedule:

Monday 11:10 to 12:00 and Thursday 11:10 to 12:00

First Lecture 17th September 2012, Last Lecture 22nd October 2012.

Additional tutorial sessions will be held at times to be announced.

Lectures will take place in the ROE Lecture Theatre.

Handouts:

Printed notes define the examinable content of the course. They will be handed out in installments. It is strongly encouraged that personal notes are taken during each lecture where key issues, and illustrations will be discussed. It is expected that students read the notes. There will be three problem sheets, for each problem sheet an accompanying tutorial will go through these problems.

Suggested references:

The main text for this course is these notes. For further reading, the following may be useful, especially the first:

- Prasenjit Saha, *Principles of Data Analysis*, $\sim$£10 from Amazon or cappella-archive.com, or downloadable free from http://www.itp.uzh.ch/$\sim$psaha/. You should consider doing both, as it is an excellent short text, and scribble on the latter.

- D.S.Sivia & J. Skilling, *Data Analysis: A Bayesian Tutorial*, Oxford University Press. Another short, and extremely clear, exposition, this time of Bayesian ideas. Price $\sim$£23.

- Robert Lupton, *Statistics in Theory and Practice*, Princeton University Press; ISBN: 0691074291, Price: $\sim$£33. This is an excellent textbook, very much in the vein of this course. Good intro.

- William H. Press, Brian P. Flannery, Saul A. Teukolsky, William T. Vetterling, *Numerical Recipes in FORTRAN (or C or C++) Example Book: The Art of Scientific Computing*, Cambridge University Press; ISBN: 0521437210, Price: $\sim$£40. This covers much more than statistics, and provides computer code for all its methods. There are C and C^{++} versions as well as Fortran, but no Java yet. Apart from actual code, its strength is in clear explanations of scientific methods, in particular Statistical Methods. Main book ($\sim$£45) is available online – see CUP web site.

- Jasper Wall and Charles Jenkins. *Practical Statistics for Astronomers*, Cambridge University Press. About $\sim$£23. Very good practical guide, with a compact style that will suit those already familiar with the basics.

- David J. C. MacKay *Information Theory, Inference and Learning Algorithms*, Cambridge University Press; Sixth Printing 2007 edition (25 Sep 2003). A good statistics guide, focusing on implementation in algorithms. $\sim$£30

PART ONE

1 Probability and statistics

1.1 Introduction

The notion of probability is *central* to the scientific endeavour. It allows us to apply logical reasoning to data in order to deduce, with a particular – and *calculable* – degree of certainty, properties of the wider universe. The general question that one could pose, and that we will answer in this course, is: How do we reason in situations where it is not possible to argue with absolute certainty?

Lack of certainty arises in many cases because of random influences: noise on measurements; intrinsic quantum randomness (e.g. radioactive decay); effective randomness of complex systems (e.g.the National Lottery). To a large extent, probability theory is a practical tool for dealing with measurements that involve random numbers of this sort. But the subject is broader than this, and also covers the case where there is no randomness, but simply incomplete knowledge. Given imperfect measurements, and limited quantities of them, how confident can we be in our understanding of the universe?

Science, and in particular Modern Astronomy and Astrophysics, is impossible without knowledge of probability and statistics. To see this let us consider how **Scientific Knowledge** evolves[1]. First of all we start with a **problem**, e.g., a new observation, or something unexplained in an existing theory. In the case of a new observation, probability and statistics are already required to make sure we are have detected something at all. We then **conjecture** (or rather we simply make something up) a set of solutions, or theories, explaining the problem. Hopefully these theories can be used to deduce testable predictions (or they are not much use), so that they can be experimentally **tested**. The predictions can be of a statistical nature, e.g., the mean value of a property of a population of objects or events. We then need probability and statistics to help us decide which theories pass the test and which do not. Even in everyday situations we have to make such decisions, when we do not have complete certainty. If we are sure a theory has failed the test it can be disregarded, and if it passes it lives to fight another day. We cannot 'prove' a theory true, since it may fail a critical test tomorrow; nevertheless, we can approach a lawyer's 'proof beyond reasonable doubt' as a theory passes many tests. At this point, the theory becomes a standard part of science, and only exceptionally strong evidence would cause it to be abandoned.

In short, we make some (usually imprecise) measurements, which gives us some data, and from these data, we make some inferences (i.e. we try to learn something).

In all of this probability and statistics are vital for:

- **Hypothesis testing**. A hypothesis is a proposition. Are the data consistent with the hypothesis? Examples of hypotheses: the Universe's properties are isotropic, i.e. don't depend on the direction you look. In a sense the detection of astronomical objects falls into this category. The hypothesis is then is the signal consistent with noise?

[1] from the Theory of Knowledge, by Karl Popper in *Conjecture and Refutations*, 1963.

- **Parameter Estimation**. We interpret the data in the context of a theoretical *model*, which includes some free parameters which we would like to measure (or, more properly, *estimate* - we won't in general get a precise answer). Example, interpreting cosmological data (microwave background temperature fluctuations, distribution of galaxies) in terms of the currently standard 'ΛCDM' model, which has parameters such as the density (relative to the critical density) of baryons, cold dark matter and dark energy $\Omega_b, \Omega_c, \Omega_\Lambda$ respectively. We would like to know what these *parameters* of the *model* (ΛCDM). (Note that the model has a lot of other parameters too).
- **Model testing**. A *model* is a theoretical (or empirical) framework in which the observations are to be interpreted. One may wish to determine whether the data favours one model over another. This is model testing. Examples might be the ΛCDM model vs the Steady-State model. Each one has some free parameters in it, but we want to ask a broader question – which model (regardless of the value of the parameters) is favoured by current data? In other words, a model is a class of explanation of a general type, which can contain many specific hypotheses (e.g. with different values of parameters).

These are the 'bread-and-butter' aims of much of data analysis.

Example: we toss a coin 6 times and get the resulting dataset: $\{T, T, H, T, T, T\}$. We can

- Test the **hypothesis** that *'The coin is fair'*.
- Construct a **model** which says that the probability that the coin will come up tails is $p(T) = q$, and heads is $p(H) = 1-q$. q is a 'parameter' of the model, and we can estimate it (and estimate how uncertain the estimate is, as well).
- Consider two **model**s: one is the one above, the other is a simpler model, which is that the coin is fair and $q \equiv 1/2$ always. Note that in this case, one model is a subset of the other. Do the data give significant evidence in favour of a non-fair coin?

In all these cases, we want a quantified answer of some sort, and this answer is calculated using the tools of statistics. We will use these tools in two ways in this course

- **As a theoretical tool:** How do we model or predict the properties of populations of objects, or study complex systems? e.g. What numbers should we use to characterise the statistical properties of galaxies, or stars? For clustering, the two-point correlation function and the power spectrum of fluctuations are usually used.
- **Experimental design:** What measurements should we make in order to answer a question to the desired level of accuracy? e.g. Given a fixed amount of observing time, how deep or wide should a galaxy survey be, to measure (say) how strongly galaxies are clustered on scales less than 20 Mpc?

In astronomy the need for statistical methods is especially acute, as we cannot directly interact with the objects we observe (unlike other areas of physics). For example, we can't re-run the same supernova over and again, from different angles and with different initial conditions, to see what happens. Instead we have to assume that by observing a large number of such events, we can collect a 'fair sample', and draw conclusions about the physics of supernovae from this sample. Without this assumption, that we can indeed collect fair samples, it would be questionable if astronomy is really a science. This is also an issue for other subjects, such as archaeology or forensics. As a result of this strong dependence on probabilistic and statistical arguments, many astronomers have contributed to the development of probability and statistics.

As a result, there are strong links between the statistical methods used in astronomy and those used in many other areas. These include particularly any science that makes use of signals (e.g. speech processing) or images (e.g. medical physics) or of statistical data samples (e.g. psychology). And in the real world, of course, statistical analysis is used for a wide range of purposes from forensics, risk assessment, economics and, its first use, in gambling.

1.2 What is a probability ?

The question "*what is a probability?*" is actually a very interesting one. Surprisingly there is not a universally agreed definition.

In astronomy (and science) we want to apply the *deductive* (consequences of general principles) reasoning of everyday experience, and pure maths (Boolean logic), to problems that require *inductive* reasoning (going from effects/observations to possible causes). This is an example of what is known as the 'epistemic' philosophical stance on probability:

- *Epistemic* is the concept of probability used to describe the property or effect of a system when the causative circumstances for that property are unknown or uncertain e.g. "the age of the Universe is 13.7 ± 0.13 billion years", "the likely cause of Solar coronal heating is magnetic reconnection".
- *Aleatory* is the concept of probability used for example in games of chance. It deals with predicting the future outcome of random physical processes. It can further be subdivided into phenomenon which are in principle predictable and those which are inherently unpredictable e.g. "it will rain tomorrow with a 40% probability", "I have a 1 million chance of winning the lottery".

This tells us different ways in which probability can be used, but not what it is. To understand the origins of this problem let us take a quick look at the historical development of probability[2]:

1654: Blaise Pascal & Pierre Fermat: After being asked by an aristocratic professional gambler how the stakes should be divided between players in a game of chance if they quit before the game ends, Pascal (he of the triangle) and Fermat (of 'Last Theorem' fame) began a correspondence on how to make decisions in situations in which it is not possible to argue with certainty, in particular for gambling probabilities. Together, in a series of letters in 1654, they laid down the basic Calculus of Probabilities (See Section 1.3).

[2] Note that a large number of key figures in this history were Astronomers.

1713: James (Jacob) Bernoulli: Bernoulli was the first to wonder how does one assign a probability. He developed the 'Principle of Insufficient Reason': if there are N events, and no other information, one should assign each event the probability $P = 1/N$. But he then couldn't see how one should update a probability after an event has happened.

1763: Thomas Bayes: The Rev. Thomas Bayes solved Bernoulli's problem of how to update the probability of an event (or hypothesis) given new information. The formula that does this is now called **Bayes' Theorem** (see Section 2.3).

1795: Johann Friederich Carl Gauss: At the age of 18, the mathematical astronomer, Gauss, invented the method of 'Least Squares' (section 2.1), derived the Gaussian distribution of errors (section 1.6), and formulated the Central Limit Theorem (section 1.9).

1820: Pierre-Simon de Laplace: Another mathematical astronomer, Laplace re-discovered Bayes' Theorem and applied it to Celestial Mechanics and Medical Statistics. Marked a return to earlier ideas that 'probability' is a lack of information

1850's: The Frequentists: Mathematicians reject Laplace's developments and try to remove subjectivity from the definition of probability. A probability is a measured frequency. To deal with everyday problems they invent **Statistics**. Notable amongst these are Boole (1854), Venn (of Diagram fame) (1888), Fisher (1932) and von Mises (1957).

1920's: The Bayesians: We see a return to the more intuitive ideas of Bayes and Laplace with the **Neo-Bayesian**. A probability is related to the amount of information we have to hand. This began with John Maynard Keynes (1921), and continued with Jeffreys (1939), Cox (1946) and Steve Gull (1980).

1989: E. Jaynes: Using Bayesian ideas Jaynes tries to solve the problem of assigning probability with the **Principle of Maximum Entropy** (section 2.4).

So what is a probability ? There are three basic interpretations about what a probability is, based on how the probabilities are assigned:

- **Frequentist (or Classical) probabilities:** Probabilities are measureable *frequencies*, assigned to objects or events. The relative frequency (probability) of an event arises from the number of times this event would occur defined relative to an 'infinite ensemble' of 'identical' experiments. This is intuitively linked to games of chance (e.g. one can conceptualise rolling a dice 'forever') but breaks down in some obvious situations, e.g., for single events, or in situations were we cannot in practice measure the frequency, we have to invent a hypothetical ensemble of events. Mathematically it requires notions of infinity and randomness which are not well defined in general.

- **Bayesian probabilities:** Probability is a 'degree-of-belief' in a proposition, allocated by an observer given the available information (data); uncertainty arising from incomplete data or noise. This is radically different to the frequentist approach, but allows us to deal with situations where no ensemble can even be imagined: e.g. 'what is the probability that there is life on Mars?'.

- **Propensity probabilities:** Here probabilities are objective properties of a system. This only arises commonly in modern physics when considering quantum processes e.g. the wavefunction of an photon means that it 'has' (i.e. a fundamental property) a 50% chance of going through either slit in the two slit experiment.

These do not obviously give the same answer to problems in probability and statistics, as they assign different probabilities to some events. However, in many instances they are the same. In this course we shall consider both the frequentist and Bayesian interpretations, which start from the same basic mathematical principles. This should provide both a 'toolbox' of practical methods to solve everyday problems in physics and astronomy, as well as an understanding of the subject as a whole.

1.3 What is a statistic?

A *statistic* is a number which is calculated from a sample of data.

The hope is that the number will tell you something about the population as a whole, or the underlying distribution from which the sample is drawn. An obvious example is the average of N data – for large N it should be a decent approximation to the mean of the underlying distribution from which the data are drawn.

1.4 The calculus of probability

The **Calculus of Probabilities** is the mathematical theory which allows us to predict the statistical behaviour of complex systems from a basic set of fundamental axioms. The axioms of probability theory stem directly, and are simply an expression of, Boolean logic. They were first formulated by Cox (1946) (and later by Kolmogorov), we summarise them here and will then explain each one in details:

1. A probability is a number between 1 and 0 : $0 \leq \ P(X) \ \leq 1$
2. The probability of X being true *or* false is 1 : $P(X) + P(\bar{X}) = 1$
3. The probability of X being true *and* Y being true is the product of the probability of X being true given that Y is true multiplied by the probability that Y is true : $P(X, Y) = P(X \mid Y)P(Y)$.

From these 3 simple rules all of statistics can ultimately be derived (!). Note that is $P(X) = 0$ then X is said to never have occurred, and if $P(X) = 1$ then X is said to be certain. The second axiom makes the obvious statement that the probability of obtaining *some* (any) result must be certain. In general probabilities come in two distinct forms: discrete, where P_i is the probability of the i^{th} event occurring, and continuous, where $P(x)$ is the probability that the event, or random variable, x, occurs.

We now revisit each of these 3 axioms in more detail

1. **The Range of Probabilities:** The probability of an event is measurable on a continuous scale, such that the probability of a discrete event i, p_i is a *real number* in the range $0 \leq p_i \leq 1$ and the probability of a variable x, $p(x)$ is a *real number* in the range $0 \leq p(x) \leq 1$

2. **The Sum Rule:** The sum of all discrete possibilities is

$$\sum_i p_i = 1. \tag{1}$$

For a continuous range of random variables, x, this becomes

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

where it is clear now that $p(x)$ is the probability of a variable lying between x and $x + dx$ is $p(x)dx$. We call $p(x)$ a **probability density function (pdf)**. The probability density function clearly must have units of $1/x$. Note that a random variable need not have a uniform pdf; random means it's not repeatable, not that all outcomes are equally likely.

3. **The Multiplication of Probabilities:** The probability of two events x and y both occurring is the product of the probability of x occurring and the probability of y occurring given that x has occurred. In notational form:

$$\begin{aligned} p(x,y) &= p(x|y)p(y) \\ &= p(y|x)p(x) \end{aligned} \tag{3}$$

where we have introduced the **conditional probability**, $p(x|y)$, which denotes the probability the of x *given* the event y has occurred, and should be read *probability of x given y.*

These basic axioms can be used to simply derive a number of foundational concepts in probability theory the most important of which we summarise here

- **The Addition of Exclusive Probabilities:** If the probabilities of n mutually exclusive events, x_1, $x_2 \cdots x_n$ are $p(x_1)$, $p(x_2) \cdots p(x_n)$, then the probability that **either** x_1 **or** x_2 **or** $\cdots x_n$ occurs is[3]

$$p(x_1 + x_2 + \cdots + x_n) = p(x_1) + p(x_2) + \cdots + p(x_n). \tag{4}$$

Note that this is not formally an axiom, since it can be derived from the sum rule.

- **The theorem of total probability:** A related result using conditional probabilities is that, if an event E can follow one of a number of independent preceding events A, B etc., then

$$p(E) = p(E|A)p(A) + p(E|B)p(B) + \cdots \tag{5}$$

(this just says we can get E either by E and A, or E and B etc.).

[3] **Maths Notes:** The logical proposition 'A.OR.B' can be written as $A + B$, or in set theory, $A \bigcup B$,which is the union of the set of events A and B. Similarly 'A.AND.B' can be written as AB or $A \bigcap B$, which is the intersection of events.

- **Bayes' Theorem:** Since the joint probability of x and y is the same whether we write it as $p(x,y)$ or $p(y,x)$, the multiplication of probabilities gives

  $$p(x,y) = p(x|y)p(y) = p(y|x)p(x) \tag{6}$$

  Hence

  $$p(y|x) = \frac{p(x|y)p(y)}{p(x)}. \tag{7}$$

  This is **Bayes' Theorem**. Note that this is not an axiom since it can be derived from the product rule.

1.5 Some examples

This probably all seems rather abstract, and it is helpful to see how it applies in some practical cases. Initially, it is easiest to consider frequentist examples where it is easy to imagine repeated trials. There is something of a tradition here in appealing to the case of an 'urn', which is a clay pot containing some numbers of various balls.

Example 1: conditional probabilities Suppose an urn holds 3 black balls and one white. We pick two balls without replacing the first one: what is the probability of getting B then W, or W then B? In the definition of conditional probability, let event x be 'B first' and y be 'W second', and we know that $p(x,y) = p(x)p(y|x)$. Now $p(x) = 3/4$ (since we pick one ball at random); having picked B, there is one chance in 3 of getting W from the remaining balls. So $p(y|x) = 1/3$ and overall $p(x,y) = (3/4)(1/3) = 1/4$. In the other order, the chance of W first is 1/4, but then the chance of getting B is 1, so the probability of B and W is 1/4 for either order. Since these are independent events, the probability of B and W in either order is 1/2. Note that this equality of probabilities is not an illustration of Bayes' theorem, which is hard to apply in this case. It is easy to deal with the probability of B first and the conditional probability of W second, having chosen B. But Bayes says that the probability of this outcome could also involve the probability of choosing B second and the conditional probability of choosing W first having chosen W first – and it isn't obvious how to work out either of these.

Example 2: the national lottery The case of drawing balls and not caring about the order is practically important in the national lottery. Suppose our urn has M numbered balls, of which we select N. All balls are numbered, and we win if all N balls are ones we chose in advance. Think of this step by step: ball 1 is OK with probability N/M; ball 2 is OK with probability $(N-1)/(M-1)$ etc. Thus

$$p(\mathrm{winning}) = \frac{N!}{M!/(M-N)!} = 1/C_N^M. \tag{8}$$

The last form of the result shows a different and quicker way of reasoning: all possible ways of choosing N from M are equally likely, but only one choice wins. Given the size of the UK, a sensible win probability is of order 10^{-7}. In practice, the choice was $N = 6$, $M = 49$, giving $p = 1/13983816$.

Example 3: Bayes with urns Now suppose we have two urns: one with 3B and 1W, and one with 10B. We pick and urn at random and draw a ball. If it is W, clearly we chose urn 1; but if it is B, what is the probability that it came from urn 1? Bayes theorem works easily here. Let x mean 'we chose urn 1' and y mean 'we got a black ball'. What we want is $p(x|y)$, which is $p(y|x)p(x)/p(y)$.

Two of these terms are easy: $p(y|x) = 3/4$ and $p(x) = 1/2$. $p(y)$ seems problematic at first sight, but from the theorem of total probability $p(y) = (3/4)(1/2) + (1)(1/2) = 7/8$. Thus the answer to our question is

$$p(x|y) = (3/4)(1/2)/(7/8) = 3/7. \tag{9}$$

1.6 Multivariate probability distributions

We have introduced the idea of a pdf for a continuous variable x, but often we have more than one variable (two will do for illustration). It is then natural to define a joint pdf $p(x, y)$ such that $p(x, y)\, dx\, dy$ is the probability of finding x in the range dx and y in the range dy. This is called a **multivariate distribution**.

- **Independence.** The variables x and y are **independent** if their joint distribution function can be written,
$$p(x, y) = p(x)\, p(y), \tag{10}$$
for all values of x and y. (Note – this is not the same as the variables being *uncorrelated* – see later.)

- **Conditional distributions.** The **conditional** distribution of x at *fixed* $y = c$ say, is defined by $p(x|y = c)$. Often it is not useful, because we may not actually know y exactly. But if we want this distribution, it is different from $p(x, y = c)$ since the conditional distribution needs to be normalized:
$$p(x|y = c) = p(x, y = c)/ \int p(x, y = c)\ dx. \tag{11}$$

- **Marginalisation.** If we have the joint pdf $p(x, y)$ and want the pdf of x alone, then we integrate over all possibilities for y. This process is called **marginalisation**:
$$\begin{aligned} p(x) &= \int dy\, p(x, y) \\ p(y) &= \int dx\, p(x, y). \end{aligned} \tag{12}$$
are the **marginal distributions** of x and y respectively. We have also $\int\int dy\, dx\, p(x, y) = 1$.

1.6.1 A classical picture of probabilities as volumes

It is instructive to think about a very simple physical example of a system that behaves according the axioms described above. Such a system is volumes of material (this is generalised in mathematics to ‘generalised volumes’ or ‘measure theory’). For example consider a 2D flat space with Cartesian coordinates x^1, x^2 then the volume over a region X is given by

$$V = \mathrm{vol}(A) \equiv \iint_X \mathrm{d}x^1 \mathrm{d}x^2 \tag{13}$$

where the integral is over the region X. If two such regions exists then the total volume is additive such that

$$\mathrm{vol}(A \cup B) = \mathrm{vol}(A) + \mathrm{vol}(B) \tag{14}$$

Now suppose A is subdivided into disjoint pieces $(A_1, \ldots, A_N)$ whose union is A and whose volumes are $V_1, \ldots, V_N$. Now the additive property of the volume integral means that $\sum_i V_i = V$ or $\sum_i V_i/V = 1$ (we have recovered axiom 2). Also because A_i are subsets of A we have $0 \leq V_i \leq V$ or $0 \leq V_i/V \leq 1$ (we have recovered axiom 1).

The important property is the additive nature of the integral which allows us to associate a probability with the normalised volume V_i/V. To complete this we note that the volume integral above can be generalised to include a density function such that

$$\mathrm{vol}(A_i) \equiv \iint_X \rho(x^1, x^2)\mathrm{d}x^1\mathrm{d}x^2, \tag{15}$$

by comparison with the continuous case above we see that the name **probability density function** has its roots in this simple analogy.

1.7 Basic properties of probability distributions

If we consider a general probability function $p(x)$ which a function of x, what general features of this function are useful?

1.7.1 Single distributions

Probability distributions can be characterized by their **moments**.

Definition:

$$m_n \equiv \langle x^n \rangle = \int_{-\infty}^{\infty} dx\, x^n p(x), \tag{16}$$

is the n^{th} moment of a distribution. The angled brackets $\langle \cdots \rangle$ denote the **expectation value**. Probability distributions are normalized so that

$$m_0 = \int_{-\infty}^{\infty} dx\, p(x) = 1 \tag{17}$$

The zeroth moment recovers axiom 2, the sum rule.

The first moment,

$$m_1 = \langle x \rangle\,, \tag{18}$$

gives the **expectation value** of x, called the **mean**: the average or typical expected value of the random variable x if we make random drawings from the probability distribution.

Centred moments are obtained by shifting the origin of x to the mean;

$$\mu_n \equiv \langle (x - \langle x \rangle)^n \rangle\,. \tag{19}$$

The second centred moment,

$$\mu_2 = \left\langle (x - \langle x \rangle)^2 \right\rangle, \tag{20}$$

is a measure of the **spread** of the distribution about the mean. This is such an important quantity it is often called the **variance**, and denoted

$$\sigma^2 \equiv \mu_2. \tag{21}$$

We will need the following, useful result later:

$$\sigma^2 = \left\langle (x - \langle x \rangle)^2 \right\rangle = \left\langle (x^2 - 2x \langle x \rangle + \langle x \rangle^2) \right\rangle = \left\langle x^2 \right\rangle - \langle x \rangle^2 . \tag{22}$$

The variance is obtained from the mean of the square minus the square of the mean. Another commonly defined quantity is the square-root of the variance, called the **standard deviation**; σ. This quantity is sometimes also called the **root mean squared (rms) deviation**[4], or **error**[5]. A Gaussian can be entirely characterised by its first and second moments.

Higher moments ($\mu_3, ..., \mu_N$) characterise the distribution further, with odd moments characterising asymmetry. In particular the third moment, called the **skewness**, $\langle x^3 \rangle$, characterises the simplest asymmetry, while the fourth moment, the **kurtosis**, $\langle x^4 \rangle$, characterises the flatness of the distribution. Rarely will one go beyond these moments.

1.7.2 Multiple distributions

For **bivariate** distributions (of two random variables), $p(x, y)$, one can also define a **covariance** (assume $\langle x \rangle = \langle y \rangle = 0$);

$$\text{Cov}(x, y) = \langle xy \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dx\, dy\, xy\, p(x, y) \tag{23}$$

and a dimensionless **correlation coefficient**;

$$r = \frac{\langle xy \rangle}{\sqrt{\langle x^2 \rangle \langle y^2 \rangle}}, \qquad \text{(Zero means assumed)} \tag{24}$$

which quantifies the similarity between two variables. r must lies between +1 (completely correlated) and −1 (completely anti-correlated). $r = 0$ indicates the variables are *uncorrelated* (but not necessarily *independent*). We shall meet the correlation coefficient later when we try and estimate if two data sets are related.

1.7.3 Variance of $x \pm y$

If x and y are **independent**, then, if $z \equiv x \pm y$ (both give the same answer), then the variance of z comes from

$$\langle z \rangle = \langle x \rangle \pm \langle y \rangle \tag{25}$$

[4] One should be cautious, as the term rms need not apply to deviations from the mean. It sometimes applies to $\sqrt{m_2}$. One should be explicit that it is a rms **deviation**, unless the mean is known to be zero.

[5] Again one should be cautions about referring to σ as the error. 'Error', or 'uncertainty' implies the distribution has a Gaussian form (see later), which in general is untrue.

and

$$\langle z^2 \rangle = \langle x^2 \rangle \pm 2\langle xy \rangle + \langle y^2 \rangle. \tag{26}$$

If x and y are **independent**, then $\langle xy \rangle = \langle x \rangle \langle y \rangle$. In this case the variance of z is

$$\begin{aligned} \sigma_z^2 &= \langle z^2 \rangle - \langle z \rangle^2 \\ &= \langle x^2 \rangle \pm 2\langle x \rangle\langle y \rangle + \langle y^2 \rangle - \Big(\langle x^2 \rangle \pm 2\langle x \rangle\langle y \rangle + \langle y^2 \rangle\Big) \\ &= \sigma_x^2 + \sigma_y^2. \end{aligned} \tag{27}$$

Note that this is independent of the form of the probability distribution – it doesn't matter, as long as the measurements are independent.

1.7.4 Transformation of random variables

The probability that a random variable x has values in the range $x - dx/2$ to $x + dx/2$ is just $p(x)dx$. Remember $p(x)$ is the **probability density**. We wish to transform the probability distribution $p(x)$ to the probability distribution $g(y)$, where y is a function of x. For a continuous function we can write,

$$p(x)dx = g(y)dy. \tag{28}$$

This just states that probability is a conserved quantity, neither created nor destroyed. Hence

$$p(x) = g(y(x)) \left| \frac{dy}{dx} \right|. \tag{29}$$

So the transformation of probabilities is just the same as the transformation of normal functions in calculus, requiring the use of the Jacobian matrix of the transformation. This was not necessarily obvious, since a probability is a special function of random variables. This transformation is of great importance in astronomical statistics. For example it allows us to transform between distributions and variables that are useful for theory, and those that are observed.

2 Basic Probability distributions and Examples

We will will discuss some common probability distributions that occur in physics and more specifically in Astronomy. The distributions we discuss here are common ones for calculating likelihoods $p(D|\theta_i)$.

2.1 Discrete probability distributions

We can use the calculus of probabilities to produce a mathematical expression for the probability of a wide range of multiple discrete events.

2.1.1 Binomial distribution

The Binomial distribution allows us to calculate the probability, p_n, of n successes arising after N independent trials.

Suppose we have a sample of objects of which a probability, p, of having some attribute (such as a coin being heads-up) and a probability, $q = 1 - p$ of not having this attribute (e.g. tails-up). Suppose we sample these objects twice, e.g. toss a coin 2 times, or toss 2 coins at once. The possible outcomes are hh, ht, th, and tt. As these are independent events we see that the probability of each distinguishable outcome is

$$\begin{aligned} P(hh) &= P(h)P(h), \\ P(ht+th) &= P(ht)+P(th) = 2P(h)P(t), \\ P(tt) &= P(t)P(t). \end{aligned} \tag{30}$$

The key realisation is that these combinations are simply the coefficients of the binomial expansion of the quantity $(P(h) + P(t))^2$.

In general, if we draw N objects instead of 2, then the number of possible permutations which can result in n of them having some attribute (e.g. 'heads') is the n^{th} coefficient in the expansion of $(p + q)^N$, the probability of each of these permutations is $p^n q^{N-n}$, and the probability of n objects having some attribute is the binomial expansion

$$P_n = C_n^N p^n q^{N-n}, \quad (0 \leq n \leq N), \tag{31}$$

where

$$C_n^N = \frac{N!}{n!(N-n)!} \tag{32}$$

are the **binomial coefficients**. The binomial coefficients can here be viewed as statistical weights which allow for the number of possible indistinguishable permutations which lead to the same outcome. This distribution is called the general **binomial**, or **Bernoulli distribution**. We can plot out the values of P_n for all the possible n (Figure 1) and in doing so have generated the predicted probability distribution which in this case is the binomial distribution whose form is determined by

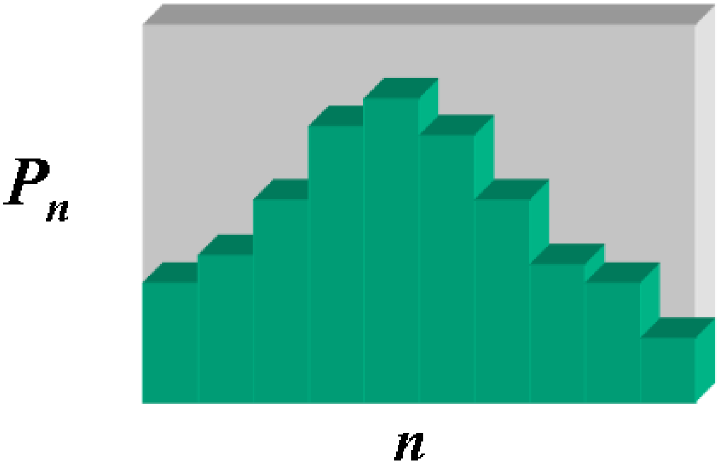

Figure 1: *Histogram of a binomial distribution.* P_n *is the probability that* n *objects in the sample have an attribute.*

N, n, p_1. If we have only two possible outcomes $p_2 = 1 - p_1$. The Binomial distribution can be generalised to a multinomial distribution function.

The mean of the binomial distribution is easy to find. Call the trial outcomes x_i and let x be 1 for a success and 0 for a failure. The number of successes is $n = \sum x_i$, so the desired mean is

$$\langle n \rangle = \sum \langle x_i \rangle = N p_1, \tag{33}$$

since $\langle x_i \rangle = p_1$ for each of N trials. A much more tedious way of obtaining the same result is to sum over the binomial distribution explicitly:

$$\begin{aligned}
\langle n \rangle &= \sum_{n=0}^{N} n P_n \\
&= \sum_{n=0}^{N} n \frac{N!}{n!(N-n)!} p_1^n p_2^{N-n} \\
\text{(Note lower limit)} &= \sum_{n=1}^{N} \frac{N!}{(n-1)!(N-n)!} p_1^n p_2^{N-n} \\
\text{(Let } m = n-1 \text{ and } M = N-1) &= N p_1 \left(\sum_{m=0}^{M} \frac{M!}{m!(M-m)!} p_1^m p_2^{M-m} \right) \\
\text{(Sum is unity)} &= N p_1
\end{aligned} \tag{34}$$

For $p_1 \neq p_2$ the distribution is asymmetric, with mean $\langle n \rangle = N p_1$, but if N is large the shape of the envelope around the maximum looks more and more symmetrical and tends towards a Gaussian distribution – an example of the **Central Limit Theorem** at work. More of this in a later lecture.

2.1.2 Poisson distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

The **Poisson distribution** occupies a special place in probability and statistics, and hence in observational astronomy. It is the archetypical distribution for point processes. It is of particular importance in the **detection** of astronomical objects since it describes **photon noise**. It essentially models the distribution of randomly distributed, independent, point-like events, and is commonly taken as the null hypothesis.

It can be derived as a limiting form of the binomial distribution. After N trials, the probability of n 'successes' of an event of probability p is, from the Binomial distribution:

$$P_n = C_n^N p^n (1-p)^{N-n}, \tag{35}$$

where instead of 2 possibilities p and q we only have success p or failure $1-p$. Let us suppose that the probability p is very small, but that in our experiment we allow N to become large, *but keeping the mean finite*, so that we have a reasonable chance of finding a finite number of successes n. That is we define $\lambda = \langle n \rangle = Np$ and let $p \to 0$ and $N \to \infty$, while keeping the mean λ =constant. Then,

$$P_n = \frac{N!}{n!(N-n)!} \left(\frac{\lambda}{N}\right)^n \left(1 - \frac{\lambda}{N}\right)^{N-n}. \tag{36}$$

We assume $N \gg n$. We now want to take the limit of this equation as $N \to \infty$.

Clearly $N!/(N-n)!$ tends to N^n as $N \to \infty$ with n finite. The more difficult final term becomes[6]

$$\left(1 - \frac{\lambda}{N}\right)^{N-n} = e^{-\lambda}, \tag{38}$$

as $N \to \infty$. So overall we find

$$P_n = \frac{\lambda^n e^{-\lambda}}{n!} \tag{39}$$

This is **Poisson's distribution** for random point processes, discovered by Siméon-Denis Poisson in 1837, and shown in Fig. 2.

2.1.3 Moments of the Poisson distribution:

Let's look at the moments of the Poisson distribution:

$$m_i = \left\langle n^i \right\rangle = \sum_{n=0}^{\infty} n^i P_n \tag{40}$$

[6] **Maths Note** The limit of terms like $(1 - x/N)^N$ when $N \to \infty$ can be found by taking the natural log and expanding to first order:

$$\ln(1 - x/N)^N = N \ln(1 - x/N) \to N(-x/N + \mathcal{O}((x/N)^2)) \to -x \tag{37}$$

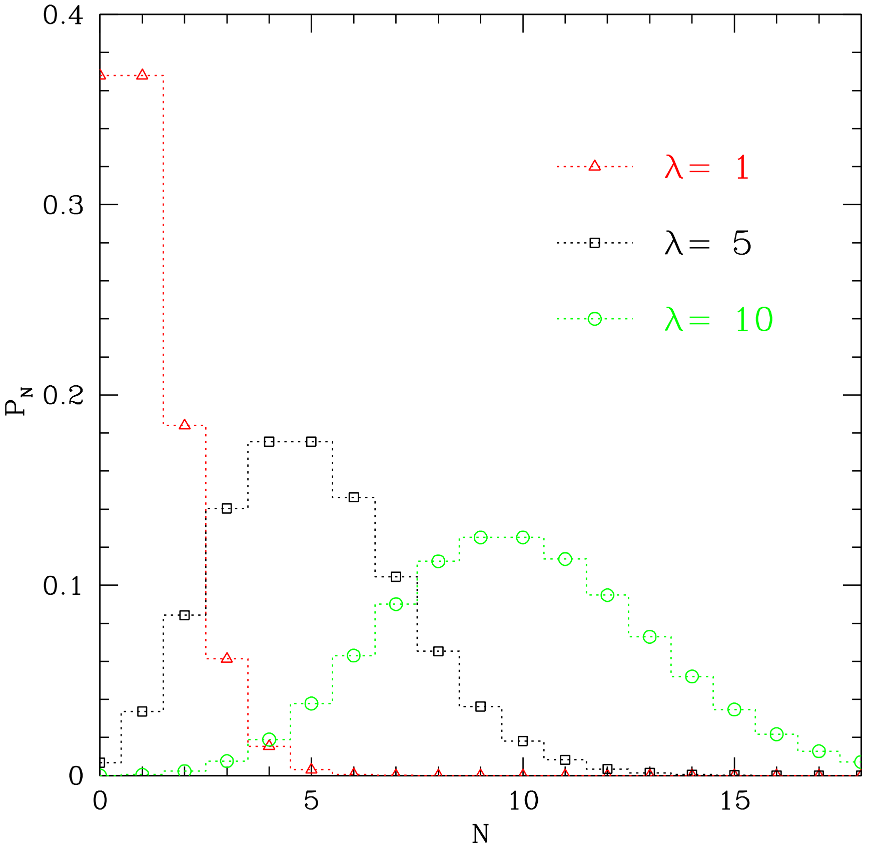

Figure 2: *Histogram of different Poisson distributions:* $\lambda = 1$ *(connected triangles),* $\lambda = 4$ *(connected squares),* $\lambda = 10$ *(connected circles).*

The mean of Poisson's distribution is Λ by definition. As with the binomial theorem, this can be checked explicitly:

$$\begin{aligned} \langle n \rangle &= \sum_{n=0}^{\infty} n \frac{\lambda^n e^{-\lambda}}{n!} \\ (0! = 1 \textit{ so } 0/0! = 0 \textit{ and } n = 0 \textit{ term gives } 0) &= \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{(n-1)!} \\ (m = n-1) &= \lambda \sum_{m=0}^{\infty} \frac{\lambda^m e^{-\lambda}}{m!} \\ (\textit{sum} = 1 \textit{ again}) &= \lambda. \end{aligned} \tag{41}$$

Now let's look at the second centred moment (i.e. the variance). As with the binomial theorem, this is most easily attacked by considering the sum of the individual trials before taking the limit on $N \to \infty$. $n = \sum x_i$ so

$$n^2 = \sum x_i^2 + \sum_{i \neq j} x_i x_j \Rightarrow \langle n^2 \rangle = N \langle x_i^2 \rangle + N(N-1) \langle x_i \rangle^2. \tag{42}$$

In the large-N limit, $N(N-1)$ is indistinguishable from N^2, so $\langle n^2 \rangle = \lambda + \lambda^2$, where we have used $x_i^2 = x_i$, since the outcome of a given trial is either 1 or 0 successes. Once again, this can be derived from the Poisson distribution directly:

$$\begin{aligned}
\mu_2 &= \left\langle (n-\langle n\rangle)^2 \right\rangle = \langle n^2\rangle - \langle n\rangle^2 \\
&= \sum_{n=0}^{\infty} n^2 \frac{\lambda^n e^{-\lambda}}{n!} - \lambda^2 \\
&= \sum_{n=1}^{\infty} \frac{n\lambda^n e^{-\lambda}}{(n-1)!} - \lambda^2 \\
&= \sum_{n=1}^{\infty} \frac{n\lambda\lambda^{n-1} e^{-\lambda}}{(n-1)!} - \lambda^2 \\
(m=n-1) &= \sum_{m=0}^{\infty} \frac{(1+m)\lambda\lambda^{m} e^{-\lambda}}{m!} - \lambda^2 \\
(\text{Take out } \lambda \text{ in first term, [the } '1']\text{. Then sum} = 1) &= \lambda + \sum_{m=0}^{\infty} \frac{m\lambda\lambda^{m} e^{-\lambda}}{m!} - \lambda^2 \\
(\text{As before, and take q} = \text{m} - 1) &= \lambda + \sum_{q=0}^{\infty} \lambda^2 \frac{\lambda^{q} e^{-\lambda}}{q!} - \lambda^2 \\
&= \lambda. \qquad (43)
\end{aligned}$$

So the variance of Poisson's distribution is also λ. **This means that the variance of the Poisson distribution is equal to its mean.** This is a very useful result.

2.1.4 Poisson distribution: a rule of thumb

Let us see how useful this result is. When counting photons, if the expected number detected is n, the variance of the detected number is n: i.e. we expect typically to detect

$$n \pm \sqrt{n} \qquad (44)$$

photons. Hence, just by detecting n counts, we can immediately say that the uncertainty on that measurement is $\pm\sqrt{n}$, without knowing anything else about the problem, and only assuming the counts are random. This is a very useful result, but beware: when stating an uncertainty like this it will normally be interpreted as the r.m.s. of a *Gaussian* error distribution (see later). Only for large n does the Poisson distribution look Gaussian (the Central Limit Theorem at work again), and we can assume the uncertainty $\sigma = \sqrt{n}$.

2.1.5 Gaussian distribution as a limit of the Poisson distribution

A limiting form of the Poisson distribution (and many others – see the **Central Limit Theorem** below) is the **Gaussian distribution**. In deriving the Poisson distribution we took the limit of the total number of events $N \to \infty$; we now take the limit that the mean value is very large. Let's write the Poisson distribution as

$$P_n = \frac{\lambda^n e^{-\lambda}}{n!}. \tag{45}$$

Now let $x = n = \lambda(1+\delta)$ where $\lambda \gg 1$ and $\delta \ll 1$. Since $\langle n \rangle = \lambda$, this means that we will also be concerned with large values of n, in which case the discrete P_n goes over to a continuous pdf in the variable x. Using Stirling's formula for $n!$:

$$x! \to \sqrt{2\pi x}\, e^{-x} x^x \quad \text{as} \quad x \to \infty \tag{46}$$

we find[7]

$$\begin{aligned} p(x) &= \frac{\lambda^{\lambda(1+\delta)} e^{-\lambda}}{\sqrt{2\pi} e^{-\lambda(1+\delta)} [\lambda(1+\delta)]^{\lambda(1+\delta)+1/2}} \\ &= \frac{e^{\lambda\delta}(1+\delta)^{-\lambda(1+\delta)-1/2}}{\sqrt{2\pi\lambda}} \\ \text{(see footnote)} &= \frac{e^{-\lambda\delta^2/2}}{\sqrt{2\pi\lambda}} \end{aligned} \tag{48}$$

Substituting back for x, with $\delta = (x-\lambda)/\lambda$, yields

$$p(x) = \frac{e^{-(x-\lambda)^2/(2\lambda)}}{\sqrt{2\pi\lambda}} \tag{49}$$

This is a **Gaussian**, or Normal[8], distribution with mean and variance of λ. The Gaussian distribution is the most important distribution in probability, due to its role in the Central Limit Theorem, which loosely says that the sum of a large number of independent quantities tends to have a Gaussian form, independent of the pdf of the individual measurements. The above specific derivation is somewhat cumbersome, and it will actually be more elegant to use the Central Limit theorem to derive the Gaussian approximation to the Poisson distribution.

2.1.6 More on the Gaussian

The Gaussian distribution is so important that we collect some properties here. It is normally written as

$$p(x) = \frac{1}{(2\pi)^{1/2}\sigma}\, e^{-(x-\mu)^2/2\sigma^2}, \tag{50}$$

[7] Maths Notes: The limit of a function like $(1+\delta)^{\lambda(1+\delta)+1/2}$ with $\lambda \gg 1$ and $\delta \ll 1$ can be found by taking the natural log, then expanding in δ to **second** order and using $\lambda \gg 1$:

$$\ln[(1+\delta)^{\lambda(1+\delta)+1/2}] = [\lambda(1+\delta)+1/2]\ln(1+\delta) = (\lambda + 1/2 + \lambda\delta)(\delta - \delta^2/2 + O(\delta^3)) \simeq \lambda\delta + \lambda\delta^2/2 + O(\delta^3) \tag{47}$$

[8] The name 'Normal' was given to this distribution by the statistician K. Pearson, who almost immediately regretted introducing the name. It is also sometimes called the Bell-curve.

so that μ is the mean and σ the standard deviation. The first statement is obvious: consider $\langle x-\mu\rangle$, which must vanish by symmetry since it involves integration of an odd function. To prove the second statement, write

$$\langle (x-\mu)^2\rangle = \frac{1}{(2\pi)^{1/2}\sigma}\,\sigma^3 \int_{-\infty}^{\infty} y^2 e^{-y^2/2}\, dy \tag{51}$$

and do the integral by parts.

Proving that the distribution is correctly normalized is harder, but there is a clever trick, which is to extend to a two-dimensional Gaussian for two independent (zero-mean) variables x and y:

$$p(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}. \tag{52}$$

The integral over both variables can now be rewritten using polar coordinates:

$$\iint p(x,y)\; dx\, dy = \int p(x,y)\; 2\pi\, r\, dr = \frac{1}{2\pi\sigma^2}\int 2\pi\, r\, e^{-r^2/2\sigma^2}\; dr \tag{53}$$

and the final expression clearly integrates to

$$P(r>R) = \exp\left(-R^2/2\sigma^2\right), \tag{54}$$

so the distribution is indeed correctly normalized.

Unfortunately, the single Gaussian distribution has no analytic expression for its integral, even though this is often of great interest. As seen in the next section, we often want to know the probability of a Gaussian variable lying above some value, so we need to know the integral of the Gaussian; there are two common notations for this (assuming zero mean):

$$P(x<X\sigma) = \Phi(X); \tag{55}$$

$$P(x<X\sigma) = \frac{1}{2}\left[1+\mathrm{erf}(X/\sqrt{2})\right], \tag{56}$$

where the **error function** is defined as

$$\mathrm{erf}(y) \equiv \frac{2}{\sqrt{\pi}}\int_0^y e^{-t^2}\; dt. \tag{57}$$

A useful approximation for the integral in the limit of high x is

$$P(x>X\sigma) \simeq \frac{e^{-X^2/2}}{(2\pi)^{1/2}X} \tag{58}$$

(which is derived by a Taylor series: $e^{-(x+\epsilon)^2/2} \simeq e^{-x^2/2}e^{-x\epsilon}$ and the linear exponential in ϵ can be integrated).

2.2 Tails and measures of rareness

So far, we have asked questions about the probability of obtaining a particular experimental outcome, but this is not always a sensible question. Where there are many possible outcomes, the chance of

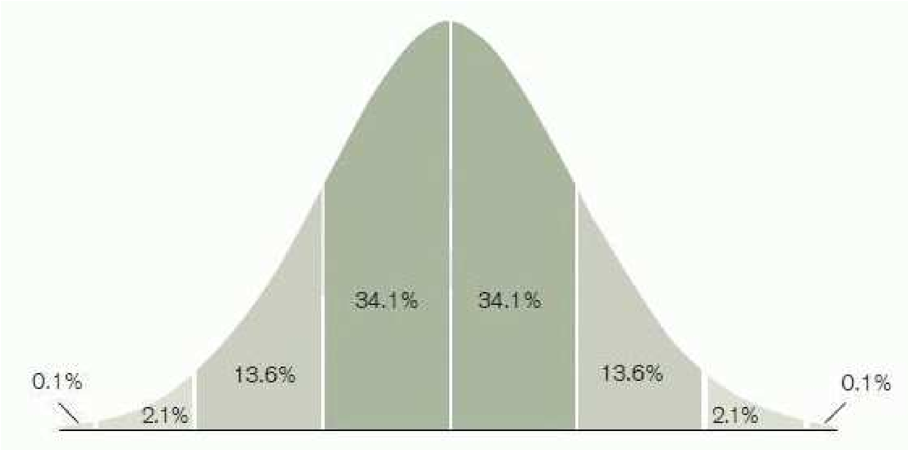

Figure 3: *The Gaussian distribution, illustrating the area under various parts of the curve, divided in units of σ. Thus the chance of being within 1σ of the mean is 68%; 95% of results are within 2σ of the mean; 99.7%of results are within 3σ of the mean.*

any given one happening will be very small. This is most obvious with a continuous pdf, $p(x)$: the probability of a result in the range $x = a$ to $x = a + \delta x$ is $p(a)\delta x$, so the probability of getting $x = a$ exactly is precisely zero. Thus it only makes sense to ask about the probability of x lying in a given range.

The most common reason for calculating the probability of a given outcome (apart from betting), is so we can test hypotheses. The lectures will give an in-depth discussion of this issue later, as it can be quite subtle. Nevertheless, it is immediately clear that we want to set possible outcomes to an experiment in some order of rareness, and we will be suspicious of a hypothesis when an experiment generates an outcome that ought to be very rare if the hypothesis were true.

Informally, we need to define a 'typical' value for x and some means of deciding if x is 'far' from the typical value. For a Gaussian, we would naturally the mean, μ as the typical value and $(x - \mu)/\sigma$ as the distance. But how do we make this general? There are two other common measures of location:

The Mode The value of x where $p(x)$ has a maximum.

The Median The value of x such that $P(> x) = 0.5$.

For the Gaussian, both these measures are equal to the mean. In general, the median is the safest choice, since it is easy to devise pathological examples where the mode or even the mean is not well defined. Following on from the median, we can define **upper and lower quartiles**, which together enclose half the probability – i.e. the values x_1 and x_2, where $P(< x_1) = 0.25$ and $P(> x_2) = 0.75$. This suggests a measure of rareness of events, which is to single out events that lie in the 'tails' of the distribution, where either $P(< x) \ll 1$ or $P(> x) \ll 1$. This can be done in a '1-tailed' or a '2-tailed' manner, depending on whether we choose to count only high excursions, or to be impressed by deviations in either direction.

The area under the tail of a pdf is called a **p value**, to emphasise that we have to be careful with meaning. If we get, say, $p = 0.05$ this means that there is a probability of 0.05 to get a value as extreme as this one, or worse, on a given hypothesis. So we need to have a hypothesis in mind to start with; this is called the **null hypothesis** – it would typically be something non-committal, such as 'there is no signal'. If we get a p value that is small, this is some evidence against the null hypothesis, in which case we can claimed to have detected a signal. Small values of p are described as giving significant evidence against the null hypothesis: for $p = 0.05$ we would say 'this result is significant at the 5% level'.

For completeness, we should mention the term **confidence level**, which is complementary to the significance level. If $P(< x_1) = 0.05$ and $P(> x_2) = 0.05$, then we would say that, on the null hypothesis, $x_1 < x < x_2$ at 90% confidence, or $x < x_2$ at 95% confidence.

As shown later in the course, the p value is not the same as the probability that the null hypothesis is correct, although many people think this is the case. Nevertheless, when p is small, you are on good grounds in disbelieving the null hypothesis.

Some of the p values corresponding to particular places in the Gaussian are listed in table 1. The weakest evidence would be a 2σ result, which happens by chance about 5% of the time. This is hardly definitive, although it may provoke further work. But a 3σ result is much less probable. If you decide to reject the null hypothesis whenever $x - \mu > 3\sigma$, you will be wrong only one time in 3000. Nevertheless, discovery in some areas of science can insist on stronger evidence, perhaps at the 5σ level (1-sided $p = 2.9 \times 10^{-7}$). This is partly because σ may not itself be known precisely, but also because many different experiments can be performed: if we search for the Higgs Boson in 100 different independent mass ranges, we are bound to find a result that is significant at about the 1% level.

Table 1: Tails of the Gaussian

x/σ	1-tail p	2-tail p
1.0	0.159	0.318
2.0	0.0228	0.0456
3.0	0.0013	0.0026

Obviously, this process is not perfect. If we make an observation and get $x - \mu \ll \sigma$, this actually favours a narrower distribution than the standard one, but broader distributions are easier to detect, because the probability of an extreme deviation falls exponentially.

2.3 The likelihood

Although we have argued that the probability of a continuous variable having exactly some given value is zero, the relative probability of having any two values, $p(x = a)/p(x = b)$ is well defined. This can be extended to the idea of relative probabilities for larger datasets, where n drawings are made from the pdf. We can approach this using the **multinomial distribution**, where we extend the binomial to something with more than two possible outcomes: e.g. we toss a six-sided dice seven times, and ask what is the probability of getting three ones, two twoes, a five and a six? Number the possible results of each trial, and say each occurs n_1 times, n_2 times etc., with probabilities p_1, p_2 etc., out of N trials. Imagine first the case where the trials give a string of 1's, followed by 2's etc.

The probability of this happening is $p_1^{n_1} p_2^{n_2} p_3^{n_3} \cdots$. If we don't care which trials give these particular numbers, then we need to multiply by the number of ways in which such an outcome could arise. Imagine choosing the n_1 first, which can happen $C^N_{n_1}$ ways; then n_2 can be chosen $C^{N-n_1}_{n_2}$ ways. Multiplying all these factors, we get the simple result

$$p = \frac{N!}{n_1! n_2! n_3! \cdots} p_1^{n_1} p_2^{n_2} p_3^{n_3} \cdots \tag{59}$$

Now consider the approach to the continuum limit, where all the p's are very small, so that the n's are either 0 or 1. Using bins in x of width δx, $p_i = p(x)\delta x$, so

$$p = N!(\delta x)^N \prod_{i=1}^{N} p(x_i) \equiv N!(\delta x)^N \, \mathcal{L}, \tag{60}$$

where $\mathcal{L}$ is the **likelihood** of the data. Clearly, when we compute the relative probabilities of two different datasets, this is the same as the likelihood ratio.

The likelihood can be used not only to compute the relative probabilities of two different outcomes for the same $p(x)$, but also the relative probabilities of the *same* outcome for two different pdfs. It is therefore a tool that is intimately involved in comparing hypotheses, as discussed more fully later in the course.

2.4 Example problems

We will now go through a number of examples where simple pdfs are applied to real astronomical problems.

2.4.1 Example: Poisson photon statistics

Typically, a star produces a large number, $N \gg 1$, of photons during a period of observation. We only intercept a tiny fraction, $p \ll 1$, of the photons which are emitted in all directions by the star, and if we collect those photons for a few minutes or hours we will collect only a tiny fraction of those emitted throughout the life of the star.

So if the star emits N photons in total and we collect a fraction, p, of those, then t

$$\begin{aligned} \lambda &= Np \text{ (the mean number detected)} \\ N &\to \infty \text{ (the mean total number emitted)} \\ p &\to 0 \text{ (probability of detection is very low)} \end{aligned} \tag{61}$$

So if we make many identical observations of the star and plot out the frequency distribution of the numbers of photons collected each time, we expect to see a *Poisson distribution* (strictly, this is not completely true, as it ignores photon bunching: when the radiation occupation number is high, as in a laser, photons tend to arrive in bursts).

Conversely, if we make one observation and detect n photons, we can use the Poisson distribution to derive the probability of getting this result for all the possible values of λ. The simplest case is when there is no source of 'background' photons (as in e.g. gamma-ray astronomy). In that case, seeing even a single photon is enough to tell us that there is a source there, and the only question is how bright it is. Here, the problem is to estimate the mean arrival rate of photons in a given time interval, λ, given the observed n in one interval. Provided n is reasonably large, we can safely take the Gaussian approach and argue that n will be scattered around λ with variance λ, where λ is close to n. Thus the source flux will be estimated from the observed number of photons, and the fractional error on the flux will be

$$\frac{\sigma_f}{f} = \sigma_{\ln f} = 1/\sqrt{n}. \tag{62}$$

When n is small, we have to be more careful – as discussed later in the section on Bayesian statistics.

2.4.2 Example: sky-limited detection

Typically the counts from the direction of an individual source will be much less than from the surrounding sky, and so our attempted flux measurement always includes sky photons as well as the desired photons from the object. For example, we may detect 5500 photons from an aperture centred on the source, and 5000 from the same sized aperture centred on a piece of 'blank sky'. Have we detected a source?

Let the counts from the aperture on the source be N_T, and from the same area of background sky N_B. N_T includes some background, so our estimate of the source counts N_S is (hat means 'estimate of'):

$$\hat{N_S} = N_T - N_B. \tag{63}$$

The question we want to address is how uncertain is this? The counts are independent and random and so each follow a Poisson distribution, so the variance on N_S is

$$\sigma_S^2 = \sigma_T^2 + \sigma_B^2 = \langle N_T \rangle + \langle N_B \rangle\,. \tag{64}$$

Thus in turn $\hat{\sigma_S^2} = N_T + N_B$. If the source is much fainter than the sky $N_S \ll N_T$, then $N_T \simeq N_B$ and the variance is approximately $2N_B$. Thus the *significance* of the detection, the signal-to-noise ratio, is

$$\text{Signal/Noise} = \frac{N_T - N_B}{\sqrt{N_T + N_B}} \simeq \frac{N_T - N_B}{\sqrt{2N_B}}. \tag{65}$$

So simply measuring the background and the total counts is sufficient to determine if a detection is made. In the above example, Signal/Noise $\simeq 500/\sqrt{10000} = 5$ (strictly, slightly less), what we would call 'a 5σ detection'.

Normally 3σ ($p \sim 0.001$) gives good evidence for a detection – but only if the position is known in advance. When we make a survey, every pixel in an image is a candidate for a source, although most of them will be blank in reality. Thus the number of trials is very high; to avoid being swamped by false positives, surveys will normally set a threshold around 5σ (for related reasons, this is the traditional threshold used by particle physicists when searching for e.g. the Higgs Boson).

2.4.3 Example: The distribution of superluminal velocities in quasars.

Some radio sources appear to be expanding faster than the speed of light. This is thought to occur if a radio-emitting component in the quasar jet travels almost directly towards the observer at a speed close to that of light. The effect was predicted by the Astronomer Royal Lord Martin Rees in 1966 (when he was an undergraduate), and first observed in 1971.

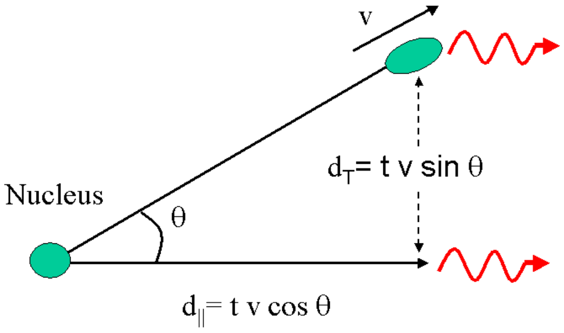

Figure 4: *Superluminal motion from a quasar nucleus.*

Suppose the angle to the line of sight is θ as shown above, and that a component is ejected along the jet from the nucleus. at $t = 0$ After some time t the ejection component has travelled a distance $d_{||} = tv\cos\theta$ along the line of sight. But the initial ejection is seen to happen later than $t = 0$, owing to the light travel time from the nucleus, so the observed duration is $\Delta t = t - d_{||}/c = t(1-(v/c)\cos\theta)$. In that time the component appears to have moved a distance $d_\perp = tv\sin\theta$ across the line of sight, and hence the apparent transverse velocity of the component is

$$v' = \frac{d_\perp}{\Delta t} = \frac{v\sin\theta}{1-(v/c)\cos\theta}. \tag{66}$$

Note that although a v/c term appears in this expression, the effect is not a relativistic effect. It is just due to light delay and the viewing geometry. Writing

$$\beta = v/c, \qquad \gamma = (1-\beta^2)^{-1/2} \tag{67}$$

we find that the apparent transverse speed β' has a maximum value when

$$\frac{\partial\beta'}{\partial\theta} = -\frac{\beta(\beta-\cos\theta)}{(1-\beta\cos\theta)^2} = 0, \tag{68}$$

when $\cos\theta = \beta$. Since $\sin\theta = 1/\gamma$ we find a maximum value of $\beta' = \gamma\beta$, where γ can be much greater than unity.

Given a randomly oriented sample of radio sources, what is the expected distribution of β' if β is fixed?

First, note that θ is the angle to the line of sight, and since the orientation is random in three dimensions (i.e. uniform distribution over the area $dA = \sin\theta\, d\theta d\phi$),

$$p(\theta) = \sin\theta \qquad 0 \le \theta \le \pi/2. \tag{69}$$

Hence,

$$\begin{aligned} p(\beta') &= p(\theta)\left|\frac{d\theta}{d\beta'}\right| = p(\theta)\left|\frac{d\beta'}{d\theta}\right|^{-1} \\ &= \frac{\sin\theta(1-\beta\cos\theta)^2}{|\beta\cos\theta-\beta^2|} \end{aligned} \tag{70}$$

where $\sin\theta$ and $\cos\theta$ are given by the equation for β'. We have chosen the limits $0 \le \theta \le \pi/2$ because in the standard model blobs are ejected from the nucleus along a jet in two opposite directions, so we should always see one blob which is travelling towards us. The limits in β' are $0 \le \beta' \le \gamma\beta$. The expression for $p(\beta')$ in terms of β' alone is rather messy, but simplifies for $\beta \to 1$:

$$\beta' = \frac{\sin\theta}{(1-\cos\theta)} \tag{71}$$

$$p(\beta') = \sin\theta(1-\cos\theta). \tag{72}$$

Squaring both sides of $\sin\theta = \beta'(1-\cos\theta)$, using $\sin^2\theta = (1-\cos\theta)(1+\cos\theta)$ and rearranging gives us $(1-\cos\theta) = 2/(1+\beta'^2)$. Substituting this and $\sin\theta = \beta'(1-\cos\theta)$ into equation (72) finally gives us

$$p(\beta') = \frac{4\beta'}{(1+\beta'^2)^2} \qquad \beta' \ge 1. \tag{73}$$

The cumulative probability for β' is

$$P(>\beta') = \frac{2}{(1+\beta'^2)} \qquad \beta' \ge 1. \tag{74}$$

so the probability of observing a large apparent velocity, say $\beta' > 5$, is $P(\beta' > 5) \approx 1/13$.

In fact, a much larger fraction of powerful radio quasars show superluminal motions, and it now seems likely that the quasars jets cannot be randomly oriented: There must be effects operating which tend to favour the selection of quasars jets pointing towards us, most probably due to an opaque disc shrouding the nucleus. Another physical effect is that jets pointing towards us at speeds close to c have their fluxes boosted by *relativistic beaming*, which means they would be favoured in a survey which was selected on the basis of their radio flux. This can be avoided by choosing the sample on the basis of some flux which is not beamed, unrelated to the jet.

3 Addition of random variables to Central Limit Theorem

3.1 Probability distribution of summed independent random variables

Let us consider the distribution of the sum of two or more random variables: this will lead us on to the **Central Limit Theorem** which is of critical importance in probability theory and hence astrophysics.

Let us define a new random variable $z = x + y$. What is the probability density, $p_z(z)$ of z? The probability of observing a value, z, which is greater than some value z_1 is

$$\begin{aligned} P(z \geq z_1) &= \int_{z_1}^{\infty} d\tilde{z}\, p_z(\tilde{z}) && (75) \\ &= \int_{-\infty}^{\infty} dy \int_{z_1-y}^{\infty} dx\, p_{x,y}(x,y) && (76) \\ &= \int_{-\infty}^{\infty} dx \int_{z_1-x}^{\infty} dy\, p_{x,y}(x,y) && (77) \end{aligned}$$

where the integral limits on the second line can be seen from defining the region in the $x - y$ plane (see Figure 5).

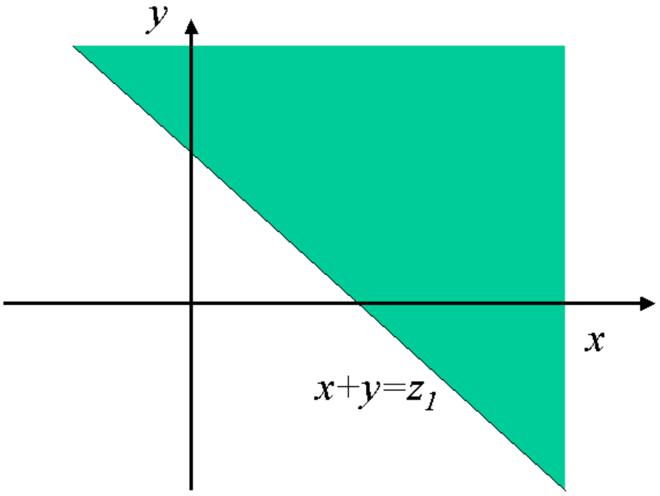

Figure 5: *The region of integration of equation (76).*

Now, the pdf for z is just the derivative of this integral probability:

$$p_z(z) = -dP(> z)/dz = \int_{-\infty}^{\infty} dx\, p_{x,y}(x, z - x). \qquad (78)$$

Or we could do the 2D integral in the opposite order, which would give

$$p_z(z) = \int_{-\infty}^{\infty} dy\, p_{x,y}(z - y, y) \qquad (79)$$

If the distributions of x and y are **independent**, then we arrive at a particularly important result:

$$p_z(z) = \int_{-\infty}^{\infty} dx\, p_x(x) p_y(z-x) \quad \text{or} \quad \int_{-\infty}^{\infty} dy\, p_x(z-y) p_y(y) \qquad \text{i.e.} \qquad p_z(z) = (p_x * p_y)(z) \tag{80}$$

If we add together two independent random variables, the resulting distribution is a convolution of the two distribution functions.

The most powerful way of handling convolutions is to use **Fourier transforms** (FTs), since the FT of a convolved function $p(z)$ is simply the product of the FTs of the separate functions $p(x)$ and $p(y)$ being convolved, i.e. for the sum of two independent random variables we have:

$$\mathcal{F}(p_z) = \mathcal{F}(p_x)\, \mathcal{F}(p_y) \tag{81}$$

3.2 Characteristic functions

In probability theory the Fourier Transform of a probability distribution function is known as the **characteristic function**:

$$\phi(k) = \int_{-\infty}^{\infty} dx\, p(x) e^{ikx} \tag{82}$$

with reciprocal relation

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk\, \phi(k) e^{-ikx} \tag{83}$$

(note that other Fourier conventions may put the factor 2π in a different place). A discrete probability distribution can be thought of as a continuous one with delta-function spikes of weight p_i at locations x_i, so here the characteristic function is

$$\phi(k) = \sum_i p_i e^{ikx_i} \tag{84}$$

(note that $\phi(k)$ is a continuous function of k). Hence in all cases the characteristic function is simply the expectation value of e^{ikx}:

$$\phi(k) = \left\langle e^{ikx} \right\rangle. \tag{85}$$

Part of the power of characteristic functions is the ease with which one can generate all of the moments of the distribution by differentiation:

$$m_n = (-i)^n \frac{d^n}{dk^n} \phi(k) \bigg|_{k=0} \tag{86}$$

This can be seen if one expands $\phi(k)$ in a power series:

$$\phi(k) = \left\langle e^{ikx} \right\rangle = \left\langle \sum_{n=0}^{\infty} \frac{(ikx)^n}{n!} \right\rangle = \sum_{n=0}^{\infty} \left\langle \frac{(ikx)^n}{n!} \right\rangle = 1 + ik \langle x \rangle - \frac{1}{2} k^2 \left\langle x^2 \right\rangle + \cdots. \tag{87}$$

As an example of a characteristic function let us consider the Poisson distribution:

$$\phi(k) = \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} e^{ikn} = e^{-\lambda} e^{\lambda e^{ik}}, \tag{88}$$

so that the characteristic function for the Poisson distribution is

$$\phi(k) = e^{\lambda(e^{ik}-1)}. \tag{89}$$

The first moments of the Poisson distribution are:

$$\begin{aligned}
m_0 &= \phi(k)\Big|_{k=0} = e^{\lambda(e^{ik}-1)}\Big|_{k=0} = 1 && (90)\\
m_1 &= (-i)^1 \frac{d^1}{dk^1}\phi(k)\Big|_{k=0} = (-i)e^{\lambda(e^{ik}-1)}\lambda e^{ik}(i)\Big|_{k=0} = \lambda && (91)\\
m_2 &= (-i)^2 \frac{d^2}{dk^2}\phi(k)\Big|_{k=0} = (-1)\frac{d^1}{dk^1}e^{\lambda(e^{ik}-1)}\lambda e^{ik}(i)\Big|_{k=0} = \cdots = \lambda(\lambda+1) && (92)
\end{aligned}$$

which is in total agreement with results found for the mean and the variance of the Poisson distribution (see Eq. 41 and Eq. 43).

Returning to the convolution equation (80),

$$p_z(z) = \int_{-\infty}^{\infty} dy\, p_x(z-y)p_y(y), \tag{93}$$

we shall identify the characteristic function of $p_z(z)$, $p_x(x)$, $p_y(y)$ as $\phi_z(k)$, $\phi_x(k)$ and $\phi_y(k)$ respectively. The characteristic function of $p_z(z)$ is then

$$\begin{aligned}
\phi_z(k) &= \int_{-\infty}^{\infty} dz\, p_z(z)e^{ikz}\\
&= \int_{-\infty}^{\infty} dz \int_{-\infty}^{\infty} dy\, p_x(z-y)p_y(y)e^{ikz}\\
&= \int_{-\infty}^{\infty} dz \int_{-\infty}^{\infty} dy\, [p_x(z-y)e^{ik(z-y)}][p_y(y)e^{iky}]\\
(\text{let } x = z-y) &= \int_{-\infty}^{\infty} dx\, p_x(x)e^{ikx} \int_{-\infty}^{\infty} dy\, p_y(y)e^{iky}. && (94)
\end{aligned}$$

which is an explicit proof of the convolution theorem for the product of Fourier transforms:

$$\phi_z(k) = \phi_x(k)\,\phi_y(k). \tag{95}$$

The power of this approach is that the distribution of the sum of a large number of random variables can be easily derived. This result allows us to turn now to the Central Limit Theorem.

3.3 The Central Limit Theorem

The most important, and general, result from probability theory is the **Central Limit Theorem**. It applies to a wide range of phenomena and explains why the Gaussian distribution appears so often in Nature.

In its most general form, the Central Limit Theorem states that the sum of n random values drawn from a probability distribution function of finite variance, σ^2, tends to be Gaussian distributed about the expectation value for the sum, with variance $n\sigma^2$.

There are two important consequences:

1. The mean of a large number of values tends to be normally distributed regardless of the probability distribution from which the values were drawn. **Hence the sampling distribution is known even when the underlying probability distribution is not.** It is for this reason that the Gaussian distribution occupies such a central place in statistics. It is particularly important in applications where underlying distributions are not known, such as astrophysics.
2. Functions such as the Binomial and Poisson distributions arise from multiple drawings of values from some underlying probability distribution, and they all tend to look like the Gaussian distribution in the limit of large numbers of drawings. We saw this earlier when we derived the Gaussian distribution from the Poisson distribution.

The first of these consequences means that under certain conditions **we can assume an unknown distribution is Gaussian**, if it is generated from a large number of events with finite variance. For example, the height of the surface of the sea has a Gaussian distribution, as it is perturbed by the sum of random winds.

But it should be borne in mind that the number of influencing factors will be finite, so the Gaussian form will not apply exactly. It is often the case that a pdf will be approximately Gaussian in its core, but with increasing departures from the Gaussian as we move into the tails of rare events. Thus the probability of a 5σ excursion might be much greater than a naive Gaussian estimate; it is sometimes alleged that neglect of this simple point by the banking sector played a big part in the financial crash of 2008. A simple example of this phenomenon is the distribution of human heights: a Gaussian model for this must fail, since a height has to be positive, whereas a Gaussian extends to $-\infty$.

3.3.1 Derivation of the central limit theorem

Let

$$X = \frac{1}{\sqrt{n}}(x_1 + x_2 + \cdots + x_n) = \sum_{j=1}^{n} \frac{x_j}{\sqrt{n}} \tag{96}$$

be the sum of n random variables x_j, each drawn from the same arbitrary underlying distribution function, p_x. In general the underlying distributions can all be different for each x_j, but for simplicity we shall consider only one here. The distribution of X's generated by this summation, $p_X(X)$, will be a convolution of the underlying distributions.

From the properties of characteristic functions we know that a convolution of distribution functions is a multiplication of characteristic functions. If the characteristic function of $p_x(x)$ is

$$\phi_x(k) = \int_{-\infty}^{\infty} dx\, p_x(x) e^{ikx} = 1 + i \langle x \rangle k - \frac{1}{2} \left\langle x^2 \right\rangle k^2 + O(k^3), \tag{97}$$

where in the last term we have expanded out e^{ikx}. Since the sum is over $x_j/\sqrt{n}$, rather than x_j we scale all the moments $\langle x^p \rangle \to \langle (x/\sqrt{n})^p \rangle$. From equation (97),we see this is the same as scaling $k \to k/\sqrt{n}$. Hence the characteristic function of X is

$$\Phi_X(k) = \prod_{j=1}^{n} \phi_{x_j/\sqrt{n}}(k) = \prod_{j=1}^{n} \phi_{x_j}(k/\sqrt{n}) = [\phi_x(k/\sqrt{n})]^n \tag{98}$$

If we assume that $m_1 = \langle x \rangle = 0$, so that $m_2 = \langle x^2 \rangle = \sigma_x^2$ (this doesn't affect our results) then

$$\Phi_X(k) = \left[1 + i\frac{m_1 k}{\sqrt{n}} - \frac{\sigma_x^2 k^2}{2n} + O\left(\frac{k^3}{n^{3/2}}\right)\right]^n = \left[1 - \frac{\sigma_x^2 k^2}{2n} + O\left(\frac{k^3}{n^{3/2}}\right)\right]^n \to e^{-\sigma_x^2 k^2/2} \tag{99}$$

as $n \to \infty$. Note the higher terms contribute as $n^{-3/2}$ in the expansion of $\Phi_X(k)$ and so vanish in the limit of large n – where we have previously seen how to treat the limit of the expression $(1 + a/n)^n$.

It is however important to note that we have made a critical assumption: all the moments of the distribution, however high order, must be finite. If they are not, then the higher-order terms in k will not be negligible, however much we reduce them as a function of n. It is easy to invent distributions for which this will be a problem: a power-law tail to the pdf may allow a finite variance, but sufficiently high moments will diverge, and our proof will fail. In fact, the Central Limit theorem still holds even in such cases – it is only necessary that the variance be finite – but we are unable to give a simple proof of this here.

The above proof gives the characteristic function for X. We know that the F.T. of a Gaussian is another Gaussian, but let us show that explicitly:

$$\begin{aligned}
p_X(X) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} dk\, \Phi_X(k) e^{-ikX} \\
&= \frac{e^{-X^2/(2\sigma_x^2)}}{2\pi} \int_{-\infty}^{\infty} dk\, e^{(-\sigma_x^2 k^2 + X^2/\sigma_x^2 - 2ikX)/2} \\
&= \frac{e^{-X^2/(2\sigma_x^2)}}{2\pi} \int_{-\infty}^{\infty} dk\, e^{-(k\sigma_x + iX/\sigma_x)^2/2} = \frac{e^{-X^2/(2\sigma_x^2)}}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi}\sigma_x} \int_{-\infty}^{\infty} dk'\, e^{-(k' + iX)^2/(2\sigma_x^2)} \\
&= \frac{e^{-X^2/(2\sigma_x^2)}}{\sqrt{2\pi}\sigma_x}.
\end{aligned} \tag{100}$$

Thus the sum of random variables, sampled from the same underlying distribution, will tend towards a Gaussian distribution, independently of the initial distribution.

The variance of X is evidently the same as for x: σ_x^2. The variance of the mean of the x_j is then clearly smaller by a factor n, since the mean is $X/\sqrt{n}$.

3.3.2 Measurement theory

As a corollary, by comparing equation (96) with the expression for estimating the mean from a sample of n independent variables,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{101}$$

we see that the estimated mean from a sample has a Gaussian distribution with mean $m_1 = \langle x \rangle$ and **standard error on the mean**

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \tag{102}$$

as $n \to \infty$.

This has two important consequences.

1. This means that if we estimate the mean from a sample, we will always tend towards the true mean,
2. The uncertainty in our estimate of the mean will decrease as the sample gets bigger.

This is a remarkable result: for sufficiently large numbers of drawings from an unknown distribution function with mean $\langle x \rangle$ and standard deviation $\sigma/\sqrt{n}$, we are assured by the Central Limit Theorem that we will get the measurement we want to higher and higher accuracy, and that the estimated mean of the sampled numbers will have a Gaussian distribution almost regardless of the form of the unknown distribution. The only condition under which this will not occur is if the unknown distribution does not have a finite variance. **Hence we see that all our assumptions about measurement rely on the Central Limit Theorem.**

3.3.3 How the Central Limit Theorem works

We have seen from the above derivation that the Central Limit Theorem arises because in making many measurements and averaging them together, we are convolving a probability distribution with itself many times.

We have shown that this has the remarkable mathematical property that in the limit of large numbers of such convolutions, the result always tends to look Gaussian. In this sense, the Gaussian, or normal, distribution is the 'smoothest' distribution which can be produced by natural processes.

We can show this by considering a non-Gaussian distribution, ie a top-hat, or square distribution (see Figure 6). If we convolve this with itself, we get a triangle distribution. Convolving again we get a slightly smoother distribution. If we keep going we will end up with a Gaussian distribution. This is the Central Limit Theorem and is the reason for its ubiquitous presence in nature.

3.4 Sampling distributions

Above we showed how the Central Limit Theorem lies at the root of our expectation that more measurements will lead to better results. Our estimate of the mean of n variables is unbiased (ie

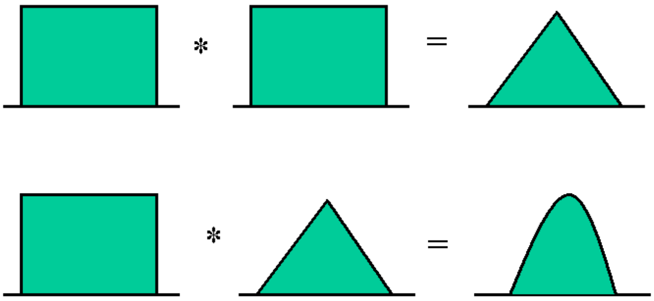

Figure 6: *Repeated convolution of a distribution will eventually yield a Gaussian if the variance of the convolved distribution is finite.*

gives the right answer) and the uncertainty on the estimated mean decreases as $\sigma_x/\sqrt{n}$, and the distribution of the estimated, or sampled, mean has a Gaussian distribution. The distribution of the mean determined in this way is known as the '**sampling distribution** of the mean'.

How fast the Central Limit Theorem works (i.e. how small n can be before the distribution is no longer Gaussian) depends on the underlying distribution. At one extreme we can consider the case of when the underlying variables are all Gaussian distributed. Then the sampling distribution of the mean will always be a Gaussian, even if $n \to 1$.

But, beware! For some distributions the Central Limit Theorem does not hold. For example the means of values drawn from a **Cauchy (or Lorentz) distribution**,

$$p(x) = \frac{1}{\pi(1+x^2)} \tag{103}$$

never approach normality. This is because this distribution has infinite variance (try and calculate it and see). In fact they are distributed like the Cauchy distribution. Is this a rare, but pathological example ? Unfortunately not. For example the Cauchy distribution appears in spectral line fitting, where it is called the Voigt distribution. Another example is if we take the ratio of two Gaussian variables. The resulting distribution has a Cauchy distribution. Hence, we should beware, that although the Central Limit Theorem and Gaussian distribution considerably simplify probability and statistics, exceptions do occur, and one should always be wary of them.

3.5 Error propagation

If z is some function of random variables x and y, and we know the variance of x and y, what is the variance of z?

Let $z = f(x,y)$. We can *propagate errors* by expanding $f(x,y)$ to first order around some arbitrary values, x_0 and y_0;

$$f(x,y) = f(x_0,y_0)+(x-x_0)\frac{\partial f}{\partial x}\bigg|_{x=x_0,y=y_0}+(y-y_0)\frac{\partial f}{\partial y}\bigg|_{x=x_0,y=y_0}+0((x-x_0)^2,(x-x_0)(y-y_0),(y-y_0)^2) \tag{104}$$

Let us assume $x_0 = y_0 = 0$ and $\langle x\rangle = \langle y\rangle = 0$ for simplicity (the answer will be general).

The mean of z is

$$\langle z\rangle = f(x_0,y_0) \tag{105}$$

and the variance is (assuming x and y are independent)

$$\begin{aligned} \sigma_z^2 &= \left\langle (z-\langle z\rangle)^2\right\rangle \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdy\,(f-\langle f\rangle)^2 p(x)p(y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdy\,(x^2f_x^2+y^2f_y^2+2xyf_xf_y)p(x)p(y) \end{aligned} \tag{106}$$

where we have used the notation $f_x \equiv \frac{\partial f}{\partial x}\Big|_{x=x_0,y=y_0}$.

Averaging over the random variables we find for independent variables with zero mean:

$$\sigma_z^2 = \left(\frac{\partial f}{\partial x}\right)^2\sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2\sigma_y^2. \tag{107}$$

This formula will allow us to propagate errors for arbitrary functions. Note again that this is valid for any distribution function, but depends on (1) the underlying variables being independent, (2) the function being differentiable and (3) the variation from the mean being small enough that the expansion is valid.

3.5.1 The sample variance

The *average* of the sample

$$\hat{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{108}$$

is an estimate of the mean of the underlying distribution. Given we may not directly know the variance of the summed variables, σ_x^2, is there a similar estimate of the variance of $\hat{x}$? This is particularly important in situations where we need to assess the significance of a result in terms of how far away it is from the expected value, but where we only have a finite sample size from which to measure the variance of the distribution.

We would expect a good estimate of the population variance would be something like

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2, \tag{109}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{110}$$

is the sample mean of n values. Let us find the expected value of this sum. First we re-arrange the summation

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum x_i^2 - \frac{2}{n^2}\sum_i\sum_k x_i x_k + \frac{1}{n^2}\sum_i\sum_k x_i x_k = \frac{1}{n}\sum x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \tag{111}$$

which is the same result we found in Section 1.4 – the variance is just the mean of the square minus the square of the mean. If all the x_i are drawn independently then

$$\left\langle \sum_i f(x_i) \right\rangle = \sum_i \langle f(x_i) \rangle \tag{112}$$

where $f(x)$ is some arbitrary function of x. If $i = j$ then

$$\langle x_i x_j \rangle = \left\langle x^2 \right\rangle \qquad\qquad i = j, \tag{113}$$

and when i and j are different

$$\langle x_i x_j \rangle = \langle x \rangle^2 \qquad\qquad i \neq j. \tag{114}$$

The expectation value of our estimator is then

$$\begin{aligned}
\left\langle S^2 \right\rangle &= \left\langle \frac{1}{n}\sum x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \right\rangle \\
&= \frac{1}{n}\sum \left\langle x^2 \right\rangle - \frac{1}{n^2}\sum_i\sum_{j\neq i} \langle x \rangle^2 - \frac{1}{n^2}\sum \left\langle x^2 \right\rangle \\
&= \left\langle x^2 \right\rangle - \frac{n(n-1)}{n^2}\langle x \rangle^2 - \frac{n}{n^2}\left\langle x^2 \right\rangle \\
&= \left(1 - \frac{1}{n}\right)\left\langle x^2 \right\rangle - \frac{n(n-1)}{n^2}\langle x \rangle^2 \\
&= \frac{(n-1)}{n}(\left\langle x^2 \right\rangle - \langle x \rangle^2) = \frac{(n-1)}{n}\sigma_x^2.
\end{aligned} \tag{115}$$

The variance is defined as $\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2$, so S^2 will underestimate the variance by the factor $(n-1)/n$. This is because an extra variance term, σ_x^2/n, has appeared due to the extra variance in our estimate in the mean. Since the square of the mean is subtracted from the mean of the square, this extra variance is subtracted off from our estimate of the variance, causing the underestimation. To correct for this we should change our estimate to

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{116}$$

which is an unbiased estimate of σ_x^2, independent of the underlying distribution. It is unbiased because its expectation value is always σ_x^2 for any n when the mean is estimated from the sample.

Note that if the mean is known, and not estimated from the sample, this extra variance does not appear, in which case equation (109) is an unbiased estimate of the sample variance.

3.5.2 Example: Measuring quasar variation

We want to look for variable quasars. We have two CCD images of one field taken some time apart and we want to pick out the quasars which have varied significantly more than the measurement error which is unknown.

In this case

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\Delta m_i - \overline{\Delta m})^2 \tag{117}$$

is the unbiased estimate of the variance of Δm. We want to keep $\overline{\Delta m} \neq 0$ (i.e. we want to measure $\overline{\Delta m}$ from the data) to allow for possible calibration errors. If we were confident that the calibration is correct we can set $\overline{\Delta m} = 0$, and we could return to the definition

$$\sigma^2 = \frac{1}{n}\sum_{i}(\Delta m)^2. \tag{118}$$

Suppose we find that one of the Δm, say Δm_i, is very large, can we assess the significance of this result? One way to estimate its significance is from

$$t = \frac{\Delta m_i - \overline{\Delta m}}{S}. \tag{119}$$

If the mean is know this is distributed as a standardised Gaussian (ie t has unit variance) if the measurement errors are Gaussian.

But if we can only estimate the mean from the data, t is distributed as Student-t. The Student-t distribution looks qualitatively similar to a Gaussian distribution, but it has larger tails, due to the variations in the measured mean and variance. In other words, Student-t is the pdf that arises when estimating the mean of a Gaussian distributed population when the sample size is small.

PART TWO

4 Bayesian statistics

4.1 Bayesian inference

We now return to the unsolved problem encountered earlier. This is that it is easy enough to calculate the **forward probability** $p(D|\theta)$, where D denotes the data resulting from an experiment, and θ represents one or more parameters in a model (e.g. $D = n$ in a Poisson experiment, where $\theta = \lambda \equiv \langle n \rangle$). But often we are actually more interested in the **inverse probability** $p(\theta|D)$ – e.g. what confidence limits would we set on a source flux density λ given the detection of n photons in a given time interval? The solution to this problem was given by Thomas Bayes in 1763, and lies at the heart of the **Bayesian** approach to statistics.

We have already met Bayes' theorem, expressed as a consequence of the two ways of expressing the joint probability of getting events x and y in terms of conditional probabilities:

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x) \Rightarrow p(y|x) = p(x|y)p(y)/p(x). \tag{120}$$

Thinking of x and y in a frequentist sense as outcomes to experiments such as drawing balls from urns, this seems fairly unremarkable as a statement about probabilities as frequencies. But the radical content of Bayes' theorem is revealed when we start to think of probabilities as also representing degrees of belief. Specifically, suppose x corresponds to obtaining some measurement value or values, D, and y corresponds to some theory parameter or parameters, θ, having a given value: we then see that Bayes' theorem solves the inverse probability problem:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}. \tag{121}$$

For future reference, we further note that this formulation will contain further unspoken information, I, in addition to parameter values θ – e.g. the assumption that we are dealing with a Poisson distribution of some mean λ. Therefore a fuller way of writing the above relation is

$$p(\theta|D, I) = \frac{p(D|\theta, I)p(\theta|I)}{p(D|I)}, \tag{122}$$

although we will often choose not to display the assumption label explicitly.

In summary, Bayes' theorem in this form allows us to make the critical switch from calculating the probability of various possible outcomes on a given hypothesis, to the probability of various hypotheses given the data. This is much more satisfactory, since it removes the need to think about all the other possible outcomes to the experiment and focus on what we learn from the one that actually happened.

There are various formal names for the terms in Bayes' theorem:

- $p(\theta|D)$ is the **posterior probability**, which represents the state of knowledge of about the system in light of the data.

- $p(D|\theta)$ is the **likelihood** of the data given a hypothesis. For many data drawn from the same pdf, the likelihood is $\mathcal{L} = \prod_i p(x_i)$.
- $p(\theta)$ is the **prior**, this contains all the information we know about the probability of a given parameter before the experiment.
- $p(D)$ is the **evidence**. This constant is irrelevant for the relative probability of competing hypotheses, but it can be computed using the theorem of total probability, summing the conditional probability of D over all hypotheses: $P(D) = \int P(D|\theta)\, p(\theta)\, d\theta$.

4.1.1 The problem of the prior

Bayes' Theorem is a powerful piece of statistical apparatus, but it can also be controversial, precisely because we are explicitly dealing with states of belief. It may be clear enough how we calculate the likelihood (e.g. probability of getting $n = 5$ in a Poisson experiment with $\lambda = 3$), but how do we set the prior distribution for λ? What if two different scientists disagree over what it should be?

One answer to this is that one of the main applications of the Bayesian formula is to perform **inference**: i.e. to estimate the parameter(s) θ, together with a measure of uncertainty on them. As the quantity of data increases, the posterior probability for θ becomes concentrated around the true value, and the form of the prior becomes unimportant, provided it can be treated as constant over some small range of θ. Nevertheless, different priors will have different effects for finite amounts of data, and it is important in any analysis to be explicit about the prior being used. Two special cases are commonly encountered:

Uniform prior. This is exactly what it implies: $p(\theta)$ is a constant. There is a problem with normalizing this if θ varies over an infinite range, but this is not a difficulty in practice, since it is only the posterior distribution that we need to normalize, and the likelihood term in Bayes' Theorem will cut off extreme values of θ.

Jeffreys prior. The uniform prior is appropriate for a quantity that can range without restriction between $-\infty$ and $+\infty$ (the mean of a Gaussian distribution, for example), but sometimes we have a positivity constraint (e.g. the σ parameter in a Gaussian). To avoid an abrupt discontinuity, one common argument is that we should adopt a prior that is uniform in $\ln\theta$, i.e. $p(\theta) \propto 1/\theta$.

To see the difference that these choices make, take a Bayesian look at the problem of estimating the mean and variance from a set of data. Suppose we have a set of N measurements x_i, which we assume are random drawings from a Gaussian distribution with mean μ and variance σ^2. The likelihood $\mathcal{L} \equiv p(D|\theta)$ is just

$$\ln \mathcal{L} = -\sum_i (x_i - \mu)^2/2\sigma^2 - (N/2)\ln[2\pi\sigma^2]. \tag{123}$$

Differentiating to maximize the posterior probability with a uniform prior gives

$$\begin{aligned} \mu_{\rm ML} &= \frac{1}{N}\sum_i x_i \\ \sigma^2_{\rm ML} &= \frac{1}{N}\sum_i (x_i - \mu_{\rm ML})^2. \end{aligned} \tag{124}$$

The trouble with this result is that it is known to be biased low: $\langle \sigma^2_{\rm ML} \rangle = [(N-1)/N]\sigma^2_{\rm true}$. If we adopt a Jeffreys prior and set the posterior to be $p(\mu, \sigma|D) \propto \mathcal{L}(\mu, \sigma|D)/\sigma$, it is easy to repeat the differentiation and show that the problem gets worse: the maximum-posterior value of σ^2 is lower by a factor $N/(N+1)$. If we wanted an unbiased estimator, we would need a prior $\propto \sigma$ (reasonably enough, in retrospect, this says that equal increments in variance are equally likely). But this is backwards: the prior should express a primitive belief, rather than trying out various options until we get one we like. This illustrates why the prior can be a controversial part of the Bayesian apparatus. But it cannot be ignored: statistical work often uses maximum-likelihood quantities, without stating that this implicitly assumes a uniform prior, and we we should be clear that this is what is happening.

4.1.2 Bayes as a rule for responding to data

A further way of dealing with the problem of the arbitrariness of the prior is to consider what happens when we take two sets of data, D_1 and D_2. The posterior distribution in the face of the totality of data is

$$p(\theta|D_1, D_2) \propto p(D_1, D_2|\theta)p(\theta). \tag{125}$$

But the likelihood will factor into the likelihoods for the two datasets separately: $p(D_1, D_2|\theta) = p(D_1|\theta)p(D_2|\theta)$, in which case the posterior can be written as

$$p(\theta|D_1, D_2) \propto p(D_2|\theta)\,[p(D_1|\theta)p(\theta)]. \tag{126}$$

Now we see that the posterior for experiment 1 functions as the prior for experiment 2, so that Bayes' Theorem expresses the process of updating our belief about θ in the light of new information. This adaptability is a natural and appealing aspect of the framework – although it does not remove the basic worry of how we should set a prior before we have any data at all.

4.2 Bayesian hypothesis testing

The discussion so far has focused on a single model, in which there is some unknown parameter. But frequently we have one or more distinct possibilities in mind, and want to make a discrimination between them. A simple example would be detection of an astronomical source. We measure a flux with some noise, and want to compare two hypotheses: H_0 (true flux is zero: no object is present); H_1 (there is a source, with true flux S). It is interesting to contrast how we would treat this problem from a frequentist and a Bayesian point of view.

The frequentist approach is focused on H_0: if this is not favoured by the data, then the only other possibility is that a source is present, so we can claim a detection (the Sherlock Holmes approach). As we know, this is done by looking at the signal-to-noise ratio for the measurement and evaluating a p-value for the tail probability. If we get a 3σ result ($p \sim 0.001$), then H_0 would commonly be rejected.

But from a Bayesian point of view, this is unsatisfactory, since we have made no statement about the prior probability of H_0. We should compute the posterior probability of H_0 being true:

$$p(H_0|D) = p(D|H_0)p(H_0)/p(D). \tag{127}$$

But the normalizing evidence, $p(D)$ involves summing over all possibilities, which includes the case that H_1 is the true hypothesis. It is therefore easier to write the same equation for H_1 and take the ratio, to get the **posterior odds** on the two hypotheses:

$$\frac{p(H_0|D)}{p(H_1|D)} = \frac{p(D|H_0)}{p(D|H_1)} \times \frac{p(H_0)}{p(H_1)}, \tag{128}$$

which we see is the product of the **likelihood ratio** and the **prior ratio**. This can yield some big differences with the frequentist approach. Suppose the prior ratio is very large (i.e. we expect that detections will be rare); in this case, a very large likelihood ratio will be required to convince us that a detection has really been made.

Furthermore, the likelihood ratio depends on the true flux, S. To see how this complicates things, suppose we were agnostic about the hypotheses, and gave them equal prior probability. If we now choose units such that the measuring noise is unity, then the posterior ratio is the same as the likelihood ratio, which is just the ratio of two Gaussian terms:

$$\frac{p(H_0|D)}{1 - p(H_0|D)} = \exp(-D^2/2 + (D-S)^2/2). \tag{129}$$

If S is very small, this says that the posterior odds are unity: reasonably enough, no data can distinguish an infinitesimal source from no source. But note that if $S \gg D$, then H_1 is disfavoured. The best case for H_1 is if $S = D$, in which case the probability of H_0 being true is $[1 + \exp(D^2/2)]^{-1}$. This is much less severe than the p-value approach. For a 3σ result, the Bayesian probability is 0.011, rather than the traditional 0.001. This is a further reason why one should be suspicious of few-sigma 'detections', beyond the point mentioned previously that such claims can arise for the single 'significant' result found in a number of independent trials.

4.2.1 Bayesian model selection (non-examinable)

The unsatisfactory feature of the above discussion is the issue of what value the unknown parameter S, the true flux, might take. Surely we can ask about the probability of a detection having been made without being specific about S? In fact, we can: this is the method of **Bayesian model selection**, which is a simple extension of our discussion so far. Recall the general form in which we wrote Bayes' theorem:

$$p(\theta|D, M) = \frac{p(D|\theta, M)p(\theta|M)}{p(D|M)}, \tag{130}$$

where we have emphasised the fact that everything is conditional on some model framework, M. Now we want to concentrate on the normalizing factor, which so far has been relegated to a secondary role or ignored altogether: the **Evidence**.

We saw that the evidence can be written as $P(D|M) = \int P(D|\theta, M)\; p(\theta|M)\, d\theta$; this is the forward probability of getting the data given the model – without regard for the specific parameter value, which is integrated over. It should be clear that we can now write out the simple form of Bayes' Theorem again, just involving D and M:

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}. \tag{131}$$

Now we can do the same thing as when we compared specific hypotheses: take the posterior ratio of the value of $p(M|D)$ for different whole classes of model. This gives a very neat way of answering questions of which class of hypothesis best fits the data; if we have two classes of model that we deem a priori to be equally probable, then the posterior ratio is just the **evidence ratio**:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)}{p(D|M_2)}. \tag{132}$$

But even if the parameter θ is not explicitly on show, it still plays a role through its prior. If either of our models has a parameter, this is integrated over by multiplying the likelihood by the prior, $p(\theta)$, and integrating $d\theta$. Suppose for simplicity that the prior is uniform over some range Δ, which is much broader than the set of values permitted by the likelihood. The evidence then scales $\propto 1/\Delta$, meaning that models with a sufficiently large Δ will always be disfavoured. This behaviour seems puzzling, but what Bayes is doing here is penalising models that are not very specific about their parameter values.

In our detection example, suppose we measure $S/N = 5$: surely a 5σ result is a firm detection? It was good enough for Higgs, after all. But now consider exactly the evidence ratio for the two models M_1: no source; M_2: a source of some flux density up to a maximum of Δ. The evidence ratio is

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{\exp[-n^2/2]}{\int \exp[-(n-S)^2/2]\; dS/\Delta}, \tag{133}$$

where for generality the measured flux is taken to be $n\sigma$. If n is reasonably large, the integral can be taken over an infinite range, so

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{\exp[-n^2/2]}{(2\pi)^{1/2}/\Delta}. \tag{134}$$

So if $n = 5$, we still prefer M_1 (no source) if $\Delta > 107,051$, and this makes sense. If we really believe that typical sources will have fluxes around 50,000, what is the measurement doing coming out as small as 5? Although that is a (very) rare outcome on the assumption of model 1, it is still the best choice. The lesson is that you still need a prior for your parameters, even when they are integrated over. In the case of the Higgs, we fortunately have the opposite situation: the (theoretical) prior for the event rate is close to what is seen, which is also a $> 5\sigma$ deviation with respect to a background-only model. With this prior, the detection criterion could really have been softened, which is why most people of a Bayesian persuasion were convinced of the reality of the Higgs in 2011, even though the official announcement came only on 4 July 2012.

5 More on model fitting and parameter estimation

Bayesian inference as discussed above is a task that is frequently of central interest in science: we often need to do things like fit a model curve through data on a 2-D graph, or surfaces in higher dimensions. The model might have some physical significance, such as fitting the Hubble diagram with the family of curves predicted by Friedmann cosmology for differing values of Ω_V and Ω_m, or to extract many (i.e. 17) cosmological parameters from the pattern of galaxies in a redshift survey, or from the pattern of fluctuations in the Cosmic Microwave Background radiation (see Figure 7). Or we may just want to find the best-fit straight line from a set of data. In either case we also usually want:

1. **To find the allowable ranges of free parameters of the model.** i.e. how do we decide on the range of models that adequately fits the data.
2. **A test of how well the model fits the data.** i.e. having found the best-fitting model, do we believe it? Do the experimental results look like something that might plausibly emerge from our model?

Let's first consider **curve and model fitting**. One of the easiest and most prevalent method is **Least Squares**.

5.1 The method of least-squares fits

We assume we have a set of data-values, D_i, measured as a function of some variable x_i, $D_i = D(x_i)$, and a model that predicts the data values, $M(x, \theta)$, where θ are some unknown (free) parameters. Least-squares minimises the sum

$$S = \sum_i (D_i - M(x_i, \theta))^2, \qquad (135)$$

with respect to the parameters, θ. We may have reason to weight some of the data values higher than others (if for example they have smaller measurement errors) in which case we minimise

$$S = \sum_i w_i (D_i - M(x_i, \theta))^2 \qquad (136)$$

where w_i are a set of arbitrary positive weights. Usually the weights are normalized: $\sum_i w_i = 1$.

If we want to minimise the uncertainty on a model, then the optimum weightings are $w_i = 1/\sigma_i^2$, where σ_i are the errors on the data (we'll show this soon). In this case S is usually known as χ^2 (chi-squared) and χ^2 has its own particular **χ^2-distribution** if the data are independent with errors which have a Gaussian distribution, with variance σ_i^2.

It should be obvious that this is closely related to the Bayesian approach of maximising the posterior, which is just maximum likelihood making the usual assumption of a uniform prior. For Gaussian data, the likelihood is

$$\mathcal{L} = \prod_i p(x_i) = \prod_i (2\pi\sigma_i^2)^{-1/2} \exp\left(-[x_i - M(x_i)]^2/2\sigma_i^2\right) = \exp\left(-\chi^2/2\right) \prod_i (2\pi\sigma_i^2)^{-1/2}, \qquad (137)$$

allowing the errors on different data to be different. But for a given set of σ_i, the ML solution clearly requires us to minimise χ^2.

5.1.1 Estimating a mean from least squares [for independent measurements]

We now consider combining many measurements of a quantity (x_i), each with its own error (σ_i), to estimate the expectation value as accurately as possible – a **minimum variance estimator**. As promised, the weighting of the data will be **inverse variance weighting**. We want to minimise the function

$$S = \sum_i w_i (x_i - \hat{\mu})^2 \qquad (w_i > 0), \qquad (138)$$

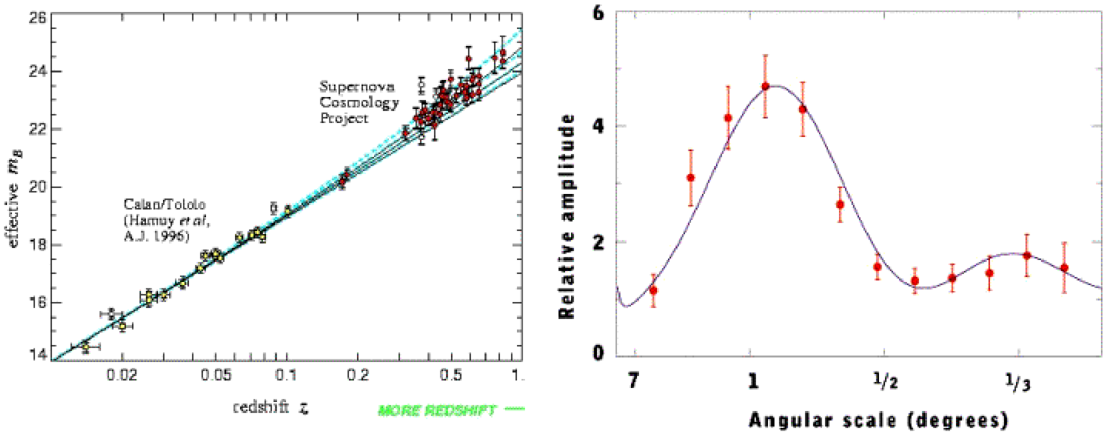

Figure 7: *LHS: Estimation of the acceleration of the Universe from Supernova Type 1a sources. RHS: Fitting the variance of temperature fluctuations in the Cosmic Microwave Background to data from the BOOMERANG balloon experiment. How do we know which curve is the best fit? And how do we know what values the parameters can take?*

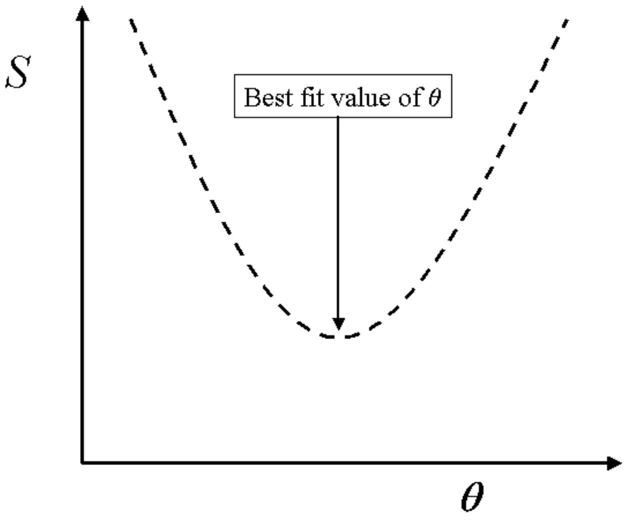

Figure 8: *Least-squares fitting. Usually we have to find the minimum of the function S by searching parameter space.*

with respect to the parameter $\hat{\mu}$, which is our *estimate* of the true expectation value $\mu \equiv \langle x \rangle$. Hence

$$\frac{\partial S}{\partial \hat{\mu}} = 2\sum_i w_i(x_i - \hat{\mu}) = 0, \tag{139}$$

which has the solution in the form of a weighted average of the data:

$$\hat{\mu} = \frac{\sum_i w_i x_i}{\sum_i w_i}. \tag{140}$$

The variance on $\hat{\mu}$ can be found by propagation of errors and with $\partial\hat{\mu}/\partial x_j = w_j/\sum_i w_i$:

$$\begin{aligned} \text{var}(\hat{\mu}) &= \sum_i \left(\frac{\partial \hat{\mu}}{\partial x_i}\right)^2 \sigma_i^2 \\ &= \frac{\sum_i w_i^2 \sigma_i^2}{(\sum_j w_j)^2} \end{aligned} \tag{141}$$

N.B.: the same result can be obtained from the formal definition of the variance: $\text{var}(\hat{\mu}) = \langle(\hat{\mu})^2\rangle - \langle\hat{\mu}\rangle^2$.

We can use Eq. 141 to find the set of weights that will minimise the error on the mean, by minimising $\text{var}(\hat{\mu})$ with respect to w_i (and remembering that $\partial w_i/\partial w_j = \delta_{ij}$):

$$\frac{\partial \text{var}(\hat{\mu})}{\partial w_i} = -\frac{2\sum_k w_k^2\sigma_k^2}{(\sum_j w_j)^3} + \frac{2w_i\sigma_i^2}{(\sum_j w_j)^2} = 0 \tag{142}$$

which implies

$$\sum_k w_k^2 \sigma_k^2 = w_i \sigma_i^2 \sum_j w_j. \tag{143}$$

This last equation is solved when

$$w_i \propto \frac{1}{\sigma_i^2}. \tag{144}$$

This set of weights is the **inverse variance weighting scheme**. With inverse variance weighting, the estimate of the mean is

$$\hat{\mu} = \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}} \tag{145}$$

and the variance on this estimate is

$$\text{var}(\hat{\mu}) = \frac{1}{\sum_i \frac{1}{\sigma_i^2}}. \tag{146}$$

If all the errors are the same, the error on our estimate of the expectation value is simply $\sigma/\sqrt{n}$. The method then reduces to **least-squares fitting**, but note that this is only appropriate if the errors are the same on all the points and all points are independent.

5.1.2 Fitting regression lines

The next step in this kind of problem is to move from one parameter to two, which gives us the important practical application of fitting a straight line. Suppose we have data values in pairs

(x_i, y_i), and our model is $y = ax + b$ – so we want to find the best slope and intercept. Suppose the errors on the y values are known: we then want a maximum-likelihood fit, which means minimising

$$S = \chi^2 = \sum_i (y_i - ax_i - b)^2/\sigma_i^2. \tag{147}$$

Setting the derivatives of S wrt a and b to zero, we get

$$\sum_i x_i(y_i - ax_i - b)/\sigma_i^2 = 0 \tag{148}$$

and

$$\sum_i (y_i - ax_i - b)/\sigma_i^2 = 0. \tag{149}$$

This is a pair of simultaneous equations in a and b; for the simple case of all errors equal, σ can be cancelled out, and the solution is

$$a = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2}; \quad b = \frac{\sum (y_i - ax_i)}{N}. \tag{150}$$

Notice that we have assumed that all the error lies in the y values, with x perfectly known – this is called a **regression of y on x**. If we chose the opposite order, then the new slope is not $a' = 1/a$, because now the errors responsible for the scatter in the plot would be assumed to be on x. In reality, there may be errors in both axes, in which case the question of the best line to fit is more complex.

5.1.3 Multiparameter estimation

We often need to handle problems with many parameters θ_i in which case the process is extended to more dimensions. If the model function $M(x, \theta) = y(x, \theta)$ can be expressed as a power series expansion of x where the free parameters θ_i are the series coefficients, then the minimum point can be found in a single step:

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots = \sum_n \theta_n x_i^n \tag{151}$$

In this case we are fitting a polynomial curve to x and y data values. Minimising S with respect to each θ we find

$$\begin{aligned} \frac{\partial S}{\partial \theta_m} &= \frac{\partial}{\partial \theta_m} \sum_j (y_j - \sum_n \theta_n x_j^n)^2 \\ &= 2\sum_j (y_j - \sum_n \theta_n x_j^n) x_j^m = 0 \end{aligned} \tag{152}$$

where $\partial\theta_n/\partial\theta_m = 0$ if $n \neq m$ and 1 if $n = m$, since the parameters are independent. Equation (152) implies that for polynomial models of the data

$$\sum_j y_j x_j^m = \sum_j \sum_n \theta_n x_j^{n+m} \tag{153}$$

The θ_n values can be found by matrix inversion. If we define $A_m = \sum_j y_j x_j^m$ and $B_{nm} = \sum_j x_j^{n+m}$ then equation (153) can be written

$$A_m = B_{nm}\theta_n \tag{154}$$

where we have assumed summation over repeated indices. This has the solution

$$\theta_n = B_{nm}^{-1} A_m \tag{155}$$

where B^{-1} is the matrix inverse of B.

5.1.4 Estimating parameter uncertainty

So far, we have concentrated on solving parameter-fitting problems of the maximum-likelihood type, and we have seen how, if the log likelihood is at worst quadratic in the parameters then the best-fitting parameters can be found via a matrix inversion. Now we have to consider how accurately these parameters have been determined.

We already solved this problem in the simplest case of determining a weighted mean, where the final answer was a simple function of the data, and we could compute its variance directly. For more complicated cases (even a regression line), this is more painful – and often we deal with a problem that is sufficiently non-linear that the best-fitting point in parameter space has to be found numerically. In such cases, it is better to resort to a simpler approximation. For a single parameter, this says that the variance in the parameter can be computed using the second derivative of the log-likelihood:

$$\mathrm{var}(\theta) \equiv \sigma^2(\theta) = -\left(\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right)^{-1}. \qquad (156)$$

to estimate the error. The idea is that the likelihood function can be approximated about its maximum, at $\theta_{\rm max}$, by a Gaussian:

$$\mathcal{L}(\theta) = \frac{e^{-\Delta\theta^2/(2\sigma^2(\theta))}}{\sqrt{2\pi}\sigma(\theta)}, \qquad (157)$$

where $\Delta\theta = \theta - \theta_{\rm max}$. Equation (156) can be verified by expanding $\ln \mathcal{L}$ to 2nd order in $\Delta\theta$ and differentiating.

This approach can be generalized to multiple parameters by the **covariance matrix**

$$C_{ij} \equiv \langle \Delta\theta_i \Delta\theta_j \rangle = H^{-1}_{ij} \qquad (158)$$

where

$$H_{ij} = -\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \qquad (159)$$

is called the **Hessian Matrix**. This derives from the expression for the N-variable multivariate (zero-mean) Gaussian (which we will not prove here):

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|C|^{1/2}} \exp[-C^{-1}_{ij} x_i x_j/2]. \qquad (160)$$

This can be used to estimate both the **conditional errors**, $\sigma^2_{\rm cond}(\theta) = 1/H_{ii}$ (no sum on i), where we assume the other parameters are known, and the **marginalised errors**, $[H^{-1}]_{ii}$. Where the parameters are correlated (off-diagonal terms in the Hessian), the marginalised errors will often be much larger than the conditional errors.

Often when designing an experiment we want to know in advance how accurate it is liable to be. Errors can still be estimated even before we have any data, because the Hessian can be averaged over ensembles of data (either analytically or, if all else fails, numerically. This yields something called the **Fisher information matrix**:

$$F_{ij} \equiv \langle H_{ij} \rangle = -\left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \right\rangle. \qquad (161)$$

The Fisher Matrix is commonly used in cosmological survey design.

6 Goodness of fit

Once we have found the least squares model, and the error on it, is our task done? A committed Bayesian will say yes: once we have committed to a model (or set of models) and written down our priors, the rest is just implementation. But in the real world we always worry (or we should worry) that there might be a possibility we didn't consider. It is therefore always a good idea to perform a sanity check and ask if the best-fitting model actually does look like the data.

6.1 The chi-squared statistic

For this task, the classical frequentist p-value is still the tool of choice, and the main question is what actual statistic we will choose as a measure of misfit. By virtue of the Central Limit Theorem, Gaussian errors are a reasonable assumption, and so a natural test statistic for n independent data values is χ^2:

$$\chi^2 = \sum_{i=1}^{n} [D_i - M_i(\theta)]^2/\sigma_i^2, \tag{162}$$

where M_i is the model prediction for datum D_i. If the data are correlated, then this generalises to

$$\chi^2 = \sum_{i,j} (x_i - \mu_i) C_{ij}^{-1} (x_j - \mu_j) \tag{163}$$

where μ_i is the expectation value of x_i and C is the *noise covariance matrix*:

$$C_{ij} \equiv \langle n_i n_j \rangle = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle. \tag{164}$$

Where the Gaussian assumption is not valid, it is still possible to make it become so by combining data. A common example is the case of discrete data values drawn from some pdf $p(x)$ in a Poissonian fashion: how can we test whether the set of x values is consistent with a particular hypothetical pdf? The x values themselves can't be used as the data; instead, we have to **bin the data**: define a finite number of ranges of x, count the observed numbers in each of m bins, O_i and compute the expected numbers, E_i. Provided $E_i > 5$ or so, the Gaussian approximation to the Poisson pdf is not so bad, and we know that the variance is E_i. Thus the expression for χ^2 in this case is

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}. \tag{165}$$

By construction, χ^2 is the sum of n Gaussian deviates; by changing variables one can show from this that the distribution of χ^2 is

$$p(\chi^2|n) = \frac{e^{-\chi^2/2}}{\Gamma(n/2)} \left(\frac{\chi^2}{2}\right)^{(n-2)/2}, \tag{166}$$

where the Gamma function generalises factorials to real and imaginary numbers:

$$\Gamma(n+1) = n! \quad ; \qquad \Gamma(n+1) = n\Gamma(n) \qquad ; \quad \Gamma(1/2) = \sqrt{\pi}.$$

Apart from the normalization factor, we are just treating χ as a radius, and writing all of our n-dimensional coordinates as χ times some analogue of polar angles, which are integrated over, leaving

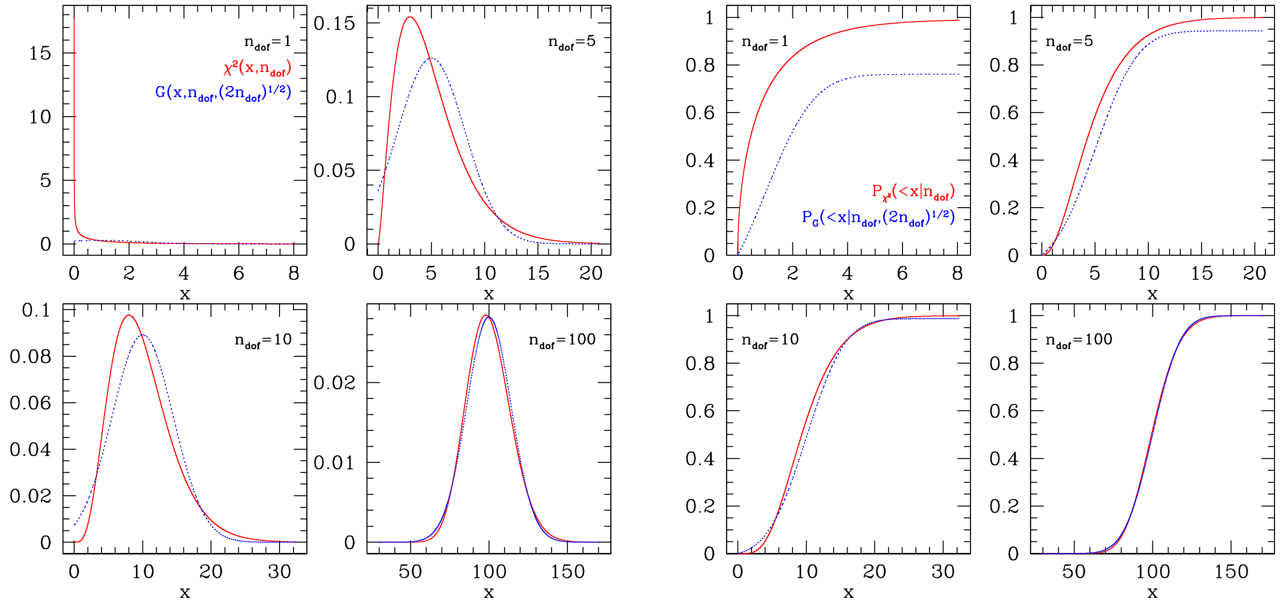

Figure 9: **LHS:** χ^2 distribution (solid line) for four different number of degrees of freedom ($\nu = n_{dof}$ as indicated in each panel) compared to its corresponding Gaussian distribution (dotted line) of mean ν and variance 2ν. **RHS:** The corresponding cumulative χ^2 and Gaussian distributions of the LHS panels.

a volume element $\propto \chi^{n-1}\, d\chi$. The derivation of the normalization in the $n = 2$ case is particularly easy, and is the trick we used to obtain the normalization of the Gaussian.

As usual, in the limit of large n, the distribution of χ^2 itself becomes Gaussian, as illustrated in Figure 9, so we only need the mean and variance. The mean of χ^2 is clearly n, since it is the sum of the squares of a set of Gaussian deviates with zero mean and unit variance: $\chi^2 = \sum_i G_i^2$. For the variance, we need

$$\left\langle \chi^4 \right\rangle = (n^2 - n) \left\langle G_i^2 \right\rangle^2 + n \left\langle G_i^4 \right\rangle = (n^2 - n) + 3n, \tag{167}$$

so that $\mathrm{var}(\chi^2) = n^2 + 2n - \langle \chi^2 \rangle^2 = 2n$.

Thus a useful rule of thumb is that for a good fit

$$\chi^2 \simeq n \pm \sqrt{2n}. \tag{168}$$

Too high a value outside this range can indicate that the model is not a good fit, or that the errors are underestimated. In the first case the model may have to be rejected as there is no point in estimating parameters for a model that does not fit the data. In the second case it may be that the errors are in reality larger than believed. This is very common in astronomy where random measurement errors are easy to determine but where systematic errors often dominate.

Too small a value ($\chi^2 \ll n - \sqrt{2n}$) is as bad as too high a value. Too small a value may indicate hidden correlations between data points, or that the errors are over-estimated (e.g. Figure 10).

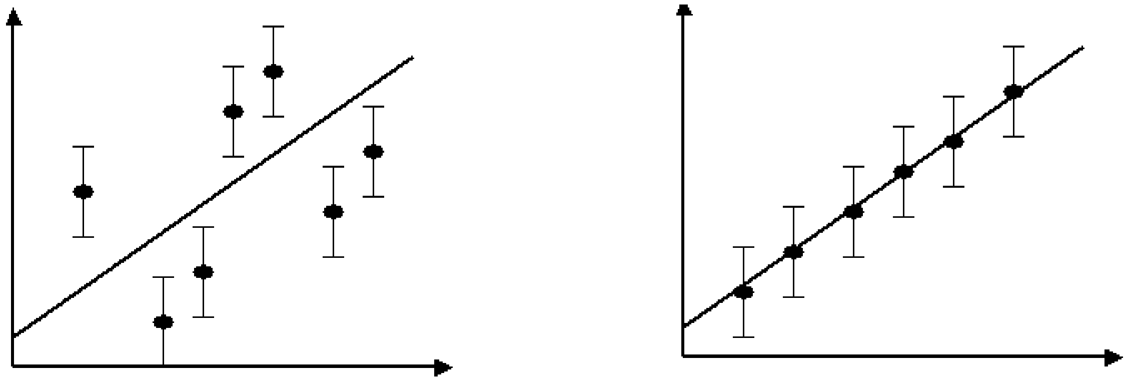

Figure 10: **LHS:** Example of a fit to data where the χ^2 value is too high. **RHS:** Example of a fit to data where the χ^2 value is too low.

6.2 Degrees of freedom

As derived above, n is the number of independent data points (or bins). But we will often want to test a model that has been adjusted so that χ^2 is minimised – e.g. if we subtract the observed mean from the data – in which case typical values of χ^2 will be suppressed below $\chi^2 = n$, and the expected distribution on the null hypothesis is modified. The effect of this can be calculated as follows.

Suppose we have n data values y_i, each of which is equal to some true value, plus a Gaussian noise term with variance σ_i^2. We calculate χ^2 by subtracting the true value (which is known, according to the null hypothesis), so to keep the algebra compact, let's choose units for the data such that these true values are all zero:

$$\chi^2 \equiv \sum_i y_i^2/\sigma_i^2. \tag{169}$$

Now suppose the data have subtracted from them some further 'model template' variation, m_i which is scaled by a parameter a:

$$y_i \to z_i \equiv y_i - am_i; \qquad \chi^2 \to \sum_i z_i^2/\sigma_i^2, \tag{170}$$

so that the value of χ^2 changes if we use the new data. As before, we can ask for the value of a that minimises χ^2, which is

$$a = \frac{\sum_j m_j y_j/\sigma_j^2}{\sum_k m_k^2/\sigma_k^2}. \tag{171}$$

For the simple case where $m_i = 1$ and all the σ's are the same, a would just be the mean value of y, for example. With this definition, the change in χ^2 is

$$\chi^2 = \sum_i z_i^2/\sigma_i^2 = \sum_i (y_i - am_i)^2/\sigma_i^2 = \sum_i y_i^2/\sigma_i^2 - a^2 \sum_i m_i^2/\sigma_i^2. \tag{172}$$

Using this, the initial multidimensional pdf can be rewritten:

$$dp \propto \exp\left[-\left(\sum_i y_i^2/\sigma_i^2\right)/2\right]\, d^n y \propto \exp\left[-\left(\sum_i z_i^2/\sigma_i^2\right)/2\right]\, \exp\left[-a^2\left(\sum_i m_i^2/\sigma_i^2\right)/2\right]\, d^n z. \tag{173}$$

Here, we have rewritten the exponential in terms of z and a, and used the fact that the Jacobian between y space and z space is a constant; the latter fact depends on the best-fitting a being a linear

Table 2: One-tailed critical values of χ^2 for various numbers of degrees of freedom, ν. Numbers in brackets are what would be computed on a Gaussian approximation to the χ^2 distribution. For a given ν, the Gaussian approximation gets worse as p becomes smaller; and although better results are obtained with larger μ, the deviations are substantial even for $\nu = 10$.

ν	$p = 0.05$	$p = 0.01$	$p = 0.001$
1	3.8(3.3)	6.6(4.3)	10.8(5.4)
2	6.0(5.3)	9.2(6.7)	13.8(8.2)
3	7.8(7.0)	11.3(8.7)	16.3(12.8)
5	11.0(10.2)	15.1(12.4)	20.5(14.8)
10	18.3(17.4)	23.2(20.4)	29.6(23.8)

function of the data, which in turn requires the model to depend linearly on a. Finally, since all z_i are also linear combinations of the y_i, we can span the space by ignoring one of the $z's$:

$$(z_1, z_2, z_3 \ldots) \rightarrow (a, z_2, z_3 \ldots). \tag{174}$$

This last set of coordinates are all linear functions of the y_i; thus again the Jacobian between the volume element in this space and $d^n y$ is a constant, and so the pdf becomes

$$dp \propto \exp\left[-\chi^2/2\right]\ \exp\left[-a^2\left(\sum_i m_i^2/\sigma_i^2\right)/2\right]\ d^{n-1}z\, da. \tag{175}$$

Now we can drop a since we want the pdf for χ^2 at given a. As before, each of the z_i will be written as the 'radius' χ times some function of polar angles, so the $(n-1)$-dimensional z space volume element is proportional to $\chi^{n-2}d\chi$ and finally

$$dp \propto \exp\left[-\chi^2/2\right]\ \chi^{n-2}\, d\chi \propto \exp\left[-\chi^2/2\right]\ (\chi^2)^{(n-3)/2}\, d(\chi^2), \tag{176}$$

which is the same as the distribution we started with, but with $n \rightarrow n-1$.

For this reason, we define the **number of degrees of freedom**, ν, to be the number of data points, n_D, minus the number of free parameters, n_θ:

$$\nu \equiv n_D - n_\theta \tag{177}$$

(since it is obvious that subtracting further independent templates from the data will lead in the same way to a further reduction in χ^2). This reduction in the number of degrees of freedom by fitting parameters is something we met earlier in the distinction between the sample and population variances. In summary, then, we test a model by computing χ^2, asking if we have reduced ν by optimizing a model, and then we compute the p-value tail probability $P(\geq \chi^2|\nu)$. Critical values of this at various significance levels are given in Table 2. As ν rises, it quickly becomes a good approximation to assume a Gaussian distribution based on the **reduced chisquared value**:

$$\chi^2_{\text{reduced}} \equiv \chi^2/\nu = 1 \pm \sqrt{2/\nu}. \tag{178}$$

7 Effect of errors on distribution functions: Malmquist and Eddington bias

So far, we have used data for model fitting and testing by deducing simple single statistics such as χ^2 from the data; we have also considered the related problem of estimating simple statistics such as the mean and variance by suitable direct averages over a set of data. But sometimes we can wish to ask more complex questions of the data; in particular, the pdf itself can be an object of interest in its own right, rather than just something that limits the accuracy with which we can measure means etc.

7.1 Luminosity functions and number counts

Any property of a population of astronomical objects can have a distribution function; the most basic cases are the ones related to the size of objects: the **mass function** or the **luminosity function**. Both of these are differential **densities**: $f(M)\,dM$ is the number density of objects with mass in the range dM, and $\phi(L)\;dL$ is a number density of objects with luminosity in the range dL (although we frequently use $\log L$ or magnitude as the independent variable. We can integrate the distribution function to get the total number of objects per unit volume, and divide by this to get a pdf:

$$p(L) = \frac{\phi(L)}{\int_0^\infty \phi(L)\;dL}, \tag{179}$$

but the normalization (total number density) is a quantity of astronomical interest.

The luminosity function is of particular importance because it relates the distribution in apparent luminosity to the distribution in apparent flux density. Suppose for simplicity that we observe a population of objects that are homogeneously distributed with constant density: what is the distribution function of flux density (number of objects per unit solid angle, per unit flux density)? Consider some area of solid angle Ω and limiting flux density S: for each luminosity, the objects can be seen out to a distance D, where

$$S = \frac{L}{4\pi D^2}, \tag{180}$$

so $D = (L/4\pi S)^{1/2}$, and the corresponding volume is $V = \Omega D^3/3 = \Omega(L/4\pi S)^{3/2}/3$. The total number of objects brighter than S is then given by multiplying by the luminosity function and integrating:

$$N = \int V\phi\;dL = \frac{\Omega}{3(4\pi S)^{3/2}} \int L^{3/2}\;dL, \tag{181}$$

which comes out as $N \propto S^{-3/2}$, independent of the form of ϕ. The differential distribution function (or **number counts**) is thus

$$\frac{dN}{dS} \propto S^{-\gamma}; \quad \gamma = 5/2; \tag{182}$$

the so-called Euclidean number counts. In practice, however, the counts of many classes of object can be approximated by a power law of a different slope. A smaller slope (fewer faint objects) arises if we are seeing to a point where the density falls (stars in the Milky Way), or where the flux-distance relation in cosmology becomes non-Euclidean. Conversely, a larger slope implies *more* objects at large distances.

In any case, there are two interesting statistical questions: (i) how do we determine the form of the number counts from data? (ii) what happens if the measured flux densities are subject to error?

7.2 Aside: ML fitting of number counts

Historically, departure of γ from the value 5/2 in the counts of extragalactic radio sources was seen as a strong argument against the steady-state cosmology in the early 1960's (even though the measured slope was actually wrong by quite a large amount).

Given observations of radio sources with flux densities S above a known, fixed measurement limit S_0, what is the best estimate for γ? The model probability distribution for S is

$$p(S)dS = (\gamma - 1)S_0^{\gamma-1}S^{-\gamma}dS \tag{183}$$

where the factors $\gamma - 1$ in front of the term arise from the normalization requirement

$$\int_{S_0}^{\infty} dS\, p(S) = 1. \tag{184}$$

An alternative way to write this model probability distribution for S would be:

$$p(S) = \frac{S^{-\gamma}}{\int_{S_0}^{\infty} S^{-\gamma}dS}. \tag{185}$$

With this pdf, the likelihood function $\mathcal{L}$ for n observed sources is

$$\mathcal{L} = \prod_{i=1}^{n}(\gamma - 1)S_0^{\gamma-1}S_i^{-\gamma} \tag{186}$$

with logarithm

$$\ln \mathcal{L} = \sum_{i=1}^{n}(\ln(\gamma - 1) + (\gamma - 1)\ln S_0 - \gamma \ln S_i). \tag{187}$$

Maximising $\ln \mathcal{L}$ with respect to γ gives

$$\frac{\partial}{\partial\gamma}\ln \mathcal{L} = \sum_{i=1}^{n}\left(\frac{1}{\gamma - 1} + \ln S_0 - \ln S_i\right) = 0, \tag{188}$$

which is minimized when

$$\gamma = 1 + \frac{n}{\sum_{i=1}^{n}\ln\frac{S_i}{S_0}}. \tag{189}$$

Suppose we only observe one source with flux twice the cut-off, $S_1 = 2S_0$, then

$$\gamma = 1 + \frac{1}{\ln 2} = 2.44 \tag{190}$$

but with a large uncertainty. Clearly, as $S_i = S_0$ we find $\gamma \to \infty$ as expected. In fact $\gamma = 2.8$ for bright radio sources at low frequencies, significantly steeper than 1.5.

7.2.1 The apparent distribution function

The problem with dealing with number counts in this way is that no account is taken of the fact that the flux densities are inexact and afflicted by measuring errors; what affect does this have on the observed number counts?

Qualitatively we can see what happens as follows. At any given flux density we count a number of sources per unit flux density, but that number contains sources that shouldn't have been counted had we measured with infinite precision, because some fainter sources will be measured as being slightly brighter than they really are. But the same process also scatters objects we should have counted out of our flux density range. Because there are more faint sources than there are bright sources, there can be a tendency to increase the numbers counted over the true value. There is also a separate, but closely related bias: because there are more fainter sources than bright ones, the objects we find at a given measured flux density will tend to have their flux densities overestimated. These two effects are called **Eddington bias** (wrong number of sources at a given measured flux density and **Malmquist bias** (wrong average flux density for a set of sources at a given measured flux density). The simplest case where it is obvious how both effects arise is if the true counts have a cutoff at some maximum flux: any object found beyond that cutoff must have had its measuring error scatter it brightwards, and clearly the counts are overestimated in this regime.

These related effects can be analysed quite simply when we realise that the effect of measuring errors is a convolution: a spike at some true flux density gets spread out with some error distribution. To keep things general, call the observable x, and let $f(x)$ be the distribution function for numbers of objects in dx. Let $p(\epsilon)$ be the pdf for getting an error ϵ in x, so that the true value of x is $x_{\rm obs} - \epsilon$. The distribution function in observed x is then

$$g(x_{\rm obs}) = \int f(x_{\rm obs} - \epsilon)\, p(\epsilon)\, d\epsilon. \tag{191}$$

If we also want the bias in $x_{\rm obs}$, we just have to average the error ϵ, weighting by the numbers at each ϵ:

$$\langle \epsilon \rangle = \frac{1}{g(x_{\rm obs})} \int \epsilon\, f(x_{\rm obs} - \epsilon)\, p(\epsilon)\, d\epsilon. \tag{192}$$

These general expressions can be made more intuitive by a Taylor expansion of $f(x_{\rm obs} - \epsilon)$, and recognising that the average of ϵ over the pdf is zero by definition. Thus

$$g(x_{\rm obs}) \simeq \int \left[f(x_{\rm obs}) - f'(x_{\rm obs})\epsilon + f''(x_{\rm obs})\epsilon^2/2 \right] p(\epsilon)\, d\epsilon = f(x_{\rm obs}) + f''(x_{\rm obs})\sigma^2/2, \tag{193}$$

where σ is the rms error; so Eddington bias depends mainly on the curvature of the distribution function. Similarly, to lowest order, the Malmquist bias is

$$\langle \epsilon \rangle \simeq \frac{1}{f(x_{\rm obs})} \int \epsilon \left[f(x_{\rm obs}) - f'(x_{\rm obs})\epsilon \right] p(\epsilon)\, d\epsilon = -[f'(x_{\rm obs})/f(x_{\rm obs})]\, \sigma^2. \tag{194}$$

So Malmquist bias depends mainly on the first derivative of the distribution function.

In astronomy, it is more normal to deal with logarithmic quantities, such as a constant error in apparent magnitude (i.e. a flux error that increases with flux). In the notation of our previous example, if $dN/dS \propto S^{-\gamma}$, then $dN/d\ln S \propto S^{-(\gamma-1)} \propto \exp[-(\gamma-1)x]$, where x is $\ln S$. The fractional Eddington bias is then

$$1 + (f''/f)\sigma^2/2 = 1 + (\gamma - 1)^2\sigma^2/2, \tag{195}$$

and the Malmquist bias is

$$\Delta \ln S = (\gamma - 1)\sigma^2. \tag{196}$$

Thus, for a power law distribution function, the effect of observational errors is always to increase the observed number of objects at a given flux; and these objects will have their fluxes overestimated provided $\gamma > 1$. Finally, if we want to translate these expressions into actual magnitudes, we need to remember $m = -2.5 \log_{10} S = -(2.5/\ln 10) \ln S$. Thus, the Malmquist bias is

$$\Delta m = -(\ln 10/2.5)(\gamma - 1)\sigma_m^2. \tag{197}$$

The term Malmquist bias tends to be used quite broadly, and especially in conjunction with the fact that more luminous objects are seen to greater distances. So the mean luminosity of a flux limited sample will be higher than the mean luminosity just averaging over the luminosity function. This case is similar mathematically, but obviously is distinct in that it doesn't involve measuring errors. Malmquist's original analysis actually involved the mean luminosity of stars of a given colour class. Colour correlates with luminosity, but imperfectly, and so the more luminous members of the class are over-represented if there is a flux limit. So here there is a scatter – but it is a 'cosmic scatter', rather than as a result of our measurements being imperfect.