# ASTRONOMICAL STATISTICS

**Andy Taylor**
Room C19, Royal Observatory; ant@roe.ac.uk

December 17, 2004



## Contents

**Suggested references:**

Textbooks on probability and statistics are notoriously: (a) boring (b) irrelevant to astronomers and physicists (c) from one viewpoint and ignore others (d) all of the above. Given this the main text for this course is these notes. However if you're after more, two text that do much better than the norm are:

- Robert Lupton, *Statistics in Theory and Practice*, Princeton University Press; ISBN: 0691074291, Price: 29.95 (from Amazon). This is a great 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 Example Book: The Art of Scientific Computing*, Cambridge University Press; ISBN: 0521437210, Price: 22.95 (from Amazon). This covers much more than statistics, and provides computer code for all its methods. There is a C and C$^{++}$ version if you're not into Fortran. Apart from actual code, its strength is in clear explanations of scientific methods, in particular Statistical Methods.

# PART ONE

# 1 PROBABILITY AND STATISTICS

## 1.1 Introduction

Science, and in particular Modern Astronomy and Astrophysics, is impossible without knowledge of probability and statistics. To see this let's 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 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, i.e., 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 we don't know if it will fail a critical test tomorrow.

In all of this probability and statistics are vital for:

- **Detection of signals:**
  How do we know if we've found a new object or detected a new signal?
  e.g. detection of astronomical objects, detection of spectral features, detection of fluctuations in temperature or polarisation in the microwave background, detection of galaxy clustering.

- **Detection of correlations:**
  A correlation between two quantities may imply a physical connection. But how do we detect any such correlation? And when should we believe its real?
  e.g. Hubble diagram, the Hertsprung-Russell diagram, galaxy colour-magnitude relations.

- **Tests of hypotheses:**
  How do we rule out a false theory? Or pass another ?
  e.g. isotropy of the Universe, physical association between astronomical objects, the existence and nature of dark matter and dark energy in the Universe.

- **Model-fitting to data and interpretation:**
  How should we compare our models with data?
  e.g. compare a cosmological theory with the cosmic microwave background or galaxy redshift survey, or compare a stellar model with the observed solar neutrino flux.

- **Estimation of parameters:**
  If a model does fit the data, how do we extract estimates of the model parameters that best fit the data?

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

e.g. the mean mass density of the Universe, the temperature, surface gravity, metallicity of stars, number and mass of neutrinos.

- **As a theoretical tool:**
  How do we model or predict the properties of populations of objects, or study complex systems?
  e.g. Statistical properties of galaxy populations, or stars, simulations of stellar clusters or large-scale structure, turbulence in the IGM.

In astronomy the need for statistical methods is especially acute, as we cannot directly interact with the objects we observe. For instance, 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 archeology 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 tricky one. Surprisingly there isn't a universally agreed definition. The answer to the question "What is a probability?" depends on how you assign probabilities in the first place. To understand the origins of this problem let's take a quick look at the historical development of probability:

**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).

**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:** Edinburgh, 1763, and 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 rejected 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 two basic philosophies 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. In some 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.

- **Bayesian probabilities:** Probability is a "degree-of-belief" of the outcome of an event, allocated by an observer given the available evidence. While this is more intuitive it can lead to problems with subjectivity. But if all observers have the same evidence and allocate the same probability, Bayesians would argue it is objective.

These do not obviously give the same answer to problems in probability and statistic, as they assign different probabilities to some events. However, in many instances they are the same, and both agree on the basics of probability theory (see next section). In this course we shall mostly work in the Frequentist framework, since the majority of probability and statistics has been formulated this way. This should provide a "toolbox" of practical methods to solve everyday problems in physics and astronomy. However, we shall encounter situations where we are forced to consider the Bayesian approach, and its deep implications.

## 1.3 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. 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 even, or **random variable**, $x$, occurs.

1. **The Range of Probabilities:** The probability of an event is measurable on a continuous scale, such that $P(x)$ 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 $p(x)$ is the **probability density**. The probability density clearly must have units of $1/x$.

3. **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[2]

$$P(x_1 + x_2 + \cdots + x_n) = P(x_1) + P(x_2) + \cdots + P(x_n) \tag{3}$$

4. **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:

$$\begin{aligned} P(x,y) &= P(x|y)P(y) \\ &= P(y|x)P(x) \end{aligned} \tag{4}$$

   where we have introduced the **conditional probability**, $P(x|y)$, which denotes the probability the of $x$ given the event $y$ has occurred. If $x$ and $y$ are independent events then

$$P(x,y) = P(x)P(y). \tag{5}$$

**Problem 1.1**
Three coins are tossed. What is the probability that they fall either all heads or all tails? (Assume $P(\text{heads}) = P(\text{tail}) = 1/2$). Try some of these suggested answers:

1. There are 8 possible equally probable combinations (HHH, HHT, HTH,.....). Two of these give us either all heads or all tails, so the probability is 1/4.

[2]**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.

2. There are 2 possibilities: either all 3 coins fall alike, or 2 fall alike and 1 is different. Hence the probability is 1/2.

3. There are 4 possibilities: 3 heads, 2 heads and 1 tail, 2 tails and 1 head, 3 tails. Hence the probability is $2/4 = 1/2$.

4. Of the 3 coins, 2 must fall alike. The other must either be the same as these or different, so probability is 1/2. Which of these arguments are wrong, and why?

## 1.4 Moments of the distribution



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

**Definition:**

$$m_n \equiv \langle x^n \rangle = \int_{-\infty}^{\infty} dx\, x^n p(x), \tag{6}$$

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

$$m_0 = \int_{-\infty}^{\infty} dx\, p(x) = 1 \tag{7}$$

(Axiom 2, The Sum Rule).

The first moment,

$$m_1 = \langle x \rangle, \tag{8}$$

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;

$$\mu_n \equiv \langle (x - \langle x \rangle)^n \rangle. \tag{9}$$

The second centred moment,

$$\mu_2 = \langle (x - \langle x \rangle)^2 \rangle, \tag{10}$$

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

$$\sigma^2 \equiv \mu_2. \tag{11}$$

We will need the following, useful result later:

$$\sigma^2 = \langle (x - \langle x \rangle)^2 \rangle = \langle (x^2 - 2x\langle x \rangle + \langle x \rangle^2) \rangle = \langle x^2 \rangle - \langle x \rangle^2. \tag{12}$$

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**; $\sigma$. This quantity is sometimes also called the **root mean squared (rms) deviation**[3], or **error**[4].

Higher moments 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.

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

$$\mathrm{Cov}(x, y) = \langle xy \rangle = \int_{-\infty}^{\infty} dx\, dy\, xy\, p(x, y) \tag{13}$$

and a dimensionless **correlation coefficient**;

$$r = \frac{\langle xy \rangle}{\sqrt{\langle x^2 \rangle \langle y^2 \rangle}}, \tag{14}$$

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.

## 1.5 Discrete probability distributions

We can use the calculus of probabilities to produce a mathematical expression for the probability of a wide range of multiple events such as discussed in Example 1.

### 1.5.1 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, $p_1$, of have some attribute (such as a coin being heads-up) and a probability, $p_2 = 1 - p_1$ 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 out comes are hh, ht, th, and tt. As these are independent events we see that the probability of each distinguishable outcome is

$$\begin{aligned} P(hh) &= P(h)P(h), \\ P(ht + th) &= P(ht) + P(th) = 2P(h)P(t), \\ P(tt) &= P(t)P(t). \end{aligned} \tag{15}$$

[3]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.

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

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, then the number of possible permutations which can result in $n$ of them having some attribute is the $n^{th}$ coefficient in the expansion of $(p_1 + p_2)^N$, the probability of each of these permutations is $p_1^n p_2^{N-n}$, and the probability of n objects having some attribute is the binomial expansion

$$P_n = C_n^N p_1^n p_2^{N-n}, \quad (0 \leq n \leq N), \tag{16}$$

where

$$C_n^N = \frac{N!}{n!(N-n)!} \tag{17}$$

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

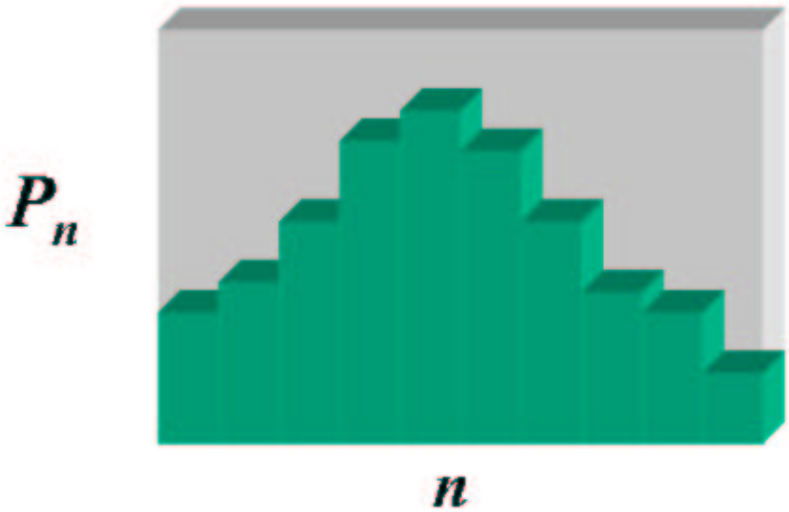


Figure 1: *Histogram of a binomial distribution*

doing so have generated the predicted probability distribution which in this case is the binomial distribution whose form is determined by $N$, $n$, $p_1$. 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

$$\begin{aligned}
\langle n \rangle &= \sum_{n=0}^{N} nP \\
&= \sum_{n=0}^{N} n \frac{N!}{n!(N-n)!} p_1^n p_2^{N-n} \\
&= \sum_{n=1}^{N} \frac{N!}{(n-1)!(N-n)!} p_1^n p_2^{N-n} \\
&= \sum_{n=1}^{N} \frac{N(N-1)!}{(n-1)!(N-n)!} p_1 p_1^{n-1} p_2^{N-n} \\
&= N p_1
\end{aligned} \tag{18}$$

For $p_1 \neq p_2$ the distribution is asymmetric, with mean $\langle n \rangle = Np_1$, but if 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 later!

### 1.5.2 The Poisson distribution

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:

The probability of n "successes" of an event of probability p is

$$P_n = C_n^N p^n (1-p)^{N-n} \tag{19}$$

after N trials.

Let us suppose that the probability p is very small, but that in our experiment we allow N to become large, while 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 $p \to 0$, $N \to \infty$, while $\lambda$ =constant. Then,

$$P_n = \frac{N!}{n!(N-n)!} \left(\frac{\lambda}{N}\right)^n \left(1 - \frac{\lambda}{N}\right)^{N-n} \tag{20}$$

Using **Stirling's approximation**, where $x! \to \sqrt{2\pi} e^{-x} x^{x+1/2}$ when $x \to \infty$, and letting $N \to \infty$, we find[5]

$$\begin{aligned} \frac{N!}{(N-n)!} &= \frac{\sqrt{2\pi} e^{-N} N^{N+1/2}}{\sqrt{2\pi} e^{-(N-n)} (N-n)^{N-n+1/2}} \\ &= \frac{e^{-n} N^{N+1/2}}{N^{N-n+1/2} (1-n/N)^{N-n+1/2}} \\ &= e^{-n} N^n e^n \\ &= N^n \end{aligned} \tag{22}$$

and

$$\left(1 - \frac{\lambda}{N}\right)^{N-n} = e^{-\lambda}. \tag{23}$$

[5]**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:

$$N \ln(1 - x/N) \to N(-x/N) = -x \tag{21}$$

Combining these results with equation (20) we find

$$P_n = \frac{\lambda^n e^{-\lambda}}{n!} \tag{24}$$

This is **Poisson's distribution** for random point processes, discovered by him in 1837.

### 1.5.3 Moments of the Poisson distribution:



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

$$m_i = \langle n^i \rangle = \sum_{n=0}^{\infty} n^i P_n \tag{25}$$

The mean of Poisson's distribution is

$$\begin{aligned} \langle n \rangle &= \sum_{n=0}^{\infty} n \frac{\lambda^n e^{-\lambda}}{n!} \\ &= \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{(n-1)!} \\ &= \sum_{n=1}^{\infty} \frac{\lambda \lambda^{n-1} e^{-\lambda}}{(n-1)!} \\ &= \lambda \end{aligned} \tag{26}$$

i.e. the expectation value of Poisson's distribution is the factor $\lambda$. This makes sense since $\lambda$ was defined as the mean of the underlying Binomial distribution, and kept constant when we took the limit. Now lets look at the second centred moment (i.e. the variance):

$$\begin{aligned} \mu_2 &= \sum_{n=0}^{\infty} (n - \langle n \rangle)^2 \frac{\lambda^n e^{-\lambda}}{n!} \\ &= \sum_{n=0}^{\infty} n^2 \frac{\lambda^n e^{-\lambda}}{n!} - \lambda^2 \\ &= \sum_{n=1}^{\infty} \frac{n \lambda^n e^{-\lambda}}{(n-1)!} - \lambda^2 \\ &= \sum_{n=1}^{\infty} \frac{n \lambda \lambda^{n-1} e^{-\lambda}}{(n-1)!} - \lambda^2 \\ &= \sum_{n=0}^{\infty} \frac{(n+1) \lambda \lambda^n e^{-\lambda}}{n!} - \lambda^2 \\ &= (\lambda + 1)\lambda - \lambda^2 \\ &= \lambda. \end{aligned} \tag{27}$$

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

### 1.5.4 A rule of thumb

Lets 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

$$n \pm \sqrt{n} \tag{28}$$

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 we are assuming the underlying distribution is Gaussian (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 $\sigma = \sqrt{n}$.



### 1.5.5 Detection of a Source

A star produces a large number, $N \gg 1$, of photons during its life. If we observe it with a telescope on Earth we can 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

$$\begin{aligned} \lambda &= Np \\ N &\to \infty \\ p &\to 0. \end{aligned} \tag{29}$$

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.

Conversely, if we make one observation and detect $n$ photons, we can use the Poisson distribution to derive the probability of all the possible values of $\lambda$: we can set confidence limits on the value of $\lambda$ from this observation. And if we can show that one piece of sky has only a small probability of having a value of $\lambda$ as low as the surrounding sky, then we can say that we have detected a star, quasar, galaxy or whatever at a particular **significance level** (i.e. at a given probability that we have made a mistake due to the random fluctuations in the arrival rate of photons). A useful rule of thumb here is

$$\lambda_S \geq \lambda_B + \nu\sqrt{\lambda_B} \tag{30}$$

where $\lambda_S$ is the mean counts from the source, $\lambda_B$, is the mean background count, and $\nu$ is the detection level. Usually we take $\nu = 3$ to be a detection, but sometimes it can be $\nu = 5$ or even $\nu = 15$ for high confidence in a detection.

### 1.5.6 Testing the Isotropy of the Universe

A key observation in cosmology is that the universe appears isotropic on large scales. Unless our galaxy occupies an extremely special place in the universe, then isotropy

implies homogeneity. Without this information a wide range of different cosmological models become possible, and homogeneity pins down the type of universe we inhabit and the manner in which it evolves.

But how do we test for isotropy? Let us suppose that we count a large number of quasars in many different parts of the sky to obtain an average surface density of quasars, per square degree. Dr. Anne Isotropy[6] is investigating quasar counts in one area of sky and finds an unusually large number of quasars, $n$ in her area of one square degree. Does this indicate anisotropy or is it a random fluctuation?

The probability of finding exactly $n$ quasars is

$$P_n = \frac{\lambda^n e^{-\lambda}}{n!}. \tag{31}$$

But being an astronomer of integrity, she admits that she would have been equally excited if she had found any number greater than the observed number $n$. So the probability of obtaining $N$ quasars or greater is

$$P(\geq N) = \sum_{n=N}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!}. \tag{32}$$

If $\lambda = 2$ and $N = 5$, $P_N = 0.036$ but $P(N \geq 5) \approx 0.053$; i.e. there is only a 5% chance of such a fluctuation occurring at random.

But nobody believes her result is significant, because it turns out that she had searched through 20 such fields before finding her result, and we should expect one out of those 20 to show such a fluctuation. The more she looks for anisotropy, the more stringent her criteria have to be!

$P(\geq N)$ has a particularly simple form if $N = 1$:

$$\begin{aligned} P(\geq 1) &= \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \\ &= 1 - e^{-\lambda}. \end{aligned} \tag{33}$$

Hence for $\lambda \ll 1$, $P(n \geq 1) \rightarrow \lambda$. The function $P(\geq N)$ is called the **cumulative probability distribution**.

### 1.5.7 Identifying Sources

Suppose we are trying to find the optical counterpart to a known X-ray source, whose position is known to some finite accuracy: we say that we know the size of the X-ray error box (usually this will be a circle!). We attempt to do this by looking within that error box on an optical image of the same region of sky. We know from counting the number of stars etc. on that image the number of chance coincidences that we expect within the error box, $\lambda$, and we can use this information to decide whether or not the existence of an optical object close to the X-ray source is evidence that the two are related. If $\lambda = 1$

[6] Sorry...

then the probability of finding one or more chance coincidences is 63 percent: this tells us that if we want to rely on this information alone we cannot expect our identification to be very reliable in this example.

## 1.6 Continuous probability distributions

So far we have dealt with discrete probability distributions (e.g. number of heads, number of photons, number of quasars). But in other problems we wish to know the probability distribution of a quantity which is continuously variable (e.g. the angle $\theta$ between two directions can have any real value between 0 and 360). We can define a function $\theta$ such that the probability that the quantity $\theta$ has values between $\theta - d\theta/2$ and $\theta + d\theta/2$ is $P(\theta)|d\theta|$, and the probability that has any value larger than is

$$p(\theta \geq \theta_1) = \int_{\theta_1}^{2\pi} d\theta \, p(\theta) \tag{34}$$

The quantity $\theta$ is known as a **random variable** and is used to denote any possible values which a quantity can have. Here $p(\theta)$ called the **probability density**. If its argument has dimensions, the probability density will be measured in units of inverse dimensions. $p(\theta \geq \theta_1)$ is again called the cumulative probability distribution, and is dimensionless.

### 1.6.1 The Gaussian Distribution

A limiting form of the Poisson distribution (and most others - see the **Central Limit Theorem** below) is the **Gaussian distribution**. Let's write the Poisson distribution as

$$P_n = \frac{\lambda^n e^{-\lambda}}{n!} \tag{35}$$

Now let $x = n = \lambda(1+\delta)$ where $\lambda \gg 1$ and $\delta \ll 1$. Using Stirlings formula for $n!$ we find[7]

$$\begin{aligned} p(x) &= \frac{\lambda^{\lambda(1+\delta)} e^{-\lambda}}{\sqrt{2\pi} e^{-\lambda(1+\delta)} [\lambda(1+\delta)]^{\lambda(1+\delta)+1/2}} \\ &= \frac{e^{\lambda\delta}(1+\delta)^{-\lambda(1+\delta)}}{\sqrt{2\pi\lambda}} \\ &= \frac{e^{-\lambda\delta^2/2}}{\sqrt{2\pi\lambda}} \end{aligned} \tag{37}$$

Substituting back for $x$ yields

$$p(x) = \frac{e^{-(x-\lambda)^2/(2\lambda)}}{\sqrt{2\pi\lambda}} \tag{38}$$

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

$$-\lambda(1+\delta)\ln(1+\delta) = -\lambda(1+\delta)(\delta - \delta^2/2) = -\lambda\delta(1+\delta/2). \tag{36}$$

This is a **Gaussian**, or Normal[8], distribution with mean and variance of $\lambda$. The Gaussian distribution is the most important distribution in probability, due to its role in the Central Limit Theorem. It is also one of the simplest.

### 1.6.2 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,

$$p(x)dx = g(y)dy. \tag{39}$$

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

$$p(x) = g(y(x)) \left| \frac{dy}{dx} \right|. \tag{40}$$

So the transformation of probabilities is just the same as the transformation of normal functions in calculus. 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.

### 1.6.3 An example: projection of a random pencil

Drop a pencil of length $L$ with random orientation onto a table. What is the probability distribution of the apparent length $x$ if we sight along the top of the table?

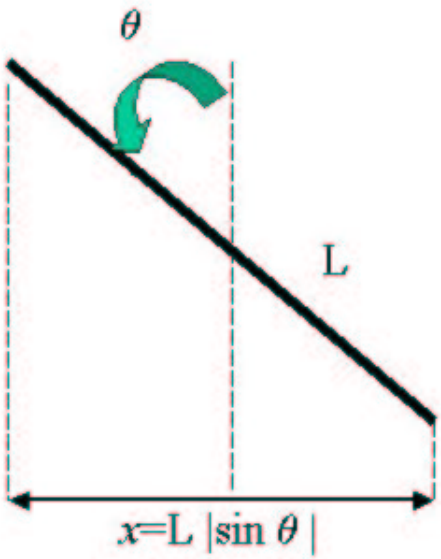


Figure 2: *A randomly orientated pencil.*

[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.

For random orientations $p(\theta) = 1/2\pi$ (uniform probability in $\theta$).

So

$$\begin{aligned} p(x) &= p(\theta)\left|\frac{d\theta}{dx}\right| \\ &= \frac{1}{2\pi}\left|\frac{1}{L\cos\theta(x)}\right| \qquad (-\pi \le \theta < \pi) \\ &= \frac{1}{2\pi}\frac{4}{L\cos\theta(x)} \qquad (0 \le \theta < \pi/2) \end{aligned} \tag{41}$$

since there are four angles, $\pm\theta$ and $\pm(\pi - \theta)$, which give the same value of $x$.

Finally, writing $\cos\theta$ as a function of $x$;

$$\cos\theta(x) = \sqrt{1 - \frac{x^2}{L^2}} \qquad (0 \le \theta < \pi/2) \tag{42}$$

we find

$$p(x) = \frac{2}{\pi\sqrt{L^2 - x^2}}. \tag{43}$$

Note:

1. Take care with limits and multiple values,
2. Normalisation, $\int dx f(x) = 1$.

### 1.6.4 The predicted 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 Sir Martin Rees in 1966 (when he was an undergrad), and first observed in 1971.

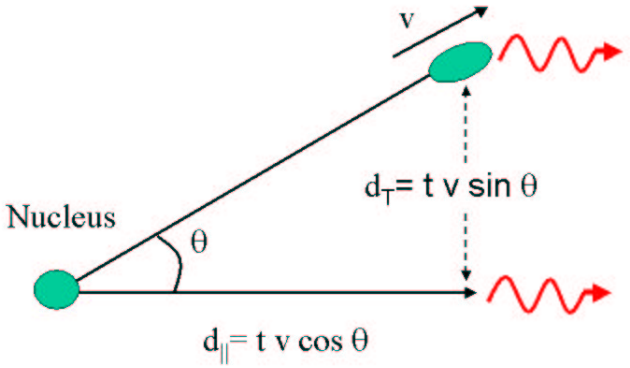


Figure 3: *Superluminal motion from a quasar nucleus.*

Suppose the angle to the line of sight is $\theta$ as shown above, and that a component is ejected along the jet from the nucleus. After some time $t$ the ejection component has

travelled a distance $d_{||} = tv\cos\theta$ along the line of sight. But light emitted from the nucleus is delayed by the extra travel time $\Delta t = t - d_{||}/c = t(1-(v/c)\cos\theta)$. 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

$$v' = \frac{d_\perp}{\Delta t} = \frac{v\sin\theta}{1-(v/c)\cos\theta}. \tag{44}$$

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

$$\beta = v/c, \qquad\qquad \gamma = (1-\beta^2)^{-1/2} \tag{45}$$

we find that the apparent transverse speed $\beta'$ has a maximum value when

$$\frac{\partial\beta'}{\partial\theta} = \frac{\beta(\beta-\cos\theta)}{(1-\beta\cos\theta)^2} = 0, \tag{46}$$

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

**Given a randomly oriented sample of radio sources, what is the expected distribution of $\beta'$ if $\beta$ is fixed?**

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

$$p(\theta) = \sin\theta \qquad\qquad 0 \le \theta \le \pi/2. \tag{47}$$

Hence,

$$\begin{aligned} p(\beta') &= p(\theta)\left|\frac{d\theta}{d\beta}\right| \\ &= \frac{\sin\theta(1-\beta\cos\theta)^2}{|\beta\cos\theta - \beta^2|} \end{aligned} \tag{48}$$

where $\sin\theta$ and $\cos\theta$ are given by the equation for $\beta'$. 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 $\beta'$ are $0 \le \beta' \le \gamma\beta$. The expression for $p(\beta')$ in terms of $\beta'$ alone is rather messy, but simplifies for $\beta \to 1$:

$$\beta' = \frac{\sin\theta}{(1-\cos\theta)} \tag{49}$$

$$p(\beta') = \sin\theta(1-\cos\theta). \tag{50}$$

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 (50) finally gives us

$$p(\beta') = \frac{4\beta'}{(1+\beta'^2)^2} \qquad\qquad \beta' \ge 1. \tag{51}$$

The cumulative probability for $\beta'$ is

$$P(>\beta') = \frac{2}{(1+\beta'^2)} \qquad\qquad \beta' \geq 1. \tag{52}$$

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.



## 1.7 The addition of random variables

### 1.7.1 Multivariate distributions and marginalisation

We can define the joint probability distribution function for two random variables $x$ and $y$ as $p(x, y)$ which is an extension of the single-variable case. Joint distributions for many random variables are known as **multivariate distributions**.

Univariate distributions for each of $x$ or $y$ can be obtained by integrating, or **marginalising** $p(x, y)$ with respect to the other variable:

$$\begin{aligned} p(x) &= \int dy\, p(x, y) \\ p(y) &= \int dx\, p(x, y). \end{aligned} \tag{53}$$

Such a distribution is called the **marginal distribution** of $x$.

The variables $x$ and $y$ are independent if their joint distribution function can be factorized,

$$p(x, y) = p(x)p(y), \tag{54}$$

for all values of $x$ and $y$ (see Section 3, axiom 4).

### 1.7.2 The probability distribution of summed 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, $p(z)$ of $z$? The probability of observing a value, $z$, which is greater than some value $z_1$ is

$$p(z \geq z_1) = \int_{z_1}^{\infty} dz\, p(z) \tag{55}$$

$$= \int_{-\infty}^{\infty} dy \int_{z_1 - y}^{\infty} dx\, p(x, y), \tag{56}$$

where the integral limits on the second line can be seen from defining the region in the $x - y$ plane (see Figure 4).

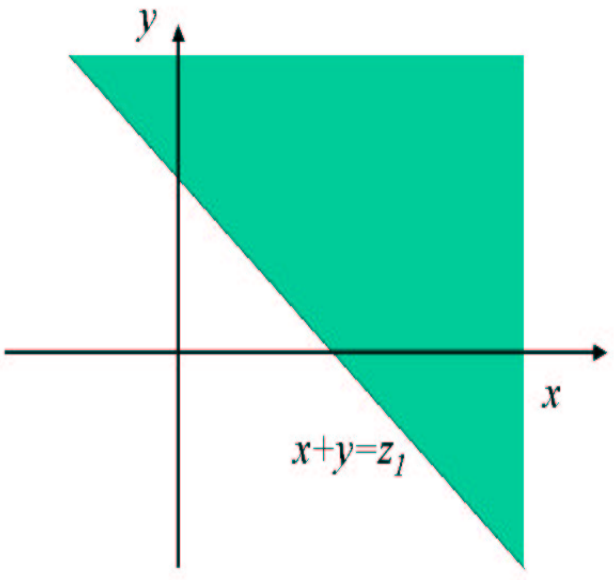


Figure 4: *The region of integration of equation (56).*

If we now change variables back to $z$, so $x = z - y$ we find

$$p(z \geq z_1) = \int_{-\infty}^{\infty} dy \int_{z_1}^{\infty} dz\, p(z - y, y). \tag{57}$$

Now, comparing with equation (55) we can write $p(z)$ as

$$p(z) = \int_{-\infty}^{\infty} dy\, p(z - y, y) \tag{58}$$

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

$$p(z) = \int_{-\infty}^{\infty} dy\, p(z - y)p(y) \tag{59}$$

**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** (F.T.'s), since the F.T. of a convolved function $p(z)$ is simply the product of the F.T.s of the separate functions $p(x)$ and $p(y)$ being convolved.

### 1.7.3 Error propagation

If $\sigma_x$ and $\sigma_y$ are the uncertainties on the independent random variables $x$ and $y$, what is the uncertainty on $z$ ? Let's assume $x$ and $y$ have zero mean. Then

$$\sigma_z^2 \;=\; \langle z^2 \rangle$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} dzdy\, z^2 p(z-y)p(y) \\
&= \int_{-\infty}^{\infty} dxdy\, (x+y)^2 p(x)p(y) \\
&= \int_{-\infty}^{\infty} dxdy\, (x^2+2xy+y^2) p(x)p(y) \\
&= \langle x^2 \rangle + \langle y^2 \rangle \\
&= \sigma_x^2 + \sigma_y^2
\end{aligned} \tag{60}
$$

Hence the variance of a sum of independent random variables is equal to the sum of their variances. This result is independent of the underlying distribution functions.

In general if $z = f(x, y)$ we can propagate errors by expanding $f(x, y)$ around some arbitrary values, $x_0$ and $y_0$;

$$f(x, y) = f(x_0, y_0) + x\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial y}. \tag{61}$$

The mean is

$$\langle z \rangle = f(x_0, y_0) \tag{62}$$

and the variance is;

$$
\begin{aligned}
\sigma_z^2 &= \langle z^2 \rangle - \langle z \rangle^2 \\
&= \int_{-\infty}^{\infty} dxdy\, (f - \langle f \rangle)^2 p(x)p(y) \\
&= \int_{-\infty}^{\infty} dxdy\, (x^2 f_x^2 + y^2 f_y^2 + 2xy f_x f_y) p(x)p(y)
\end{aligned} \tag{63}
$$

where we have used the notation $f_x \equiv \partial f/\partial x$.

Averaging over the random variables we find

$$\sigma_z^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2. \tag{64}$$

This formula will allow us to propagate errors for arbitrary function. 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.

## 1.8 Characteristic functions

Let's return to the problem of dealing with convolutions of functions. As we noted above, the easiest way to handle convolutions is with Fourier Transforms. In probability theory the Fourier Transform of a probability distribution function is known as the **characteristic function**:

$$\phi(k) = \int_{-\infty}^{\infty} dx\, p(x) e^{ikx} \tag{65}$$

with reciprocal relation

$$p(x) = \int_{-\infty}^{\infty} \frac{dk}{2\pi}\, \phi(k) e^{-ikx} \tag{66}$$

(note the choice of where to put the factor $2\pi$ is not universal). Hence the characteristic function is also the expectation value of $e^{ikx}$;

$$\phi(k) = \langle e^{ikx} \rangle. \tag{67}$$

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

$$m_n = \left[ \frac{d}{d(ik)} \right]^n_{k=0} \phi(k). \tag{68}$$

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

$$\phi(k) = 1 + ik\langle x \rangle - \frac{1}{2}k^2\langle x^2 \rangle + \cdots. \tag{69}$$

As an example of a characteristic function lets consider the Poisson distribution.

$$\phi(k) = \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} e^{ikn} = e^{-\lambda} e^{\lambda e^{ik}}. \tag{70}$$

Hence the characteristic function for the Poisson distribution is

$$\phi(k) = e^{\lambda(e^{ik}-1)}. \tag{71}$$

Returning to the convolution equation (59),

$$p(z) = \int_{-\infty}^{\infty} dy\, p(z-y)p(y), \tag{72}$$

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

$$\begin{aligned} \phi_z(k) &= \int_{-\infty}^{\infty} dz\, p(z) e^{ikz} \\ &= \int_{-\infty}^{\infty} dz \int_{-\infty}^{\infty} dy\, p(z-y)p(y)e^{ikz} \\ &= \int_{-\infty}^{\infty} dz \int_{-\infty}^{\infty} dy\, (p(z-y)e^{ik(z-y)})(p(y)e^{iky}) \end{aligned} \tag{73}$$

Fourier transforming in the last equation we find

$$\phi_z(k) = \phi_x(k)\phi_y(k), \tag{74}$$

as expected from the properties of Fourier transforms.

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.

## 1.9 The Central Limit Theorem

The most important, and general, result from probability theory is the **Central Limit Theorem**. It is

- **Central** to probability and statistics - without it much of probability and statistics would be impossible,
- **A Limit Theorem** because it is only asymptotically true in the case of a large sample.

Finally, it is extremely general, applying 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, $\sigma^2$, 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, Poisson, $\chi^2$ or Student-t distribution 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. For a non-astrophysical example, the distribution of human heights is Gaussian, because the total effects of genetics and environment can be thought of as a sum of influences (random variables) that lead to a given height. The height of surface of the sea has a Gaussian distribution, as it is perturbed by the sum of random winds. The surface of planets has a Gaussian distribution due to the sum of all the factors that have formed them. The second consequence means that **many distributions**, under the right circumstances, **can be approximated by a Gaussian**. This is very useful since the Gaussian has many simple properties which we shall use later.

### 1.9.1 Derivation of the central limit theorem

Let

$$X = \frac{1}{\sqrt{n}}(x_1 + x_2 + \cdots + x_n) \tag{75}$$

be the sum of $n$ random variables $x_i$, each drawn from the same arbitrary underlying distribution function, (in general the underlying distributions can be different for each $x_i$, but for simplicity we shall only consider only one). The distribution of $X$'s generated by this summation, let's call it $p(X)$, 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 $p(x)$ is

$$\phi_x(k) = \int_{-\infty}^{\infty} dx\, p(x) e^{ikx} = 1 + i\langle x\rangle k - \frac{1}{2}\langle x^2\rangle k^2 + O(k^3), \tag{76}$$

where in the last term we have expanded out $e^{ikx}$. Since the sum is over $x_i/\sqrt{n}$, rather than $x_i$ we scale all the moments $\langle x^p\rangle \to \langle (x/\sqrt{n})^p\rangle$. From equation (76),we see this is the same as scaling $k \to k/\sqrt{n}$. Hence the characteristic function of $X$ is

$$\Phi_X(k) = [\phi_x(k/\sqrt{n})]^n \tag{77}$$

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

$$\Phi_X(k) = \left[1 - \frac{\sigma_x^2 k^2}{2n}\right]^n \to e^{-\sigma_x^2 k^2/2} \tag{78}$$

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. We know that the F.T. of a Gaussian is another Gaussian, but let's show that (by "completing the square"):

$$\begin{aligned} p(X) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, \Phi_X(k) e^{-ikX} \\ &= \frac{e^{-X^2/(2\sigma_x^2)}}{2\pi}\int_{-\infty}^{\infty} dk\, e^{(-\sigma_x^2 k^2 + X^2/\sigma_x^2 - 2ikX)/2} \\ &= \frac{e^{-X^2/(2\sigma_x^2)}}{2\pi}\int_{-\infty}^{\infty} dk\, e^{(X/\sigma_x - ik\sigma_x)^2/2} \\ &= \frac{e^{-X^2/(2\sigma_x^2)}}{\sqrt{2\pi}\sigma_x}. \end{aligned} \tag{79}$$

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

### 1.9.2 Measurement theory

As a corollary, by comparing equation (75) with the expression for estimating the mean from a sample of $n$ variables,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{80}$$

we see that the estimated mean from a sample has a Gaussian distribution with mean $m_1 = \langle x\rangle$ and variance $\sigma^2 = \mu_2/n$ 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 vanish 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.**

### 1.9.3 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 5). 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.

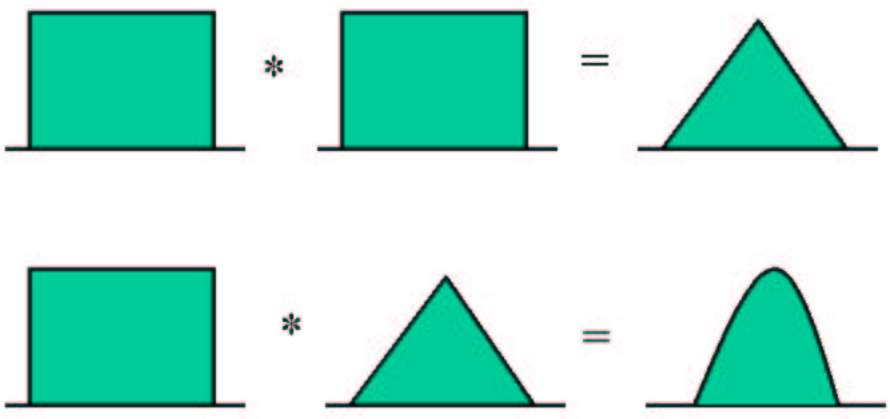

Figure 5: *Repeated convolution of a distribution will eventually yield a Gaussian if the variance of the convolved distribution is finite.*

## 1.10 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 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**,

$$p(x) = \frac{1}{\pi(1+x^2)} \tag{81}$$

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.

### 1.10.1 The sample variance

The mean of the sample is an estimate of the mean of the underlying distribution. Given we may not directly know the variance of the summed variables, $\sigma_x^2$, is there a similar estimate of the variance of $X$? 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

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x \rangle)^2, \tag{82}$$

where

$$\langle x \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{83}$$

is the sample mean of $n$ values. Let us find the expected value of this sum. First we

re-arrange the summation

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x\rangle)^2 = \frac{1}{n}\sum x_i^2 - \frac{2}{n}\sum_i\sum_k x_i x_k + \frac{1}{n^2}\sum_i\sum_k x_i x_j = \frac{1}{n}\sum x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \tag{84}$$

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

$$\langle \sum_i f(x_i)\rangle = \sum_i \langle f(x_i)\rangle \tag{85}$$

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

$$\langle x_i x_j\rangle = \langle x^2\rangle \qquad\qquad i = j, \tag{86}$$

and when $i$ and $j$ are different

$$\langle x_i x_j\rangle = \langle x\rangle^2 \qquad\qquad i \neq j. \tag{87}$$

The expectation value of our estimator is then

$$\begin{aligned}
\langle S^2\rangle &= \left\langle \frac{1}{n}\sum x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \right\rangle \\
&= \frac{1}{n}\sum\langle x^2\rangle - \frac{1}{n^2}\sum_i\sum_{j\neq i}\langle x\rangle^2 - \frac{1}{n^2}\sum\langle x^2\rangle \\
&= \langle x^2\rangle - \frac{n(n-1)}{n^2}\langle x\rangle^2 - \frac{n}{n^2}\langle x^2\rangle \\
&= \left(1 - \frac{1}{n}\right)\langle x^2\rangle - \frac{n(n-1)}{n^2}\langle x\rangle^2 \\
&= \frac{(n-1)}{n}(\langle x^2\rangle - \langle x\rangle^2).
\end{aligned} \tag{88}$$

The variance is defined as $\sigma^2 = \langle x^2\rangle - \langle x\rangle^2$, so $S^2$ will underestimate the variance by the factor $(n-1)/n$. This is because an extra variance term, $\sigma^2/n$, 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

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \langle x\rangle)^2 \tag{89}$$

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

Note that if the mean is know, and not estimated from the sample, this extra variance does not appear, in which case equation (82) is an unbiased estimate of the sample variance.

### 1.10.2 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

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\Delta m_i - \overline{\Delta m})^2 \tag{90}$$

is the unbiased estimate of the variance of $\Delta m$. 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

$$\sigma^2 = \frac{1}{n}\sum_{i}(\Delta m)^2. \tag{91}$$

Suppose we find that one of the $\Delta m$, say $\Delta m_i$, is very large, can we assess the significance of this result? One way to estimate its significance is from

$$t = \frac{\Delta m_i - \overline{\Delta m}}{S}. \tag{92}$$

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[9]. 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.

[9]The name 'Student' comes about because the first derivation was part of an undergraduate exam.

# PART TWO

# 2 STATISTICAL INFERENCE

## 2.1 Model fitting and parameter estimation

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 $\Omega_V$ and $\Omega_m$, 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 6). 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. **A test of how well the model fits the data.** ie. do we believe our model of the data?
2. **To find the allowable ranges of free parameters of the model.** ie. how do we decide on the range of models that adequately fits the data.

There are several methods which could be chosen.

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

### 2.1.1 The method of least-squares fits

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

$$S = \sum_i (D_i - M(x_i, \theta))^2, \tag{93}$$

with respect to the parameters, $\theta$. 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

$$S = \sum_i w_i (D_i - M(x_i, \theta))^2 \tag{94}$$

where $w_i$ are a set of arbitrary weights.

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

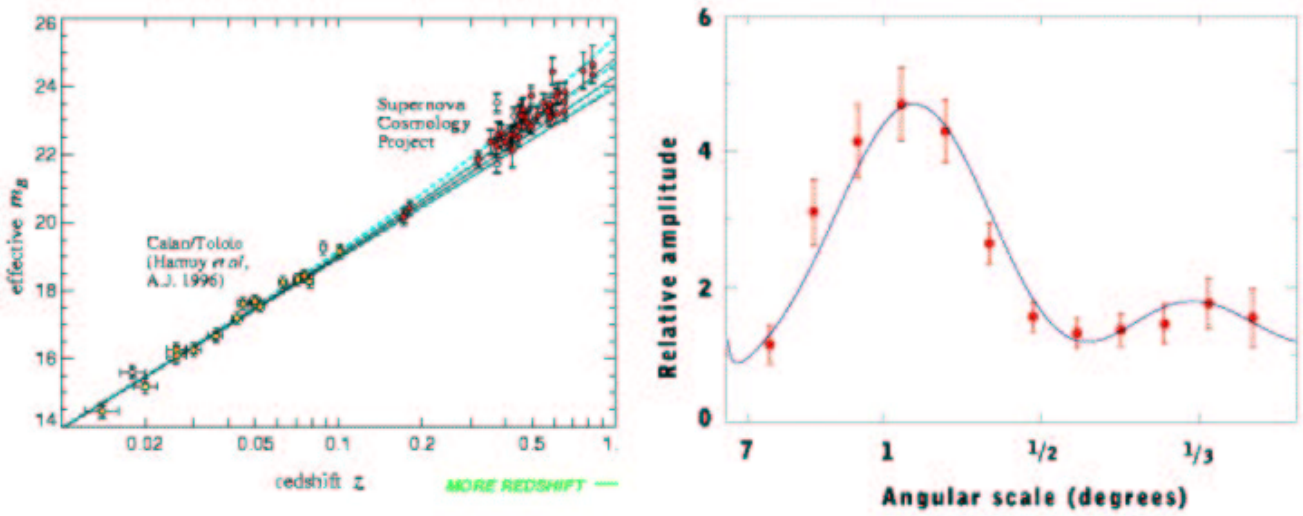


Figure 6: *LHS: Estimation of the acceleration of the Universe from Supernova Type 1a sources. RHS: Fitting the variance of temperature fluctuations in the Cosmic Microwave Background to data from the BOOMERANG balloon experiment. How do we know which curve is the best fit? And how do we know what values the parameters can take?*

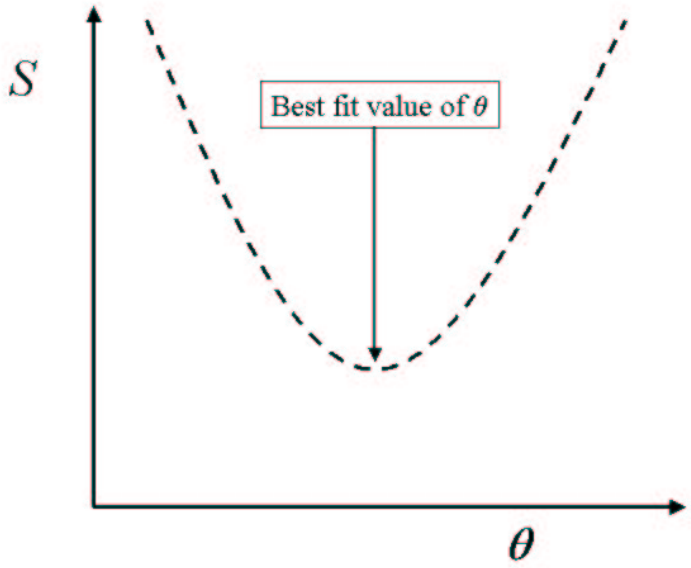


Figure 7: *Least-squares fitting. Usually we have to find the minimum of the function S by searching parameter space.*

In general it is not possible to find the best-fitting $\theta$ values in one single step: we have to evaluate $S$ for a range of values of $\theta$ and choose the values of $\theta$ which give the **smallest** value of $S$. Hence the name Least Squares.

### 2.1.2 Estimating a mean from Least Squares

What is the optimal (ie smallest error) estimate of the mean, $\langle x \rangle$, of a set of observations, $x_i$, and what is its variance? In this case we want to minimise the function

$$S = \sum_i w_i (x_i - \langle x \rangle)^2, \tag{95}$$

with respect to $\langle x \rangle$. Hence

$$\frac{\partial S}{\partial \langle x \rangle} = 2 \sum_i w_i (x_i - \langle x \rangle) = 0, \tag{96}$$

which has the solution

$$\langle x \rangle = \frac{\sum_i w_i x_i}{\sum_i w_i}. \tag{97}$$

The variance on this can be found by **Propagation of Errors** (see Section 1.7.3):

$$\begin{aligned} \sigma^2(\langle x \rangle) &= \sum_i \left( \frac{\partial \langle x \rangle}{\partial x_i} \right)^2 \sigma_i^2 \\ &= \frac{\sum_i w_i^2 \sigma_i^2}{(\sum_j w_j)^2} \end{aligned} \tag{98}$$

We can use this to find the set of weights which will minimise the error on the mean, by minimising $\sigma^2(\langle x \rangle)$ with respect to the $w_i$:

$$\frac{\partial \sigma^2(\langle x \rangle)}{\partial w_i} = -\frac{2 \sum_i w_i^2 \sigma_i^2}{(\sum_j w_j)^3} + \frac{2 w_i \sigma_i^2}{(\sum_j w_j)^2} \tag{99}$$

which implies

$$\sum_i w_i^2 \sigma_i^2 = w_i \sigma_i^2 \sum_j w_j. \tag{100}$$

This last equation is solved when

$$w_i = \frac{1}{\sigma_i^2}. \tag{101}$$

This set of weights is the **minimum variance weighting scheme**.

### 2.1.3 Multiparameter estimation

In many cases we will be concerned with many parameters $\theta_i$ in which case the process is extended to more dimensions: the following graph shows contours of $S$ as a function of two free parameters $\theta_1$ and $\theta_2$:

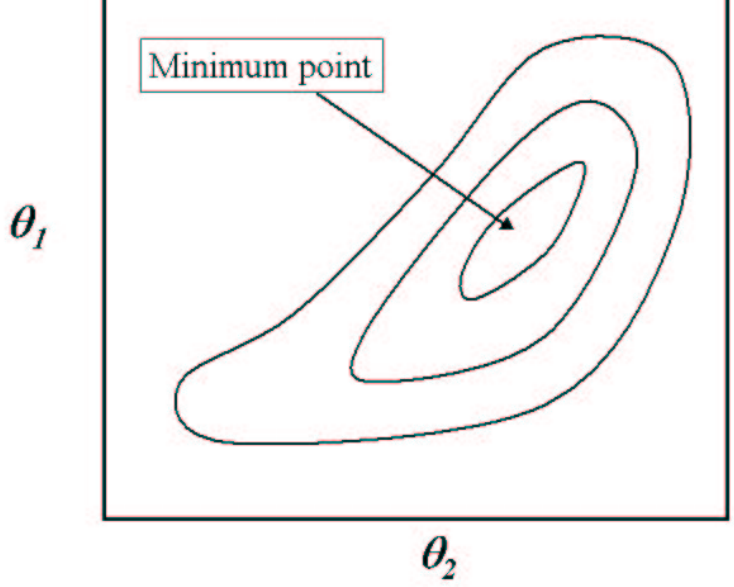


Figure 8: *Multi-parameter Least-squares fitting. Usually we have to find the minimum of the function $S$ by searching an $n$-dimensional parameter space. In practise this can be very difficult.*

If the model function $M(x,\theta) = y(x,\theta)$ can be expressed as a power series expansion of $x$ where the free parameters $\theta_i$ are the series coefficients, then the minimum point can be found in a single step:

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots = \sum_n \theta_n x_i^n \tag{102}$$

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

$$\begin{aligned} \frac{\partial S}{\partial \theta_m} &= \frac{\partial}{\partial \theta_m} \sum_j (y_j - \sum_n \theta_n x_j^n)^2 \\ &= 2 \sum_j (y_j - \sum_n \theta_n x_j^n) x_j^m = 0 \end{aligned} \tag{103}$$

where $\partial\theta_n/\partial\theta_m = 0$ if $n \neq m$ and 1 if $n = m$, since the parameters are independent. Equation (103) implies that for polynomial models of the data

$$\sum_j y_j x_j^m = \sum_j \sum_n \theta_n x_j^{n+m} \tag{104}$$

The $\theta_n$ 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 (104) can be written

$$A_m = B_{nm}\theta_n \tag{105}$$

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

$$\theta_n = B_{nm}^{-1} A_m \tag{106}$$

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

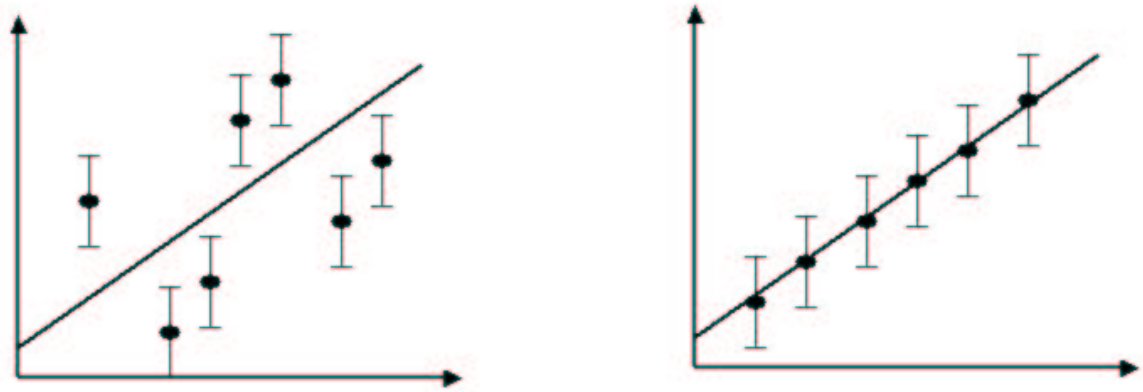

Figure 9: *LHS: Example of a fit to data where the $\chi^2$ value is too high. RHS: Example of a fit to data where the $\chi^2$ value is too low.*

### 2.1.4 Goodness of fit

Once we have found the least squares solution, we must now ask is the fit acceptable? To do this we find the distribution and ask if the cumulative probability, $P(> \chi^2)$ , is acceptable. If the measurement errors are Gaussian then by changing variables we find the distribution of $\chi^2$ is[10],

$$p(\chi^2|\nu) = \frac{e^{-\chi^2/2}}{\Gamma(\nu/2)} \left(\frac{\chi^2}{2}\right)^{\nu/2-1}, \tag{108}$$

where $\nu$ is the **number of degrees of freedom**, defined as the number of data points, $n_D$, minus the number of free parameters, $n_\theta$; $\nu = n_D - n_\theta > 0$. This can be used to test the **goodness of fit** by seeing if the value of $\chi^2$ is consistent with the known errors.

### 2.1.5 A rule of thumb for goodness of fit

A useful **rule of thumb** is that for a good fit

$$\chi^2 = \nu \pm \sqrt{2\nu}. \tag{109}$$

This approximation assumes that $\chi^2$ is large enough that $p(\chi^2|\nu)$ is nearly a Gaussian distribution with mean $\nu$ and variance $2\nu$ (the Central Limit Theorem in action again!).

[10]**Maths note:** The Gamma function is a standard function generalising factorials to real and imaginary numbers:

$$\begin{aligned} \Gamma(n+1) &= n! \\ \Gamma(n+1) &= n\Gamma(n) \\ \Gamma(1/2) &= \sqrt{\pi} \end{aligned} \tag{107}$$

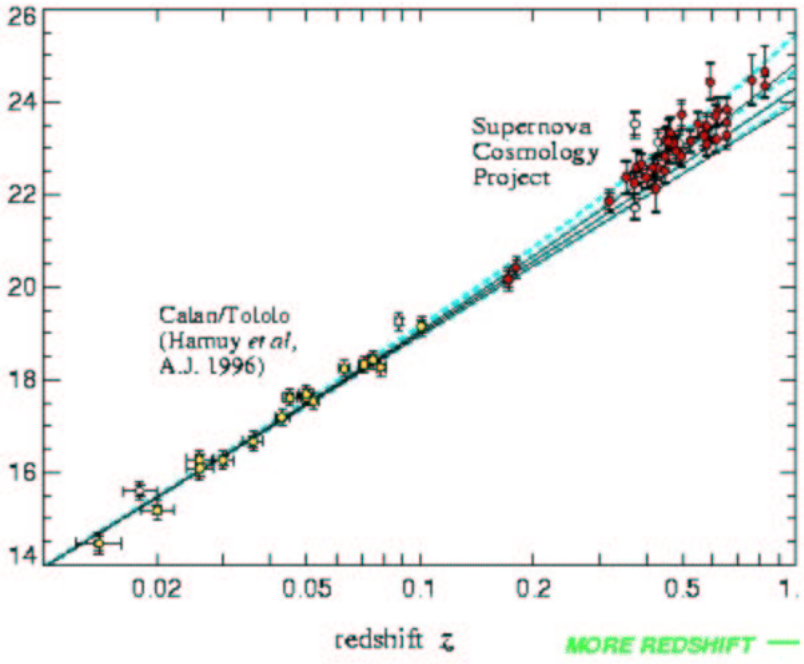


Figure 10: *The accelerating universe: line fitting to SNIa data.*

**Too high a value** ($\chi^2 \gg \nu + \sqrt{2\nu}$) 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** ($\chi^2 \ll \nu - \sqrt{2\nu}$) is as bad as too high a value. Too small a value may indicate hidden correlations between data points, or that the errors are overestimated. This should be investigated before proceeding further.

In general one should always check the value of $P(\geq \chi^2|\nu)$, or $\chi^2/\nu = 1 \pm \sqrt{2/\nu}$, rather than try "chi-by-eye" (ie guessing it's ok).

### 2.1.6 Confidence regions for minimum $\chi^2$

Having found the best-fitting values for a set of parameters, we also want to assign **limits** to the values which those parameters may have.

e.g. In fitting a Friedmann cosmological model to the Hubble diagram of Type Ia supernovae, we can find some best-fitting value for $\Omega_V$, the vacuum energy-density. What is the statistical uncertainty in the value of $\Omega_V$ due to the random errors in the data? If we can find two values of $\Omega_V$, $\Omega_{V,\rm min}$ and $\Omega_{V,\rm max}$ such that the probability that the observed values of $\chi^2$ or greater, $P(\geq \chi^2)$, should only occur with some probability, $x$ (say 0.32) then these values define the $(1-x)\times$ 100% **confidence region** (say, 68%) for $\Omega_V$.

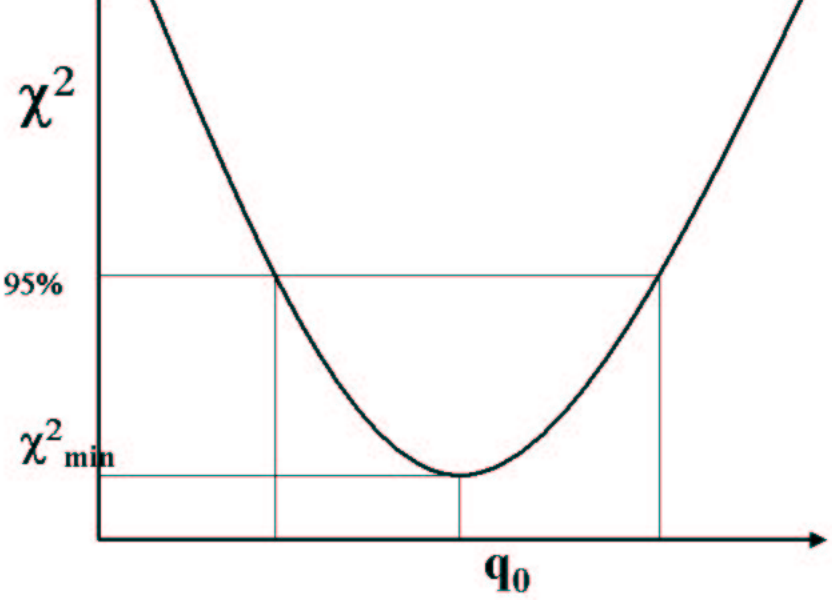


Figure 11: *Confidence regions for estimating $\Omega_V$ from supernova data.*

In the above example (Figure 11) the 68% confidence limits for $\Omega_V$ are 0.5 and 0.9: $\Omega_V$ lies within that range with 68% confidence: there is a 32% chance that the true value of $\Omega_V$ is outside that range, with the errors on the data conspiring to make the data appear consistent with $\Omega_V = 0.7$.

Again, these limits only have any meaning if we are confident that the model is correct. On the Hubble diagram for example, a good fit to the data can be found for $\Omega_V = 0.7 \pm 0.2$ (68% confidence level), but this would be meaningless if the effects of evolution change the luminosity of supernova at high redshifts so that the Hubble diagram is actually determined much more by the evolution than by the value of $\Omega_V$.

Bearing this in mind, **how do we evaluate these confidence limits?**

It can be shown, assuming Gaussian statistics, that the $x$% confidence region for one free parameter is bounded by the value of $\chi^2$ such that if $\chi^2 = \chi^2_{\rm min} + \Delta\chi^2$ then $\Delta\chi^2$ is determined by the distribution of $\chi^2$ for one degree of freedom, where here the number of degrees of freedom equals just the number of free parameters.

For example the probability that the solution lies within $\Delta\chi^2 = 1$ is $P(x \leq \Delta\chi^2 | \nu = 1)$ which is 68.3% or 1-$\sigma$ uncertainty for Gaussian errors. This can be extended to larger numbers of free parameters and degrees of confidence. For instance for two parameter we use the $\Delta\chi^2 = 2.3$ threshold for 68% confidence regions.

We can be even more stringent on how we quote errors. For a single parameter we may want to quote the range of values where there is a 95% chance that the true values lies. In this case we must find the range of solutions which lies within $\Delta\chi^2 = 4$. For 99% confidence regions we require $\Delta\chi^2 = 7$, for a single parameter.

Note that in the goodness-of-fit test, the number of degrees of freedom, $\nu$, for the distribution of $\chi^2$ was given by the number of data points minus the number of free parameters to be fitted. In the case of parameter estimation, the number of degrees of freedom for $\Delta\chi^2$ is just the number of free parameters.

## 2.2 Maximum Likelihood Methods

Minimum-$\chi^2$ is a useful tool for estimating goodness-of-fit and confidence regions on parameters. However it can only be used in the case where we have a set of measured data values with a known, Gaussian-distributed, scatter about some model values. A more general tool is to use the technique of **Maximum Likelihood**.

The "likelihood function" is the joint probability distribution, $p(D_1, \cdots, D_n|\theta_1, \cdots, \theta_m)$ of $n$ measured data values, $D_i$, drawn from a model distribution with $m$ free parameters $\theta_j$:

$$L(D_1, \cdots, D_n|\theta_1, \cdots, \theta_m) = p(D_1, \cdots, D_n|\theta_1, \cdots, \theta_m). \tag{110}$$

If the data values are independent then;

$$\begin{aligned} L(D_1, \cdots, D_n|\theta_1, \cdots, \theta_m) &= p(D_1|\theta_1, \cdots, \theta_m) \cdots p(D_n|\theta_1, \cdots, \theta_m) \\ &= \prod_{i=1}^{n} p(D_i|\theta_1, \cdots, \theta_m). \end{aligned} \tag{111}$$

The technique of maximum likelihood **maximises** this function with respect to the free parameters $\theta$. That is, it chooses the values of parameters which most closely simulate the distribution of the observed data values. In practice, it is easier to maximise $\ln(L)$ rather than $L$: i.e.

$$\frac{\partial}{\partial\theta} \ln L = 0. \tag{112}$$

The method is particularly useful when the type of models being investigated are not those that predict a value for each of a set of measured quantities, but rather those that predict the statistical distribution of values in a sample. In this case, one may not know the error on any individual measurement. It is also very useful when the distribution of the data is known to be non-Gaussian. In this case Maximum Likelihood is both very versatile and simple to apply.

### 2.2.1 The flux distribution of radio sources

The distribution of flux densities of extragalactic radio sources are distributed as a power-law with slope $-\alpha$, say. In a non-evolving Euclidean universe $\alpha = 3/2$ (can you prove this?) and departure of $\alpha$ from the value 3/2 is evidence for cosmological evolution of radio sources. This was the most telling argument against the steady-state cosmology in the early 1960's (even though they got the value of a wrong by quite a long way).

Given observations of radio sources with flux densities $S$ above a known, fixed measurement limit $S_0$, what is the best estimate for $\alpha$?

The model probability distribution for $S$ is

$$p(S)dS = (\alpha - 1)S_0^{\alpha-1}S^{-\alpha}dS \tag{113}$$

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

$$\int_{S_0}^{\infty} dS\, p(S) = 1. \tag{114}$$

So the likelihood function $L$ for $n$ observed sources is

$$L = \prod_{i=1}^{n} (\alpha - 1) S_0^{\alpha-1} S_i^{-\alpha} \tag{115}$$

with logarithm

$$\ln L = \sum_{i=1}^{n} (\ln(\alpha - 1) + (\alpha - 1) \ln S_0 - \alpha \ln S_i). \tag{116}$$

Maximising $\ln L$ with respect to $\alpha$:

$$\frac{\partial}{\partial \alpha} \ln L = \sum_{i=1}^{n} \left( \frac{1}{\alpha - 1} + \ln S_0 - \ln S_i \right) = 0 \tag{117}$$

we find the minimum when

$$\alpha = 1 + \frac{n}{\sum_{i=1}^{n} \ln \frac{S_i}{S_0}}. \tag{118}$$

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

$$\alpha = 1 + \frac{1}{\ln 2} = 2.44 \tag{119}$$

but with a large uncertainty. Clearly, as $S_i = S_0$ we find $\alpha \to \infty$ as expected. In fact $\alpha = 1.8$ for bright radio sources at low frequencies, significantly steeper than 1.5.

### 2.2.2 Goodness-of-fit and confidence regions from maximum likelihood

In cases where we are fitting a set of data values with known Gaussian errors, then maximum-likelihood is precisely equivalent to minimum-$\chi^2$, since in this case $\ln L = -\chi^2/2$. The distribution of $-2 \ln L$ is then the $\chi^2$-distribution and hence we can readily calculate a goodness of fit.

In cases such as the example in Section 2.2.1, however, the probability distribution assumed for the model is not Gaussian – it is the power-law model for the source counts. In this case we do not know what the expected distribution for the ln(likelihood) actually is, and hence **we cannot estimate a goodness-of-fit**.

In the limit of large numbers of observations, we can still estimate confidence regions for parameters, however (another example of the operation of the **Central Limit Theorem**) in that we expect the distribution of $-1/2 \ln L$ to have the $\chi^2$ distribution with the number of degrees of freedom equal to the number of free parameters (c.f. the estimation of confidence regions in the case of minimum $\chi^2$).

In practice one would instead plot out the likelihood surface in parameter space. In the case of two parameter this will generally be an ellipse.

### 2.2.3 Estimating parameter uncertainty

Although plotting out the likelihood contours around the maximum is a good indicator of the parameter uncertainty, it can sometimes be computationally expensive, especially for large data sets, and many parameters. An alternative is to use the second derivative of the log-likelihood:

$$\sigma^2(\theta_i) = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1} \tag{120}$$

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

$$L(\theta) = \frac{e^{-\Delta\theta^2/(2\sigma^2(\theta))}}{\sqrt{2\pi}\sigma(\theta)} \tag{121}$$

where $\Delta\theta = \theta - \theta_{\max}$. Equation (120) can be verified by expand $\ln L$ to 2nd order in $\Delta\theta$ and differentiating. This can be generalized to multiple parameters by the covariance matrix

$$\langle \Delta\theta_i \Delta\theta_j \rangle = F_{ij}^{-1} \tag{122}$$

where

$$F_{ij} = -\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \tag{123}$$

is called the **Fisher Information Matrix**, after the Statistician R.A. Fisher. This can be used to estimate both the **conditional errors**, $\sigma^2_{\rm cond}(\theta) = 1/F_{ii}$, where we assume the other parameters are known, and the **marginalised errors**, $[F^{-1}]_{ii}$. These two quantities satisfy the relation

$$1/F_{ii} \leq [F^{-1}]_{ii}. \tag{124}$$

The Fisher Matrix is commonly used in cosmology in both survey design and analysis.

## 2.3 Hypothesis testing

### 2.3.1 Introduction

The concept of testing goodness-of-fit of a model, and of estimating confidence regions for parameters, is part of the more general philosophical problem of hypothesis testing:

**Given a set of data, can we distinguish between two (or more) hypotheses or models which predict those data values?**

In some sense, we wish to compare the goodness of fit of the two hypotheses. There tend to be two schools of thought among statisticians, the Frequentists and Bayesians we met at the beginning, over how to do this, and the subject is very controversial. The problem arises from Bayes' Theorem, published posthumously in 1763, which describes how to new information changes a probability distribution. It was really the first attempt to use statistical axioms to make decisions on hypotheses. If we follow the logical consequences of Bayes' work, we find that the tests of goodness-of-fit of models described above are fundamentally incomplete in their treatment of the problem.

### 2.3.2 Bayes Theorem

Suppose we are given some information about a physical experiment, which we shall denote by $I$. We carry out the experiment and get a result, or set of data, denoted $D$. Now we want to choose between two hypotheses $H_1$ and $H_2$. From the axioms we met in Lecture 1, we can say that the conditional joint probability of obtaining the result $D$ **and** of $H_i$ being correct, given our prior information $I$, is

$$p(H_i D|I) = p(D|I)p(H_i|DI). \tag{125}$$

That is, the joint probability of the hypothesis and data (result) being true, given the initial information, $p(H_1 D|I)$, is the probability of obtaining that data set given only the initial information, $p(D|I)$, multiplied by the probability of $H_i$ being correct given the prior information $I$ and data $D$, $p(H_i|DI)$ (this is known as the a **posteriori probability**). This is not very useful, but if we swap $H_i$ and $D$ we find

$$p(H_i D|I) = p(H_i|I)p(D|H_i I), \tag{126}$$

which now says that this joint probability is the **likelihood**, $p(D|H_i I)$ (the probability of getting a data set, $D$, from hypothesis $H_i$ given $I$), multiplied by the **prior probability**, $p(H_i|I)$, of the hypothesis $H_i$ being true in the absence of any data other than $I$. Thus

$$p(H_i|DI) = \frac{p(H_i|I)p(D|H_i I)}{p(D|I)}. \tag{127}$$

This is **Bayes' Theorem**. If we want to choose between the two hypotheses $H_1$ and $H_2$, we should evaluate $p(H_1|DI)$ and $p(H_2|DI)$ and choose the most probable result. The factor, $p(D|I)$, called the **evidence**, is the same for $H_1$ and $H_2$, so we can use the likelihood ratio,

$$\frac{p(H_1|DI)}{p(H_2|DI)} = \frac{p(H_1|I)p(D|H_1 I)}{p(H_2|I)p(D|H_2 I)} \tag{128}$$

to choose the higher probability.

If we assume that the two prior probabilities are equal, then we just choose the hypothesis with the highest likelihood. In fact, it is common to ignore Bayes' Theorem entirely and simply determine the likelihoods, and proponents of such tests do not consider the possible non-equality of the prior probabilities. Indeed, Bayes postulated that in the absence of any information to the contrary we should assume equal prior probabilities (Bernoulli's "Principle of Insufficient Reason").

### 2.3.3 Updating the probability of a hypothesis

If we now obtain some more information, perhaps from a new experiment, then we can use Bayes' theorem to update our estimate of the probabilities associated with each hypothesis. The problem reduces to that of showing that adding the results of a new experiment to the probability of a hypothesis is the same as doing the two experiments first, and then seeing how they both effect the probability of the hypothesis. In other words it should not matter how we gain our information, the effect on the probability of the hypothesis should be the same.

We start with Bayes' expression for the posteriori probability of a hypothesis,

$$p(H|DI) = \frac{p(H|I)p(D|HI)}{p(D|I)}. \tag{129}$$

Let say we do a new experiment with new data, $D'$. We can then express this by transforming equation (129) by $D \to D'$ and letting the old data become part of the prior information $I \to DI$. Bayes' theorem is now

$$p(H|D'DI) = \frac{p(H|DI)p(D'|HDI)}{p(D'|DI)}. \tag{130}$$

We now notice that the new prior in this expression is just the old posteriori probability from equation (129), and that the new likelihood is just

$$p(D'|DHI) = \frac{p(D'D|HI)}{p(D|HI)}. \tag{131}$$

Substituting in the old posteriori probability and this expression for the new likelihood we find

$$p(H|D'DI) = \frac{p(H|I)p(D'D|HI)}{p(D'D|I)} \tag{132}$$

which has the same form as equation (129), the outcome from the initial experiment, but now with the new data incorporated, ie. we have shown $D \to D'$ and $I \to DI$ is the same is $D \to D'D$. This shows us that it doesn't matter how we add in new information. **Bayes' theorem gives us a natural way of improving our statistical inferences as our state of knowledge increases.**

### 2.3.4 The prior distribution

In Lecture 1 we discussed that Frequentists and Bayesians have different definitions of probability, which can lead to different results. These discrepancies occur particularly in parameter estimation: a Bayesian would multiply the likelihood function for a given parameter value by a prior probability. The maximum likelihood estimate of $\alpha$ in Section 2.2.1 is Bayesian if the prior probability is assumed to be a constant per unit $\alpha$.

But the Bayesian estimate will change if we assume a different prior, such as constant probability per unit $\ln \alpha$. We may do this if $\alpha$ has to be positive, or if we want to search for the most likely $\alpha$ over many orders of magnitude. In this case

$$p(\ln \alpha) = \frac{1}{\alpha} p(\alpha). \tag{133}$$

Here our initial problem of how we assign probabilities becomes most apparent. It is this arbitrariness that is unattractive to Frequentists, who ignore the prior. The Bayesian point of view is that the prior exists and so we must choose something.

But it is as well to be aware that ignoring priors is equivalent to making a specific choice of prior by default. Hypothesis testing and parameter estimation is a minefield of philosophical uncertainty! Some astronomers use both Bayesian and Frequentist methods

to satisfy both. Another view is that if the choice of prior really matters, there cannot be enough information in the data to decide which is the most likely hypothesis.

In the next section we look at one application where Bayesian methods have become widely accepted.

## 2.4 Imaging process and Bayes' Theorem

### 2.4.1 Using the prior



A common problem in image processing is to correct an image for the two main processes which degrade its appearance and hence our interpretation of the scene which it captures. These two processes are the **convolution** (or smoothing, or blurring) of an image by some instrumental response, and the **addition of noise** to the image (which depending on the application may arise before the blurring process, as in radioastronomy interferometer images, or after the blurring process, as in images from optical telescopes).

Even if we knew precisely the blurring function of the system and its noise properties, there are two reasons why we cannot precisely reconstruct what the scene should look like if viewed in the absence of these factors.

- First, it may not be possible to unambiguously deconvolve the blurring. Remember that convolution by a function is equivalent in Fourier space to multiplying the scene's Fourier transform by the Fourier transform of the blurring function. Hence deconvolution is the process of dividing the transform of the observed image by the transform of the blurring function. In general that process will not be defined at all spatial frequencies, and some spatial frequencies in the result will have undefined values. Even in regions where the division is defined, the addition of noise to the system is likely to have an unstable effect in the deconvolving process.

- Second, the noise is a random statistical process, and so we cannot unambiguously calculate the intensity at a particular point in a scene, given the observed image: we can only calculate probabilities that the scene has any particular value at each point.

So image reconstruction can be seen to be a process of statistical inference, in which we have to estimate the most likely (in some sense not yet defined) values for the scene given what we know about the problem: we want to find the best-fitting model scene. Now in this case, straightforward application of the maximum likelihood technique to choose the best-fitting model will not work: we have $n$ pixels of data, but each of those has to be a free parameter in the model. If we want to make progress we need to add in some extra information, and Bayes' theorem gives us a way of doing that. We should choose the model scene which maximises

$$p(\text{scene}|\text{initial information})p(\text{image data}|\text{scene, initial information}). \qquad (134)$$

The second term is the likelihood, but we now have the ability to bring additional information to the problem by means of the prior.

What can we say about the scene before we attempt to measure an image of it? Remember that one of the problems in the deconvolution is that at some spatial frequencies the Fourier transform may be undefined. Equivalently, the final deconvolved image could have sinusoidal waves of arbitrary amplitude imposed upon it, unless we take steps to suppress them. Similarly, the presence of noise would allow large fluctuations to exist at high spatial frequencies in the image, which in reality would not be appropriate. So one way to carry out the scene reconstruction is to enforce a prior which penalises solutions which are not smooth.

A number of schemes exist to do this: perhaps the most publicised is the **Maximum Entropy** prior.



### 2.4.2 Maximum Entropy

Imagine constructing a scene by randomly putting $N$ photons into $m$ pixels. In each pixel there will be $n_i = Np_i$ photons. The number of ways in which we can randomly place $N$ identical photons in $m$ pixels is

$$\frac{N!}{n_1!n_2!\cdots n_m!} \rightarrow \exp\left(-\sum_i n_i \ln\left(\frac{n_i}{N}\right)\right) \tag{135}$$

where the right hand side limit can be derived by taking the natural log and using Stirling's formula;

$$\begin{aligned} \ln N! - \sum_i \ln n_i! &\rightarrow -N + N\ln N - \sum_i(-n_i + n_i \ln n_i) \\ &= -N + \sum_i n_i \ln N + N - \sum_i n_i \ln n_i \\ &= -\sum_i n_i \ln\left(\frac{n_i}{N}\right). \end{aligned} \tag{136}$$

We have assumed each photon is indistinguishable. The quantity in the outer brackets in equation (135), is known as the **image entropy**, $H$, by analogy with the entropy of statistical mechanics, which can also be deduced by considering the multiplicity of microstates. A prior probability can thus be assigned for each scene which is just proportional to $\exp[H]$. This is the usefulness of Max-Ent, that it gives us a way of choosing the prior in the case of image processing.

For large $N$ this prior is equal to the number of ways the scene could have been **randomly generated**. In this "random scene" model the prior is a maximum when every pixel has an equal number of photons: the "simplest" scene, the one which adds in the least amount of structure consistent with the observed data, is preferred over more complicated scenes. One feature of this choice of prior is that the reconstructed scene must be positive: you can't have a scene with negative numbers of photons. For many images this is a good property to embody into the prior, but this may not be appropriate for other types of image where negative values may be allowed.

In practice, we may have more information about the scene than assumed above, in which case we can incorporate that information also into the maximum entropy prior.

Maximum entropy does not force us to favour only a completely uniform intensity distribution.

**Maximum entropy** has found favour largely because it works, but there are clearly other choices of prior which could be made. Note that the maximum entropy prior does not contain any information about the intensity of pixels relative to their neighbours, such as you might require if you wanted to adopt the smoothest solution to the scene restoration. If the pixels in the scene were randomly permuted in position, the image entropy would be unaffected. This suggests that, for images with complex structure, a choice of prior which **did** incorporate information about pixel inter-relationships might be more effective.

The idea of Maximum Entropy has been extended by E. Jaynes, who has suggested that it is a way of subjectively solving the problem of how to choose a prior in any situation, and has further elaborated on the connection with statistical physics . However, to maintain rigour in this approach, the methods can be rather complex. Again using this approach become a matter of personal choice.



## 2.5 Non-parametric statistics

A **non-parametric statistical test** is one in which the outcome **does not depend on the probability distribution** which generated the data. Some statisticians prefer to call such tests "distribution-free". By implication, it is not necessary to know that the probability distribution that generated the data values (and their errors) was in order to apply a non-parametric test. This is a vital property in many areas (from psychology to astronomy) where the distributions are not known. The tests are usually quick and easy to apply – another practical advantage.

If the underlying distribution is known then a parametric test may be more powerful, but usually this advantage is small and is more than offset by the advantage of the non-parametric tests. Non-parametric tests are also often valid for very small samples.

There are three important tests which I shall deal with here,

1. **The $\chi^2$ Goodness-of-Fit Test.**
2. **The Kolmogorov-Smirnov Test**
3. **The Spearman Rank Correlation Coefficient.**

### 2.5.1 The $\chi^2$-goodness-of-fit test

We have already met the parametric $\chi^2$ goodness-of-fit test, in which data values $x_i$ are compared with model predictions. If the errors on the data values are **Gaussian distributed**, then the quantity

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \lambda_i)^2}{\sigma_i^2} \tag{137}$$

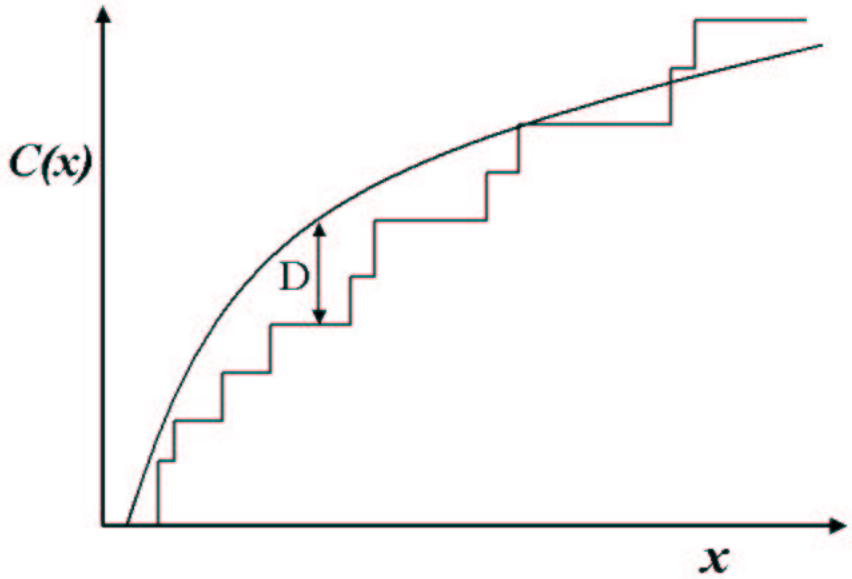


Figure 12: *The K-S test comparing the cumulative distribution of a data set with a theoretical cumulative distribution.*

(where $\lambda_i$ are model predictions; $\sigma_i^2$ are model variances) has a $\chi^2$ distribution with $n$ degrees of freedom, or $(n - n_p)$ if there are $n_p$ free parameters whose values have been estimated from the data by a $\chi^2$ procedure.

The **non-parametric version** of this test is obtained by **binning the data values** $x_i$ and then comparing the numbers of values in each bin with the numbers predicted by the model, rather than comparing the $x$ values themselves. If the model predicts $n_i$ values, then the probability of obtaining the observed number, $N_i$, is given by the **Binomial Distribution**. In the limit of large $n_i$ the binomial distribution tends to the normal distribution with variance $n_i$, so for large $n_i$ we expect

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - n_i)^2}{n_i} \tag{138}$$

to tend towards the $\chi^2$ distribution, regardless of the model distribution which produced the observed values. Thus for large $n_i$ this test is distribution free.

In fact, the distribution tends towards the $\chi^2$ distribution remarkably quickly: $n_i$ need only be as large as 5 in each bin for the test to be accurate.

The expected value of is about the number of degrees of freedom. Quite often, the model predictions have to be normalised to the total number of observed values, in which case one degree of freedom is lost: $k$ bins result in $(k - 1)$ degrees of freedom. Again, estimating free parameters reduces the number of degrees of freedom further.

Perhaps the worst point about the $\chi^2$ test is that the data have to be binned, resulting in a loss of information and hence power. Also, small samples cannot adequately be tested.

### 2.5.2 The Kolmogorov-Smirnov test

A very powerful test which can be used for small samples, does not require binning, and is completely distribution free is the **Kolmogorov-Smirnov (KS) test**. The one-sample test tests for a difference between an observed and a model probability distribution. It does this by looking for the largest difference between the normalised cumulative frequency distributions.

Suppose there are $n$ data values and that $C_n(x)$ is their cumulative distribution. If the $n$ events are located at $x_i$, where $i = 1, \cdots, n$, then $C_n(x)$ is the fraction of data points to the left of a given value $x$. This function is constant between consecutive points and jumps by a constant value, $1/n$, at each $x_i$. Comparing this distribution with the model cumulative distribution, $M(x)$, the K-S test is just

$$D = \max_{-\infty<x<\infty}|C_n(x) - M(x)|, \tag{139}$$

ie it is just the maximum difference between the cumulative data and the model. This type of statistic is most sensitive to differences in the mean of two distributions.

The KS test is clearly insensitive to a rescaling of $x$. For example it is the same using $x$ or $\ln x$. In addition the distribution of $D$ can be calculated as the probability of finding the difference $D$ at the $i^{th}$ value is just a binomial distribution. It is beyond this course to calculate the distribution of $D$ so we shall merely quote it here for completeness:

$$P(> D) = 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2ni^2D^2} \tag{140}$$

for large $n$.

The test is very powerful and distribution free. It can be used to place confidence limits on the possible values of a cumulative distribution, and it can be generalised to a two-sample test for comparing two observed distributions to see whether they have been drawn from different model distributions.

### 2.5.3 The Spearman rank correlation coefficient

One of the most powerful techniques used in statistical astronomy is to infer relationships between quantities by searching for a correlation between them
*e.g. is there a link between radio and X-ray emission from quasars?*
Plot $L_X$ (X-ray luminosity) against $L_R$ (radio luminosity) for a sample (Figure 13). If a statistically significant correlation exists, we may infer that $L_X$ and $L_R$ are related – although as we shall see later we have to be very careful making that inference.

Standard parametric correlation coefficients can be very useful, but they need to make some assumption about the distribution of the scatter about the correlation. If the scatter is due to Gaussian measurement errors, a parametric test, such as the parametric $\chi^2$ test, may be fine.

But in astronomy the scatter on such graphs is usually caused by an **intrinsic dispersion** in the physical properties of the objects rather than by measurement errors.

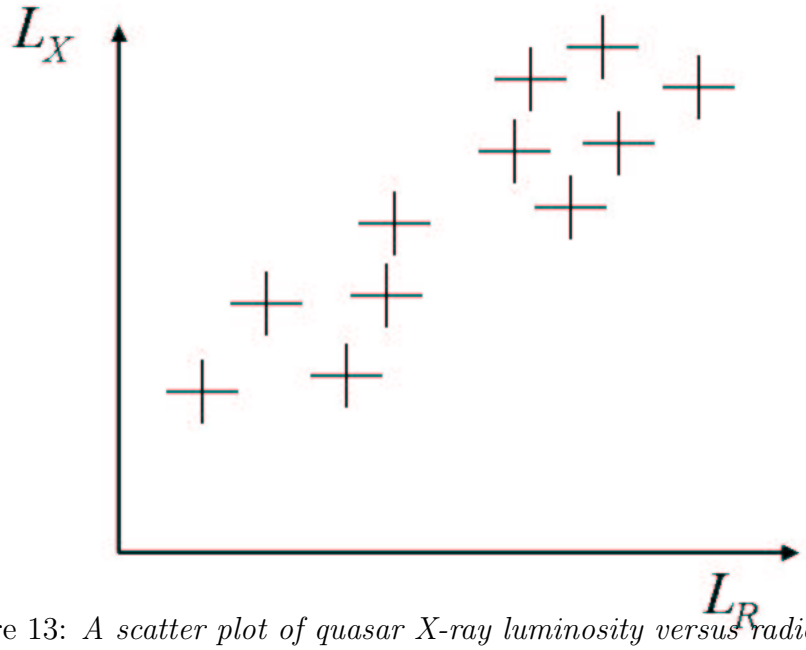


Figure 13: *A scatter plot of quasar X-ray luminosity versus radio luminosity.*

For example in the quasars' correlation above there is a scatter of nearly a factor 10 in luminosity - far greater than the measuring error. And of course we have no idea what the distribution of that "cosmic scatter" might be. So we need a **non-parametric correlation coefficient**.

Suppose we have data values $X_i$ and $Y_i$, with means $\langle X \rangle$ and $\langle Y \rangle$ . Let us define the new variables $x_i = X_i - \langle X \rangle$ and $y_i = Y_i - \langle Y \rangle$. A general expression for their **correlation coefficient** (see Section 4) is

$$r = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_j y_j^2}} \tag{141}$$

The **Spearman rank correlation coefficient** is obtained by replacing the $x_i$ values by the rank, $\tilde{x}_i$ (or the integer order that the values lies in) in the $x$ direction, and similarly replace $y_i$ by its rank $\tilde{y}_i$. The correlation coefficient, $r_s$ obtained from these new ranked variables is the Spearman rank correlation coefficient:

$$r_s = \frac{\sum_i \tilde{x}_i \tilde{y}_i}{\sqrt{\sum_i \ \tilde{x}_i^2 \sum_j \tilde{y}_j^2}} \tag{142}$$

This may be re-cast into a more convenient form by defining $d_i$, the difference between the rank in the $x$ direction and the rank in the $y$ direction,

$$d_i = (\tilde{x}_i - \tilde{y}_i). \tag{143}$$

In this case

$$\begin{aligned} \sum_i d_i^2 &= \sum_i (\tilde{x}_i - \tilde{y}_i)^2 \\ &= \sum_i (\tilde{x}_i^2 + \tilde{y}_i^2 - 2\tilde{x}_i \tilde{y}_i) \\ &= \sum_i (\tilde{x}_i^2 + \tilde{y}_i^2) - 2r_s \sqrt{\sum_i \tilde{x}_i^2 \sum_j \tilde{y}_j^2}. \end{aligned} \tag{144}$$

rearranging this last term we find

$$r_s = \frac{\sum_i (\tilde{x}_i^2 + \tilde{y}_i^2 - d_i^2)}{2\sqrt{\sum_i \tilde{x}_i^2 \sum_j \tilde{y}_j^2}} \tag{145}$$

Since the original quantities $\tilde{X}$ and $\tilde{Y}$ are ranks we can calculate that

$$\begin{aligned} \sum_i \tilde{X}_i &= n(n+1)/2 \\ \sum_i \tilde{X}_i^2 &= n(n+1)(2n+1)/6 \end{aligned} \tag{146}$$

so that

$$\begin{aligned} \sum_i \tilde{x}_i^2 &= \sum_i (\tilde{X}_i - \langle \tilde{X} \rangle)^2 \\ &= \sum_i \tilde{X}_i^2 - \frac{1}{n}\left(\sum_i \tilde{X}_i\right)^2 \\ &= \frac{n(n^2-1)}{12}. \end{aligned} \tag{147}$$

Putting this all together we find

$$r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}. \tag{148}$$

As with other correlation coefficients, $-1 \leq r_s \leq 1$.

The significance of any value of $r_s$ depends on the sample size, and can be evaluated from basic probability theory. Suppose the null hypothesis is that there is no correlation: there is no relation in the population between the $x$ and $y$ values. So for any given rank order of the $y$ values any order of the $x$ values is just as likely as any other. For $n$ sample members, there are $n!$ possible rankings of the $x$ values for any order of $y$ values. So the probability of occurrence of any particular ranking of the $x$ scores with any particular ranking of the $y$ scores is $1/n!$.

For each of these combinations there will be a value of $r_s$, so the probability of occurrence of any value of $r_s$, assuming the null hypothesis, is the number of permutations giving rise to that value divided by $n!$.

e.g.

1. $n = 2$: there are only two possible values for $r_s$: $r_s = \pm 1$. Each can occur with probability 1/2.
2. $n = 3$, $r_s$ can be $-1$, $-1/2$, $1/2$, 1 with respective probabilities 1/6, 1/3, 1/3, 1/6 .

**The significance of $r_s$:**
This process can be carried out for any $n$, although for large $n$ it can be shown that

$$t = \frac{r_s}{\text{sampling error of } r_s} = r_s \left(\frac{n-2}{1-r_s^2}\right)^{1/2} \tag{149}$$

has the **Student-t distribution** with $\nu = n - 2$ degrees of freedom. Again in the limit of large numbers, $n \gg 1$, we can appeal to the Central Limit Theorem and see that $1/t$ is the fractional uncertainty on $r_s$.

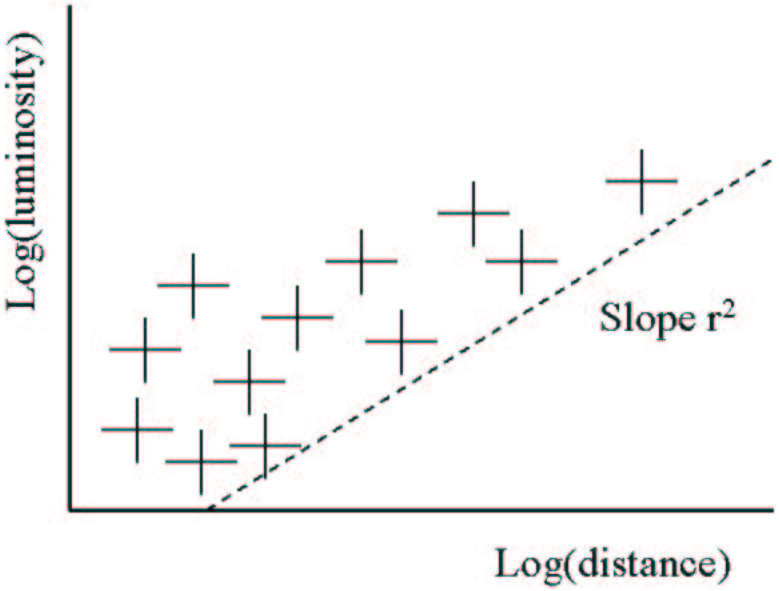


Figure 14: *The distribution of intrinsic luminosities as a function of distance.*

## 2.6 Inference from correlations

Having found a **statistically-significant correlation** between two variables, we would then like to infer that there is a **physical link** between them (e.g. X-ray emission from quasars is caused by inverse-Compton scattering in the radio-emitting region).

But we have to be very careful at this point, because artificial correlations can be and often are induced by the method of sample selection, or by each variable being correlated with a hidden third variable. These effects are known as **selection effects** and are very serious in almost all astronomical data.

The most common selection effects arise in samples in which the members have been selected as being brighter than some apparent brightness (magnitude, flux density or whatever). But any given apparent brightness corresponds to higher and higher intrinsic luminosities as the sample members are located at greater distances. Let us look at the distribution of intrinsic luminosities in such a sample as a function of distance (Figure 14).

We will not include any objects to the right of the dashed line as they will fall below our apparent brightness limit. The dashed limit has slope 2 in a Euclidean universe since

$$L_{\min} = 4\pi r^2 S_{\min} \tag{150}$$

where $r$ is the distance to the object and $S_{\min}$ is the sample flux density limit. We can see that we have induced a strong correlation between luminosity and distance which is purely a selection effect. This is known as **Malmquist bias**[11].

In practice the artificial correlations are often even stronger than suggested by the figure above, since usually objects of higher luminosity are rarer than those of lower

[11] In fact Eddington, in 1915, was the first to consider statistical biases which arise using a distance dependent observable such as apparent magnitude. Malmquist gave a classical discussion of the effect in the context of stellar magnitudes and star counts in 1920.

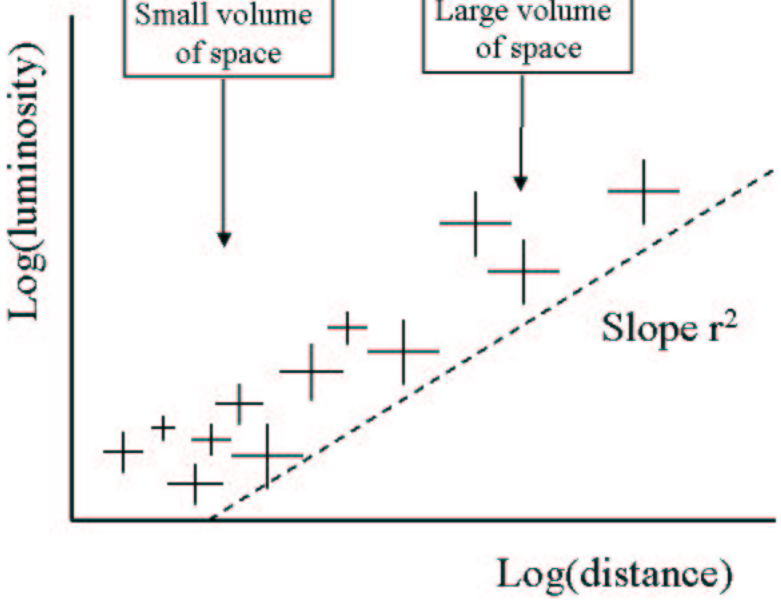


Figure 15: *The distribution of intrinsic luminosities for a sample with a range of luminosities as a function of distance. The apparent correlation is due to a selection effect.*

luminosity, so we need to sample a larger volume of space in order for them to appear in our sample. But as we look to higher distances that is precisely what we are doing: the volume of a shell of thickness $\Delta r$ in distance is $4\pi r^2 \Delta r$. So our correlation now looks like Figure 15.

This figure shows the observed redshift-luminosity relation for 3C radio galaxies selected as having radio flux densities brighter than 10 Jy.

This selection effect also rears its ugly head when correlating two luminosities (e.g. $L_R$ and $L_X$). If sample members are included only if they are brighter than some X-ray flux density and brighter than some radio flux density, then since both luminosities are artificially correlated with distance they become artificially correlated with each other: at larger distances an object has to have higher $L_R$ and higher $L_X$ to be included in the sample. This artificial correlation clearly has slope unity since the distance bias is the same for $L_X$ and $L_R$.

There are ways of dealing with this problem. Suppose $L_R$ appears correlated with $L_X$ with slope 1 and a selection effect is suspected: plot the flux densities $S_X$ and $S_R$ against each other to remove the artificial distance bias. If there is still a correlation then it is probably not that particular selection effect. Another method is to include objects in the sample even if they have upper limits to one of the luminosities, and then perform the correlation assuming the true luminosity of any object with an upper limit is in the place which most weakens the correlation.

In general other selection effects may occur in the data. Sometimes these are very subtle and are not realised for many years after publication of the results.

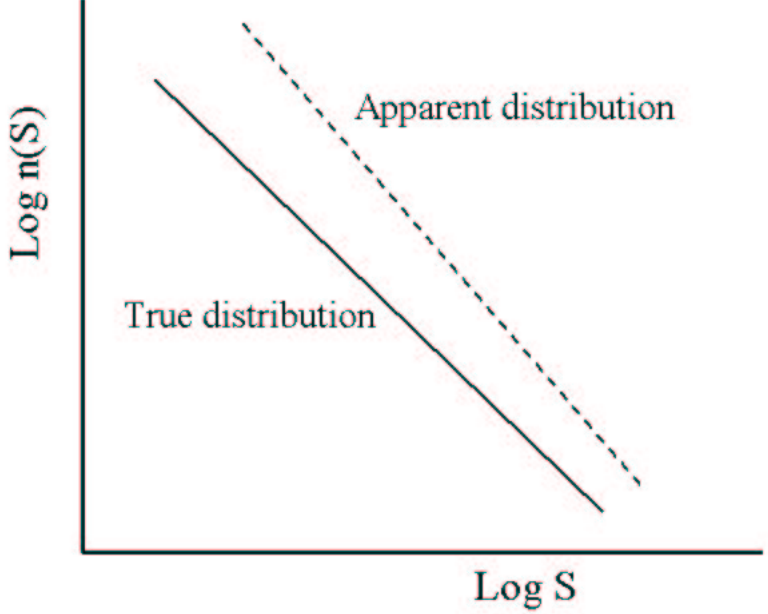


Figure 16: *The shift in the apparent distribution of flux counts of a population of sources due to measurement errors.*

### 2.6.1 Other Malmquist sampling effects

We met above the most basic form of Malmquist bias – the effect of including more luminous objects at larger distances in a sample limited by flux density. Similar biases arise in other situations, and are generally referred to as "Malmquist effects", although this general term covers a wide range of types of bias.

Let us look at the bias in the number of radio sources which are counted down to some limit of flux density, due to the presence of measurement errors. 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, the net effect is to increase the numbers counted over the true value, and to bias the measured brightnesses to brighter values.

Suppose the true (differential) number of sources per unit flux density is

$$n(S) = n_o S^{-\alpha}. \tag{151}$$

To get to the observed distribution we **convolve** this function with the error distribution, say

$$f(\Delta S) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\Delta S^2/2\sigma^2} \tag{152}$$

So the observed number counts are

$$\begin{aligned} n'(S) &= \frac{n_0}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dS'\, S'^{-\alpha} e^{-(S'-S)^2/2\sigma^2} \\ &= \frac{n_0}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} d\Delta S \left(1 + \frac{\Delta S}{S}\right)^{-\alpha} e^{-\Delta S^2/2\sigma^2}. \end{aligned} \tag{153}$$

For small $\Delta S \ll S$, we can expand the brackets in the second line

$$n'(S) = n(S) \int_{-\infty}^{\infty} d\Delta S \left(1 + \frac{\alpha \Delta S}{S} + \frac{\alpha(\alpha-1)\Delta S^2}{2S^2} + \cdots\right) \frac{e^{-\Delta S^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \tag{154}$$

The integral of the first term is unity; the second term is zero because $\langle \Delta S \rangle = 0$, and the third term, with $\langle \Delta S^2 \rangle = \sigma^2$, gives approximately the effect of Malmquist bias:

$$\Delta n = \frac{\alpha(\alpha-1)\sigma^2}{2S^2} n(S) \tag{155}$$

or

$$\Delta \ln n = \frac{\Delta n}{n} = \frac{\alpha(\alpha-1)\sigma^2}{2S^2} \tag{156}$$

This is the change in height of the $\log(n) - \log(S)$ curve due to the presence of measurement errors:

This shift is equivalent to a bias in the mean brightness of objects if they are selected as having apparent brightness $S'$;

$$\begin{aligned} \Delta \ln S' &= -\frac{\Delta \ln n(S)}{\alpha} \\ &= -\frac{(\alpha-1)\sigma^2}{2S^2} \end{aligned} \tag{157}$$

or, in magnitudes ($m = -2.5 \log_{10} S$)

$$\Delta m \approx -\frac{5 \log_{10} e \alpha_m \sigma_m^2}{2S^2} \tag{158}$$

if the measurement error in magnitudes is $\sigma_m$ and the slope of the $\log(n) - m$ relation is $\alpha_m = 0.4\alpha$.

### 2.6.2 Example: Malmquist bias in a quasar survey

A quasar survey has selected quasars with magnitudes $B < 16.2$ over 3 sr of sky, but with large measurement errors, $\sigma_m \approx 0.25$. The slope of the $\log(n) - m$ relation for quasars is very steep, due to cosmological evolution, and has a value $\alpha_m \approx 0.9$. Hence the quasars which are found in the survey are, on average, measured as being about 0.07 magnitudes brighter than they really are, and the number of quasars found is a factor about 40% too high.

## 2.7 Monte-Carlo Methods

To finish with, we shall consider an alternative approach to error analysis which has grown with the availability of fast computing. Many of the analytic approaches to probability and statistics we have developed in this course have made use of the Central Limit Theorem. This has allowed us simplify the problem to that of the study of the Gaussian distribution, which we understand well.

However, in many cases the Central Limit Theorem may not apply (we may not have enough samples to approach the limit of $n \to \infty$) or we may not believe the underlying distribution is a Gaussian, but still want to estimate confidence regions. Or it may just be that we have a very complicated system for which it is hopeless to try and propagate all of the uncertainties analytically.

To deal with such circumstances, astronomers have turned to a set of methods called **"Monte-Carlo" methods**[12]. The idea here is that we make an ensemble of computer realizations of whatever system we are studying. For each member of the ensemble, the components of the system are randomly chosen from each of their underlying distributions.

For example, say we want to model realistic errors on the variance on different scales, estimated from Cosmic Microwave Background (CMB) sky, as seen by some particular experiment. We would randomly generate a random temperature field with Gaussian temperature variations, on a pixelized sky (lets not worry how, but it is quite easy). We might then want to add foreground microwave emission due to synchrotron radiation generated in our Galaxy. There may be uncertainties on this foreground which we could model by sampling each pixel value from another distribution, maybe another Gaussian. Then we would add noise appropriate to the instrumental response in each pixel, sampling the noise from e.g. again a Gaussian with a variance equal to the variance due to noise in the instrument. We could build up a model for the observed CMB fluctuations by adding more and more layers. We can generate an ensemble of such observations by repeating this process, but selecting each effect at random from its distribution each time.

Having generated such an ensemble, it can then be used to generate estimates of the CMB variance on different scales including all of these effects without having to appeal to the Central Limit Theorem. The ensemble of estimated variances can then be used to estimate the distribution of measured variances, including all of these effects, and hence the uncertainty of a measurement.

As neat as this sounds, and it can be very useful for modelling the statistical properties of non-Gaussian and nonlinear effects, we should note some limitations of the Monte-Carlo method. The most basic is that we must make many choices about what the underlying distributions actually are. If we understand things well, this can be fine. But, as is quite often, if we don't a simple Gaussian will be chosen. We should also be aware that we will neglect all of the things we don't know about, for example systematics effects.

[12]The name "Monte-Carlo" comes from the emergence of the method during card games at Los Alamos during the making of the atomic bomb, when the mathematician John von Neumann realized such methods would help him win.