Statistical Astronomy I: A3/A4 Lectures, October 2000

COURSE CONTENT (10 lectures)

Sec. 1: Mathematical Building Blocks

1.1 *The theory of probability.* Probability as the relative frequency of outcomes. Law of Addition. Conditional Probability. Law of multiplication.

1.2 *Statistical independence.*

1.3 *Probability distributions.* Discrete random variables; Poisson distribution. Continuous random variables. Cumulative distribution function. Uniform and normal distributions.

1.4 *Expectation and other measures of a distribution.* Expected, or mean, value. Median. Mode. Variance. Skewness. Kurtosis. Variance of a function of random variable.

1.5 *Variable transformations.*

1.6 *Probability integral transform.*

1.7 *Multivariate distributions.* Joint PDF of two or more random variables. Marginal distributions. Conditional distributions. Bayes' theorem.

1.8 *Statistical independence revisited.*

1.9 *The bivariate normal distribution.*

Sec. 2: Statistical Building Blocks

2.1 *The sampling distribution.*

2.2 *Parameter estimation.* Definition of a statistic. Estimators. Bias of an estimator. Risk of an estimator. The sample mean. The law of large numbers. The central limit theorem.

2.3 *The principle of maximum likelihood.* Likelihood functions. Maximum likelihood estimators.

2.4 *The principle of least squares.* Least squares fits to straight lines. Linear regression.

2.5 *Goodness of fit.* The χ^2 distribution.

2.6 *Fitting general models.*

Sec. 3: Hypothesis Tests

3.1 *Simple hypothesis tests.* Test statistics. Null and alternative hypotheses. Critical regions.

3.2 **Incorrect decisions**. Type I and II errors. Power of a hypothesis test.

3.3 *Level of significance.*

3.4 *Two-tailed tests.*

3.5 *Goodness of fit for discrete distributions.* The binomial distribution.

3.6 *Are two distributions equal?.* The Kolmogorov-Smirnoff test.

3.7 *The student's 't' test.*

3.8 *Testing the difference of means.*

3.9 *Testing the ratio of variances.*

3.10 *Testing hypotheses on the correlation coefficient.*

Sec. 4: Point and Interval Estimation

4.1 *Defining confidence intervals.*

4.2 *Interpreting confidence intervals.*

Martin Hendry, October 2000.

Introduction

An experimental science such as astronomy involves measuring many diverse types of physical quantities (e.g. apparent magnitudes and diameters, colours, parallaxes, wavelengths). These quantities can often, in turn, be used to infer the values of other physical quantities which we *cannot* measure directly (e.g. luminosities, intrinsic sizes, temperatures, distances, velocities). This inference process generally involves making certain model assumptions. For example, to infer the distance of a star of a given spectral type, we might assume that **all** stars of that spectral type have the same absolute magnitude. We can then deduce the star's distance by combining the measured *apparent* magnitude with the assumed *absolute* magnitude.

The inference process is often made difficult because the physical quantities which we observe cannot be measured with arbitrary precision. In addition, the model assumptions which we make will only be valid to some given level of precision. For example, stars of the same spectral type will not have **exactly** the same absolute magnitude, and moreover the observed apparent magnitude of a given star will be subject, at some level, to a **measurement error**. These two effects mean that our inferred distance for the star will also be uncertain.

The aim of this course will be to develop the necessary tools to model the errors and uncertainties which arise in this inference process. The theory which describes the mathematical nature and behaviour of errors is known as **probability theory**. The branch of mathematics which describes how errors affect the process of inferring physical quantities from observational data is known as **statistics**. In Sections 1 and 2, we will study the 'building blocks' of probability and statistics respectively. In Section 3 we will then discuss in detail how we can use statistics to test our physical theories and determine, for example, which of two competing theories (or 'models') is in better agreement with a particular set of observational data. We call this process **hypothesis testing**. Finally in Section 4 we will consider a method of estimating inferred quantities which is very widely used in the astronomy and astrophysics literature (and indeed throughout the literature of almost any quantitative science); the method of estimation based on **confidence intervals**.

SECTION 1 : Mathematical Building Blocks

1.1 : Probability

The theory of probability is a branch of *pure* mathematics. This means that we could deduce laws and theorems which describe how to manipulate probabilities, starting from a set of axioms – in a manner similar to the theory of arithmetic. Many modern textbooks take this approach, since it allows the mathematical machinery of **measure theory** to be directly applied.

The disadvantage to such an approach is that it is both abstract and complicated. We will instead try to develop ideas about probability which are more **intuitive**. (c.f. how we learn to count at school).

1.1.1 : Counting (combinatorial) Definition of Probability

Suppose we observe some event (e.g. a physical experiment) for which there are a finite number, n, of possible outcomes. Suppose the outcomes can be grouped together according to some well-defined attribute or characteristic. e.g.:-

Event	example attributes
Tossing a coin	head, tail
Throwing a dice	1, 2, 3, odd number, even number

Suppose that attribute A occurs in m of the n possible outcomes. Then we could define the **probability** of an outcome having attribute A – which we write simply as $P(A)$ – as:-

$$P(A) \quad = \quad \frac{\text{number of outcomes with attribute } A}{\text{total number of outcomes}} \quad = \frac{m}{n}$$

For example, if a coin is equally likely to fall as a head or tail, we say that:-

$$P(\text{head}) \;=\; P(\text{tail}) \;=\; 1/2$$

1.1.2 : 'Frequentist' Definition of Probability

How do we 'know' from the outset that a coin is equally likely to fall as a head or a tail? In truth we do *not* know this *a priori*, but our intuition might lead us to reason as follows. Suppose we toss the coin a large number of times and, in the long run, the coin falls as a head half of the time and as a tail half of the time; we could then regard a head and a tail as equally probable outcomes. This intuitive idea of what happens to the coin when it is tossed a large number of times forms the basis of what is known as the **frequentist** definition of probability.

More generally, suppose we perform an experiment N times. (This can be something as simple as tossing a coin, or something as complex as measuring the Hubble constant). We define the **relative frequency** of an outcome with attribute A_i as:-

$$\text{rel. freq.}(A_i) \quad = \quad \frac{\text{number of outcomes with attribute } A_i}{\text{total number of outcomes}} \quad = \quad \frac{n(A_i)}{N}$$

We then define the probability of outcome A_i as

$$P(A_i) \quad = \quad \lim \frac{n(A_i)}{N} \qquad \text{as} \quad N \to \infty$$

Aside: Later in the course we will consider how to *test* e.g. whether a coin is **fair** (i.e. $P(\text{head}) = P(\text{tail}) = 1/2$) by asking how close the experimentally determined ratio, $n(\text{head})/N$, should be to $1/2$, for a given number of 'experiments', N, in order to be confident that the coin is fair. We will see that we can never be **absolutely** sure that the coin is fair, but statistics allows us to make quantitative statements about how **likely** it is that the coin is fair. Here the assumption of a fair coin ais an example of a **hypothesis**, which we can test by tossing the coin a large number of times. Based on our accumulated data, we then decide to either accept or reject the hypothesis of a fair coin.

Tossing a coin is an example of a simple event because there are only two possible outcomes and these are mutually exclusive. Generally, however, we must deal with events which are *not* simple, but rather are *composite* – i.e. combinations of two or more simple events. Consider, for example, a pack of cards. (Assume that the probability of drawing each card is 1/52). Three such composite events would be:-

- Probability of drawing an ace **or** a spade
- Probability of drawing an ace **and** a spade
- Probability of drawing an ace **then** a spade

To handle such events we need laws for combining probabilities. We will not *prove* these, but justify them by counting arguments – essentially using our well-known ideas about the intersection and union of sets.

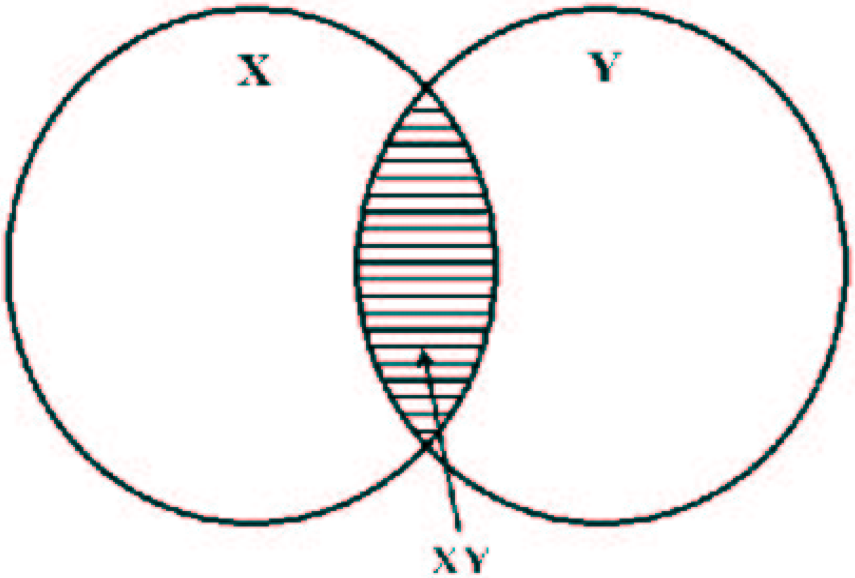
1.1.3 : Law of Addition

Let X and Y be two different sets of outcomes of an experiment. Let $X+Y$ denote the set of outcomes which occur in *either* X *or* in Y, and XY the set of outcomes which occur in *both* X *and* Y. Then

$$P(X+Y) \quad = \quad P(X) \; + \; P(Y) \; - \; P(XY)$$

We can justify this equation by counting arguments. Suppose we carry out the experiment N times. let $n(X)$, $n(Y)$, $n(X+Y)$ and $n(XY)$ denote the number of elements in the sets X, Y, $X+Y$ and XY respectively (c.f. Figure 1).

Figure 1: Venn diagram showing two intersecting sets of outcomes



Simple counting gives

$$n(X+Y) \; = \; n(X) \; + \; n(Y) \; - \; n(XY)$$

Dividing by N and letting $N \to \infty$, we obtain the law of addition. Thus, in order to determine the probability that an outcome belongs to set X *or* set Y, we add the probability that the outcome belongs to set X to the probability that it belongs to set Y. But this means that we have counted *twice* those outcomes which belongs to both X and Y, so we need to *subtract*, $P(XY)$.

Ex: $P(\text{ace } \textbf{or} \text{ spade}) \;=\; P(\text{ace}) \;+\; P(\text{spade}) \;-\; P(\text{ace } \textbf{and} \text{ spade})$
$= 4/52 + 13/52 - 1/52$
$= 16/52 = 4/13$

1.1.4 : Conditional Probability

Consider an experiment which is repeated n times – i.e. we have a total of n outcomes. Let n_1 of these outcomes have some attribute A_1, n_2 have another attribute A_2 and n_{12} have attributes A_1 *and* A_2. Then,

$$P(A_1) \;=\; \frac{n_1}{n} \quad (\text{strictly}\, P(A_1) = \lim \frac{n_1}{n}\,, \;\; \text{as } N \to \infty)$$

Also

$$P(A_2) \;=\; \frac{n_2}{n}$$

and

$$P(A_1 \,\text{and}\, A_2) \;=\; \frac{n_{12}}{n}$$

We can write this last equation as

$$P(A_1 \,\text{and}\, A_2) \;=\; \frac{n_{12}}{n_1}\frac{n_1}{n} \;=\; \frac{n_{12}}{n_1}\,P(A_1)$$

n_{12}/n_1 is the relative frequency of those outcomes which have attribute A_1, which **also** have attribute A_2.

In the limit as $n_1 \to \infty$, n_{12}/n_1 is defined as the **conditional probability** of the outcome having attribute A_2, *given* that it has attribute A_1. It is usually written as $P(A_2|A_1)$.

1.1.5 : Law of Multiplication

In the above notation

$$P(A_1 \,\text{and}\, A_2) \;=\; P(A_1\, A_2) \;=\; P(A_1)\, P(A_2|A_1) \;=\; P(A_2)\, P(A_1|A_2)$$

Thus

$$P(A_2|A_1) = \frac{P(A_1\, A_2)}{P(A_1)}$$

which is often how conditional probabilities are defined in practice.

1.2 : Statistical Independence

Let X and Y denote two sets of outcomes of an experiment. In keeping with the notation already introduced,

$$P(X) = \text{Prob(outcome } \in X)\,, \quad P(Y) = \text{Prob(outcome } \in Y)$$

$$P(X\,Y) = \text{Prob(outcome } \in X \,\text{and}\, Y)$$

We then say that X and Y are **independent** sets of outcomes if and only if

$$P(XY) \;=\; P(X)\,P(Y)$$

We can justify this result as follows. If X and Y are independent, then knowledge that the outcome belongs to X has no effect on the probability that the outcome belongs to Y. This means that

$$P(Y|X) \;=\; P(Y)$$

But since for *any* X and Y we have

$$P(XY) \;=\; P(X)P(Y|X)$$

it follows that, for X and Y independent

$$P(XY) \;=\; P(X)\,P(Y)$$

This equation defines statistical independence.

Extension to more than two sets of outcomes is straightforward. For example:-

$$P(XYZ) \;=\; P(X|YZ)\,P(YZ) \quad = P(X|YZ)\,P(Y|Z)\,P(Z)$$

If X, Y and Z are independent, then

$$P(XYZ) \;=\; P(X)\,P(Y)\,P(Z)$$

The outcomes which we have considered so far have been **qualitative** (head, tail, ace, spade etc). This is a useful means of introducing definitions of probability and independence, but we now need a description of probability which deals with **quantitative** outcomes.

1.3 : Probability Distributions

An observed event with several possible outcomes is called a *random* event. When the outcome is a numerical quantity (e.g. a physical measurement such as length, time, apparent magnitude, wavelength) it is called a **random variable** (RV).

1.3.1 : Discrete Probability Distributions

If a RV can take only a finite[1] number of values then it is a discrete RV. We can associate with each possible outcome, r, a probability, $p(r)$. The set of all $p(r)$ is called the **probability distribution** of the discrete random variable, r.

1.3.2 : Poisson Distribution

A **Poisson** RV is a discrete RV describing, e.g., the number of photons counted in a given time by a CCD. We denote the probability of counting r photons in time interval t by $p(r,t)$, although some textbooks use the notation $P_r(t)$. A Poisson RV is defined by the following three postulates.

a The probability of an event occuring in time interval, t, is independent of the past history of events prior to t

b For small interval, δt, there is an intrinsic **rate**, (i.e. number of events per unit time) $\mu\,(>0)$ such that the probability of a single event in δt, $p(1,\delta t) = \mu\delta t + o(\delta t)$.

c The probability of two or more events happening at the same time is zero, i.e. $p(r,\delta t) = o(\delta t)$, for all $r \geq 2$.

Here $o(\delta t)$ represents any function such that $o(\delta t)/\delta t \to 0$ as $\delta t \to 0$.

These postulates imply that the probability distribution function of a Poisson RV takes the form

$$p(r,t) \quad = \quad \frac{(\mu t)^r}{r!}\,e^{-\mu t}$$

[1] or countably infinite, although this mathematical subtlety need not concern us in this course

We can prove this result by *induction*; although this proof is not examinable, a short summary is provided on a handout (see website). Note that

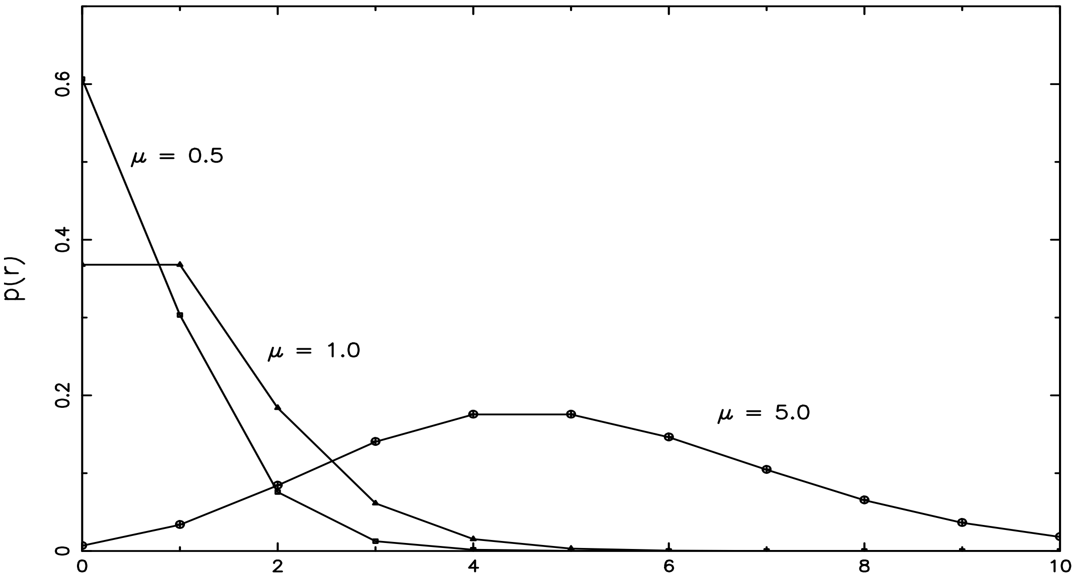
$$\sum_{r=0}^{\infty} p(r,t) \quad = \quad \sum_{r=0}^{\infty} \frac{(\mu t)^r}{r!}\, e^{-\mu t} \quad = e^{-\mu t} \sum_{r=0}^{\infty} \frac{(\mu t)^r}{r!} \quad = \quad 1$$

as required, since r must take *some* value between 0 and ∞. It is very often the case that the time interval, t, is simply taken to be be unity, in which case we can write

$$p(r) \quad = \quad \frac{\mu^r}{r!}\, e^{-\mu}$$

Figure 2 shows a plot of the Poisson distribution for several different values of μ. Note that the shape of the PDF changes significantly with increasing μ: for small values of μ the PDF is monotonic decreasing, whereas for larger values of μ it takes on more of a bell shape.

Figure 2: Poisson distribution, $p(r)$, for different values of μ



Note that we could also define the Poisson RV in space – e.g. the probability of finding r galaxies in a given volume or projected area of sky could be modelled as a spatial Poisson RV. In that case, the rate parameter, μ, would have dimensions of inverse volume, or inverse area, instead of inverse time. We will consider another common discrete probability distribution, the **binomial distribution**, later in the course.

1.3.3 : Continuous Distributions

Suppose a RV, X, can take *any* real value in a given interval – i.e. we have an uncountably infinite number of possible outcomes. We call X a **continuous RV**. Examples of continuous random variables include the apparent or absolute magnitude of stars or galaxies, distances, redshifts, orbital inclinations, etc. In fact, almost *all* quantitative physical measurements in astronomy can be regarded as continuous RVs.

Many textbooks denote a RV by a capital letter – often in bold face – and use the corresponding small letter to denote a particular *observed value*, or **realisation**, of the RV. Whenever convenient, we will adopt this notation.

What is $P(X = x)$? We have a potential paradox here. If we sum probabilities over x, we would have

$$\sum_x P(X = x) \quad = \infty \quad > 1$$

if $P(X = x) \neq 0$ for an infinite number of values of x. Of course, a probability cannot be greater than unity, far less equal to infinity!

This is simply telling us that the probability of X being *exactly* equal to any fixed value is zero. Instead we measure the probability of X lying in a small interval, $(x, x + dx)$. In the limit as $dx \to 0$, we have

$$P\left(X \in (x, x + dx)\right) \quad = \quad p(x)dx$$

Here $p(x)$ is known as the **probability density function** (PDF) but is **NOT** itself a probability. In particular, we can certainly have $p(x) > 1$, but always

$$\int_{-\infty}^{\infty} p(x)dx \quad = \quad 1$$

Thus the probability that X lies in the interval (a, b) is given by

$$P(a < X < b) \quad = \quad \int_a^b p(x)dx$$

In general we can always define a RV, X, on the entire real line, $(-\infty, \infty)$. We simply define $p(x) = 0$ outside the range of physically meaningful values of x.

1.3.4 : Cumulative Distribution Function

Consider a RV, X. The function

$$P(t) = P(X < t) = \int_{-\infty}^{t} p(x)dx$$

is called the **cumuluative distribution function** (CDF) of X. Thus the CDF measures the probability that X takes a value less than t. Note that $P(-\infty) = 0\,, \quad P(\infty) = 1$.

1.3.5 : Examples of Continuous RVs

(1) Simplest example of a continuous RV is the **uniform distribution**, usually denoted by $U(a, b)$, defined on the interval (a, b), with $a \neq b$. The uniform distribution has PDF

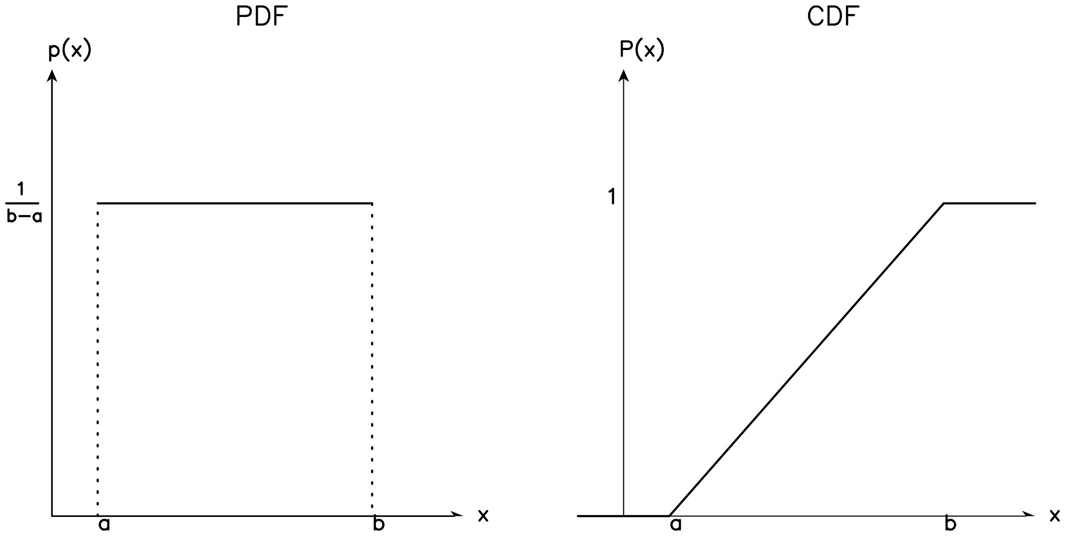
$$p(x) = \begin{cases} 1/(b-a) & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

and CDF

$$P(x) = \begin{cases} 0 & x \leq a \\ (x-a)/(b-a) & a < x < b \\ 1 & x \geq b \end{cases}$$

These functions are shown in Figure 3.

Figure 3: PDF and CDF of the uniform RV, $U(a, b)$



(2) The most important continuous RV is the normal, or Gaussian, distribution, usually denoted by $N(\mu, \sigma)$. It has PDF

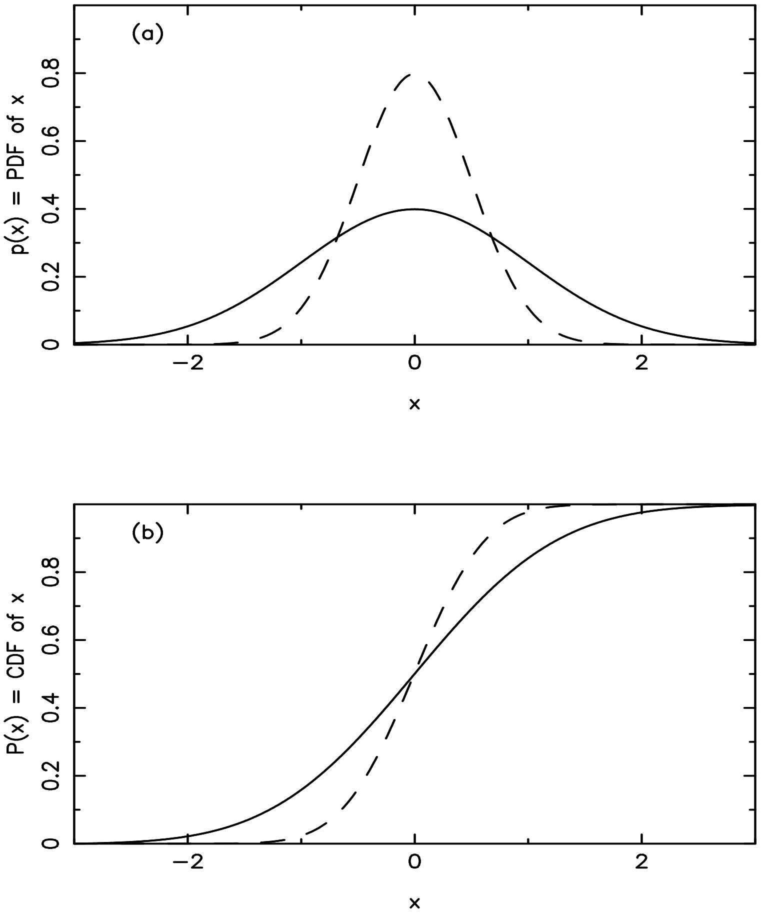
$$p(x) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$

This is a bell-shaped curve, symmetrical about $x = \mu$. The parameter σ is a measure of the width of the PDF. There is no analytic form for the cdf of the normal distribution, although it is often denoted by $\Phi(t)$. Thus

$$\Phi(t) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{t} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]dx$$

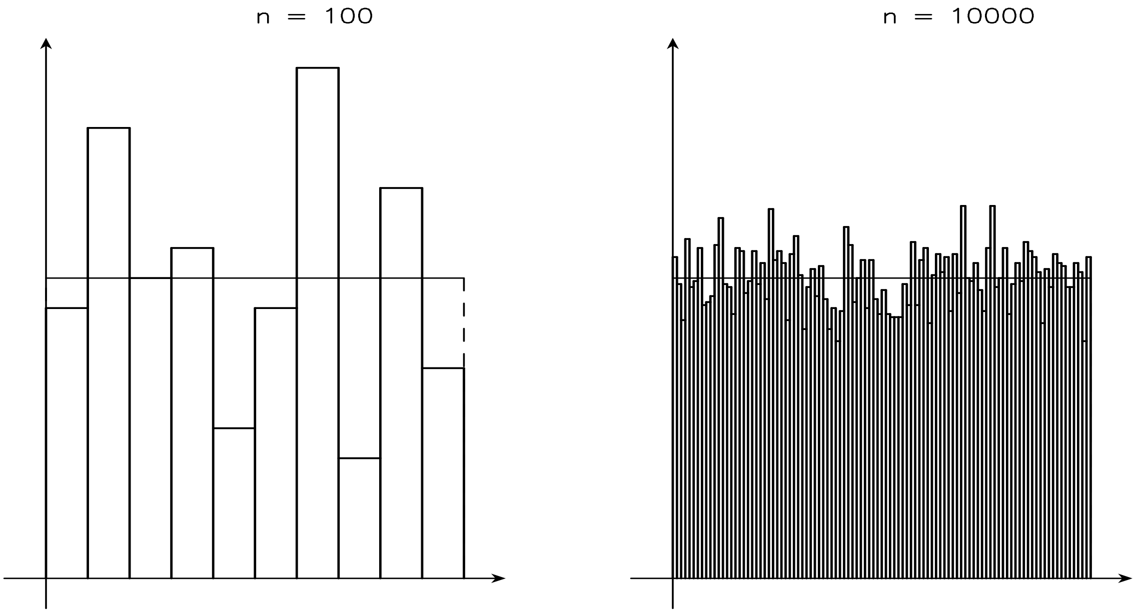
The value of $\Phi(t)$ is tabulated in many statistics textbooks, or in numerical packages. The PDF and CDF of the normal distribution for several different values of σ is shown in Figure 4. Both the normal and uniform distributions are very important for theoretical reasons, as we will see later.

Figure 4: PDF (a) and CDF (b) of the normal distribution for $\sigma = 0.5$ (dashed) and $\sigma = 1.0$ (solid)



How, in practice, do we determine the PDF of a RV? The most intuitive approach is to make use of our 'frequentist' definition of probability. Suppose we make repeated measurements of our physical quantity, i.e. we repeat our experiment a very large number of times. We then record the measured values in a **histogram**, normalised so that the total area under the histogram is equal to unity. In the limit as the number of experimental 'trials' tends to infinity (and where the width of the histogram bins tends to zero), the heights of the histogram bins (the 'relative frequency' of the different outcomes) 'traces out' the PDF of the RV. This is illustrated for the simple case of a RV uniform on the interval $(0, 1)$ in Figure 5, below. (Here the sequence of 'experiments' have been generated on computer using a random number generator program. Note that as the number of trials increases the histogram more accurately approximates the 'flat' PDF.

Figure 5: Histogram approximations to a uniform RV, with PDF $U(0, 1)$



1.4 : Expectation and Other Measures of a Distribution

1.4.1 : Expected value

The **expectation** or **expected value** of a continuous RV, X, is defined as its integral over the pdf of X. It is usually denoted by $E(X)$. Thus

$$E(X) \quad = \quad \int_{-\infty}^{\infty} x\, p(x) dx$$

Similarly, the expected value of a discrete (e.g. Poisson) RV is defined by

$$E(X) \quad = \quad \sum_{x=0}^{\infty} x\, p(x)$$

The expected value is also known as the **mean**, and is often written as $\overline{x}$, or $<x>$.

1.4.2 : Median value

The **median** of a RV, X, is the value, x_{med}, which divides the CDF into two equal halves. Thus x_{med} satisfies

$$\int_{-\infty}^{x_{\text{med}}} p(x) dx \quad = \quad 0.5$$

If the PDF is symmetric about the mean, then the mean and median are identical.

1.4.3 : Modal value

The **mode** of a RV, X, is the value of X at the maximum of the PDF. Thus x_{mode} satisfies

$$\frac{\partial p(x = x_{\text{mode}})}{\partial x} \; = \; 0$$

Obviously the mode may not be uniquely defined. For example, for $U(a, b)$, $\partial p / \partial x = 0$ for all $x \in (a, b)$.

1.4.4 : Variance

The **variance** of X is defined as (for a continuous RV)

$$\text{var}(X) \quad = \quad \int_{-\infty}^{\infty} (x - \overline{x})^2\, p(x) dx$$

with the analogous expression for a discrete RV. The variance is usually denoted by σ^2, while $\sigma = \sqrt{(\sigma^2)}$ is called the **standard deviation**.

For either continuous or discrete RVs the following equation holds

$$\mathrm{var}(X) \quad = \quad E(X^2) - [E(X)]^2$$

(The proof of this result is left as an exercise).

1.4.5 : Examples

The following table summarises the mean value and the variance of the uniform and normal distribution. (Proofs are left as an exercise; all results are quite straightforward to derive).

X	$p(x)$	$E(X)$	$\mathrm{var}(X)$
Poisson	$\frac{(\mu)^x}{x!}\, e^{-\mu}$	μ	μ
Uniform	$1/(b-a)$	$(a+b)/2$	$(b-a)^2/12$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$	μ	σ^2

The next two measures of a distribution are particular expectation values.

1.4.6 : Skewness and Kurtosis

The (normalised) **skewness** of X, is defined by

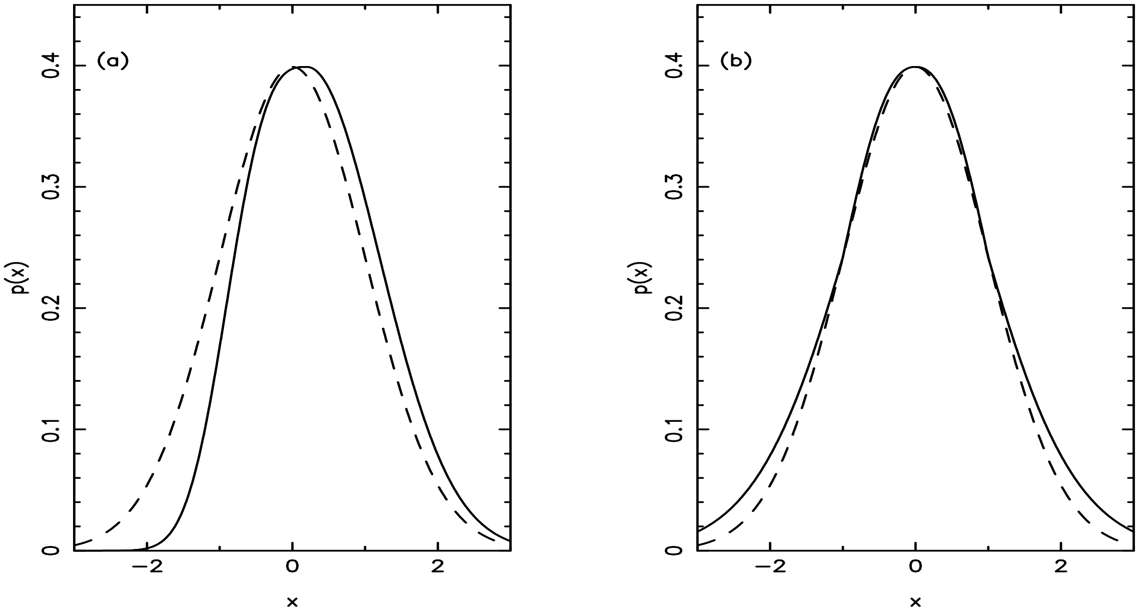
$$\mathrm{skew}(X) \quad = \quad E[(X-\overline{x})^3]/[\mathrm{var}(X)]^{\frac{3}{2}}$$

In a similar manner, the (normalised) **kurtosis** of X is defined by

$$\mathrm{kurt}(X) \quad = \quad E[(X-\overline{x})^4]/[\mathrm{var}(X)]^2 \quad - \quad 3$$

For a normally distributed RV skew(X) and kurt(X) are identically zero. The measured skewness and kurtosis of a sample of real data is often used as a test of whether those data are *drawn* from a normal distribution (see Section 2). If skew$(X) > 0$ then the PDF of X is 'positively' lopsided. If kurt$(X) > 0$ then the PDF of X has wider tails than a normally distributed RV (see Figure 6).

Figure 6: Examples of PDFs with positive skewness (a) and kurtosis (b). PDFs are shown as solid curves; Gaussian distributions are shown as dashed curves for comparison.



1.4.7 : Variance of a Function of a RV

The variance, var$[f(X)]$, of an arbitrary function of X can be approximated to second order by the following expression

$$\mathrm{var}[f(X)] \quad = \quad \mathrm{var}(X) \left(\frac{\partial f}{\partial x}\right)^2_{x=\overline{x}}$$

This expression is often used to assign an error to a function of a random variable. For example, suppose an experiment involves measuring a RV, X, but one wishes to determine the variance of $Y = X^2$. The above formula tells us that

$$\mathrm{var}[Y] \quad = \quad \mathrm{var}(X)\,[2\overline{x}]^2 \quad = \quad 4\mathrm{var}(X)\overline{x}^2$$

(For a proof of this result, see handout on webpage). To determine the *distribution* of Y we need to define a **variable transformation**.

1.5 Variable Transformations

Let the RV X have pdf $f(x)$ and let $h(x)$ denote some function of X. (e.g. if x is the colour of a star, then $h(x)$ could be the temperature). $H = h(x)$ is itself a RV with pdf $g(h)$, say. How are $f(x)$ and $g(h)$ related? We have that

$$f(x)\,dx \quad = \quad \text{prob}(x < X < x + dx)$$

and we require to find $g(h)$ such that

$$g(h)\,dh \quad = \quad \text{prob}(h < H < h + dh)$$

Suppose first that $h(x)$ is one-to-one, i.e. x maps to a unique h and vice versa. Hence the inverse function $x = x(h)$ exists, and we can write f as a function of h, i.e.

$$f(x) \equiv f(x(h))$$

This function is *not* $g(h)$, however, since we also have to transform the infinitesimal dx (just as with changing variables in integration). Thus

$$dx = \left|\frac{dx}{dh}\right| dh$$

where the modulus is required because probability is never negative. Combining the above expressions:-

$$f(x)\,dx \quad = \quad f(x(h)) \left|\frac{dx}{dh}\right| dh \quad = \quad g(h)\,dh$$

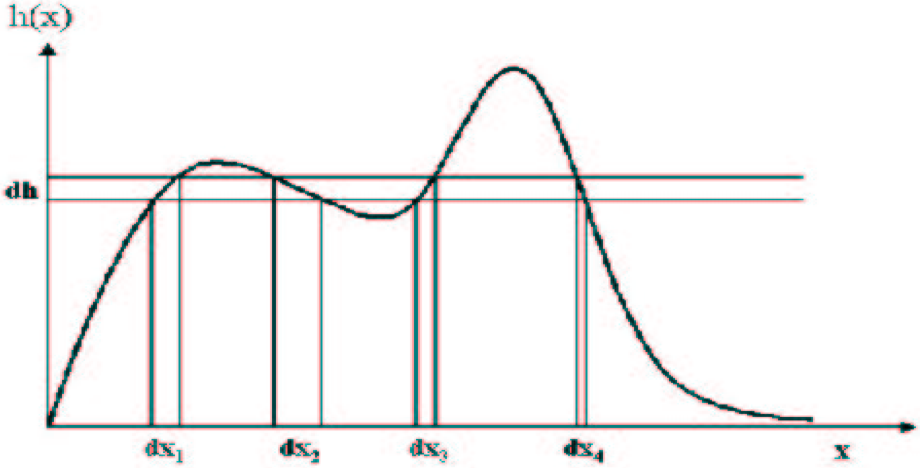
If $h(x)$ is *not* one-to-one then we must sum over all values of x for which $h(x) = h$, or more precisely over the small intervals, dx_i, corresponding to dh (see Figure 7). Thus

$$\text{prob}(h < H < h{+}dh) = \text{prob}(x_1 < x < x_1{+}dx_1){+}\text{prob}(x_2 < x < x_2{+}dx_2){+}\text{prob}(x_3 < x < x_3{+}dx_3){+}...$$

It then follows that

$$g(h)\,dh \quad = \quad \sum_{h(x_i)=h} f(x_i(h)) \left|\frac{dx}{dh}\right|_{x_i(h)} dh$$

Figure 7: Variable transformation when h is not one-to-one



1.6 : Probability Integral Transform

One variable transformation merits special consideration. Suppose X has PDF $f(x)$ and CDF $F(x)$. Define $h(x) \equiv F(x)$, which is one-to-one. Then:-

$$\begin{aligned} g(h)\,dh &= f(x(h)) \left|\frac{dx}{dh}\right| dh \\ &= f(x(h)) \left|\frac{dh}{dx}\right|^{-1} dh \end{aligned} \tag{1}$$

Since $h(x) = F(x)$, $dh/dx = f(x)$. Thus

$$g(h)\,dh \quad = \quad \frac{f(x)}{f(x)}\,dh \quad = \quad 1.dh$$

i.e. the pdf of H is the uniform distribution, $U(0,1)$, (since $0 < F(x) < 1$).

This important result shows that we can always transform the PDF of any RV into the simple form of $U(0,1)$, **provided** we know the CDF of the original RV. This approach can be used in generating random numbers numerically (see e.g. Numerical Recipes, Chap 17.).

1.7 : Multivariate Distributions

Thus far we have considered only the properties of distributions of a single (univariate) RV. We now extend to the **multivariate** case of two or more RVs.

1.7.1 : Joint PDF

The **joint PDF** of two RVs, X_1 and X_2 is $p(x_1, x_2)$. Then,

$$\text{Prob}(a_1 < X_1 < b_1 \text{ and } a_2 < X_2 < b_2) \quad = \quad \int_{a_1}^{b_1} \int_{a_2}^{b_2} p(x_1, x_2)\, dx_1 dx_2$$

Extension to more than two RVs is carried out in the obvious way.

1.7.2 : Marginal Distributions

The **marginal PDF**, $p_1(x_1)$ of X_1 is defined by

$$p_1(x_1) \quad = \quad \int_{-\infty}^{\infty} p(x_1, x_2)\, dx_2$$

and is a PDF in the usual sense that

1. $p_1(x_1) \geq 0$, for all x_1
2. $\text{Prob}(a < X_1 < b) \quad = \quad \int_a^b p_1(x_1) dx_1$
3. $\int_{-\infty}^{\infty} p_1(x_1) dx_1 \quad = \quad 1$

Similarly, the marginal PDF of X_2 is

$$p_2(x_2) \quad = \quad \int_{-\infty}^{\infty} p(x_1, x_2)\, dx_1$$

In general, given any multivariate PDF, we may find the marginal PDF of any subset of the $X_1, ..., X_n$ by integrating over all other variables. e.g.

$$p_{13}(x_1, x_3) \quad = \quad \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} p(x_1, ..., x_n)\, dx_2 dx_4 dx_5 ... dx_n$$

1.7.3 : Conditional Distributions

Consider the joint PDF, $p(x_1, x_2)$, of X_1 and X_2. Suppose we observe X_1 to have the value x_1, but do not observe X_2. We want a function that describes the PDF of X_2, given that $X_1 = x_1$ (usually simply stated as 'given x_1'). This function is known as the **conditional** PDF of X_2, written as $p(x_2|x_1)$, and defined by

$$p(x_2|x_1) = \frac{p(x_1, x_2)}{p_1(x_1)}$$

i.e. the conditional PDF is obtained by dividing the joint PDF of X_1 and X_2 by the marginal PDF of x_1 (provided $p_1(x_1) \neq 0$). Similarly

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p_2(x_2)}$$

Note that we can write

$$\begin{aligned} p(x_1, x_2) &= p(x_1|x_2)p_2(x_2) \\ &= p(x_2|x_1)p_1(x_1) \end{aligned}$$

This is known as **Bayes' formula**.

Extension to more than 2 RVs is again straightforward. For example,

$$p(x_1, x_3|x_2, x_4) = \frac{p(x_1, x_2, x_3, x_4)}{p_{24}(x_2, x_4)}$$

1.8 : Statistical Independence

If the conditional PDF of X_2 given x_1 does *not* depend on x_1, this means that X_1 and X_2 are statistically independent, since the observed value of X_2 is unaffected by the observed value of X_1.

Equivalently, X_1 and X_2 are independent if and only if the joint PDF of X_1 and X_2 can be written as the product of their marginal PDFs, i.e.

$$p(x_1, x_2) = p_1(x_1)\, p_2(x_2)$$

Again, we extend in the obvious way. The RVs $X_1, ..., X_n$ are **mutually independent** if and only if their joint PDF can be written as the product of their marginal pdfs. i.e.

$$p(x_1, x_2, ..., x_n) = p_1(x_1)p_2(x_2)...p_n(x_n)$$

1.9 : The Bivariate Normal Distribution

Let X and Y be RVs with the following joint PDF

$$p(x,y) \quad = \quad \frac{1}{2\pi\sigma_{\rm x}\sigma_{\rm y}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}Q(x,y)\right]$$

where the quadratic form, $Q(x,y)$ is given by

$$Q(x,y) \quad = \quad (\frac{x-\mu_{\rm x}}{\sigma_{\rm x}})^2 - 2\rho(\frac{x-\mu_{\rm x}}{\sigma_{\rm x}})(\frac{y-\mu_{\rm y}}{\sigma_{\rm y}}) + (\frac{y-\mu_{\rm y}}{\sigma_{\rm y}})^2$$

Then $p(x,y)$ is known as the **bivariate normal PDF** and is specified by the 5 parameters $\mu_{\rm x}$, $\mu_{\rm y}$, $\sigma_{\rm x}$, $\sigma_{\rm y}$ and ρ. This PDF is used often in astronomy to model the joint PDF of two random variables. Figure 8, for example shows the joint distribution of apparent magnitude and the logarithm of the 21cm line width (denoted by P), which is often modelled as a bivariate normal PDF in statistical studies of the Tully-Fisher distance relation for spiral galaxies.

The first 4 parameters of the bivariate normal PDF are, in fact, equal to the following expectation values:-

1. $E(X) = \mu_{\rm x}$
2. $E(Y) = \mu_{\rm y}$
3. $\text{var}(X) = \sigma_{\rm x}^2$
4. $\text{var}(Y) = \sigma_{\rm y}^2$

The parameter ρ is known as the **correlation coefficient** and satisfies

$$E[(X-\mu_{\rm x})(Y-\mu_{\rm y})] \quad = \quad \rho\sigma_{\rm x}\sigma_{\rm y}$$

Note that if $\rho = 0$ then X and Y are statistically independent.

$E[(X-\mu_{\rm x})(Y-\mu_{\rm y})]$ is known as the **covariance** of X and Y and is often denoted by $\text{cov}(X,Y)$.

The marginal PDFs of X and Y are just the univariate normal PDFs, i.e.

$$p_x(x) = N(\mu_{\rm x},\sigma_{\rm x}) \qquad p_y(y) = N(\mu_{\rm y},\sigma_{\rm y})$$

The conditional PDF of Y given x is also a univariate normal PDF, viz:-

$$p(y|x) \quad = \quad N(\mu_{\rm y} + \frac{\sigma_{\rm y}}{\sigma_{\rm x}}\rho(x-\mu_{\rm x}), \sigma_{\rm y}\sqrt{1-\rho^2})$$

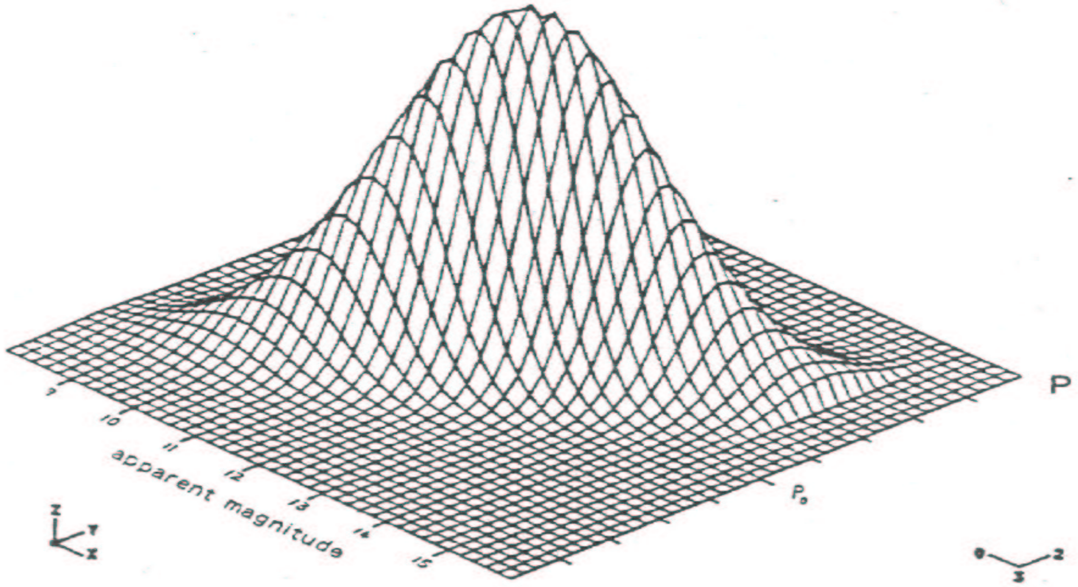
with the corresponding expression for $p(x|y)$.

$\mu_{\rm y} + \frac{\sigma_{\rm y}}{\sigma_{\rm x}}\rho(x-\mu_{\rm x})$ is often referred to as the **conditional expectation** (value) of Y given x, and the equation

$$y = \mu_{\rm y} + \tfrac{\sigma_{\rm y}}{\sigma_{\rm x}}\rho(x-\mu_{\rm x})$$

is called the **regression line** of Y on X. We will say more about regression in Section 2.

Figure 8: The bivariate normal PDF



The bivariate normal PDF (and indeed any bivariate distribution function) can be represented as a **plot of isoprobability contours**. These contour curves, closely analogous to the contours on an OS map, denote those points on the (x, y) plane where the function $p(x, y)$ is constant. Examples of isoprobability contours for bivariate normal pdf's with different values of ρ, and the corresponding regression lines of Y on X for these pdf's, are shown on the handout provided on the website.

SECTION 2 : Statistical Building Blocks

In Section 1 we considered various mathematical aspects of probability theory. We now apply some of those mathematical tools to study the *statistics* of real data samples.

2.1 : The Sampling Distribution

Consider a RV, X, with pdf $p(x)$. Suppose we observe n different *realisations* (values) of X. We call the set $\{X_1, ..., X_n\}$ a **random sample from the population with pdf** $p(x)$. The joint pdf, $g(x_1, ..., x_n)$, is known as the **sampling distribution** of the random sample. We can think of this joint pdf in terms of the 'histogram' picture which we discussed in Section 1; i.e. if we were to repeatedly random sample sets of n numbers from the pdf, $p(x)$, and construct an n-dimensional histogram of the sampled values, then in the limit as the number of samples tends to infinity the 'shape' of the histogram will approximate the sampling distribution, $g(x_1, ..., x_n)$. In this course we will consider only random samples in which all the elements are **independently and identically distributed** (usually written as iid). This means that the sampled value of $X = x_1$ is independent of $X = x_2$ and so on. In other words, the elements of the random sample are statistically independent of each other. It then follows that

$$g(x_1, ..., x_n) \quad = \quad p(x_1)p(x_2)...p(x_n)$$

i.e. the joint pdf of the random sample is product of the individual pdfs.

2.2 : Parameter Estimation

Suppose we wish to study a population which is known (or assumed) to have a pdf, $p(x;\theta)$. This notation indicates that the pdf is dependent upon a (possibly unknown) parameter, θ. If we observe a random sample from the population, $\{X_1, ..., X_n\}$ say, how can we estimate the parameter, θ? How do we decide how 'good' our estimate of θ is (or even what we mean by this question?).

2.2.1 : Statistics

A **statistic** is a function of observable random variables which does **not** depend upon any unknown parameters. Thus if we have a random sample, $\{X_1, ..., X_n\}$, from the population with pdf $p(x;\theta)$ then any function of $\{X_1, ..., X_n\}$ which does **not** depend on θ is an example of a statistic.

Suppose, for example, that $X \sim N(\mu, \sigma)$, where μ and σ are not known *a priori*. Then $X - \mu$ is **not** a statistic, since it depends on the value of the parameter, μ. The key idea in parameter estimation is to use statistics to estimate the unknown parameters of a pdf.

2.2.2 : Estimators

A statistic with which we estimate the value of a parameter is known as an **estimator** of that parameter. Estimators are usually denoted by a caret, or 'hat', e.g. $\hat{\theta}$ is an estimator of θ.

Note that $\hat{\theta}$ is **not** a function of θ (if it depended on the value of θ then it would be redundant as an estimator of θ!). Note that $\hat{\theta}$ is, however, a RV since it is a function of the RVs $\{X_1, ..., X_n\}$. Hence we can (in principle, at least) determine the pdf of $\hat{\theta}$ in terms of the sampling distribution, $g(x_1, ..., x_n)$. This means that the pdf of $\hat{\theta}$ depends upon the **true** value of the parameter, θ. We can therefore write the pdf of $\hat{\theta}$ as $p(\hat{\theta}; \theta)$, and we can use the properties of $p(\hat{\theta}; \theta)$ to decide whether $\hat{\theta}$ is a 'good' estimator.

Consider the following illustrative example. (We take an example from cosmology, although similar examples from any other branch of astronomy could be presented, since it is not the astronomical details but the statistical details which are important here).

Suppose we are measuring the redshift of a nearby galaxy in (say) the Virgo cluster. We do this, of course, by identifying features in the spectrum of the galaxy and comparing their wavelengths with the laboratory values. Thus, if we denote the *true* redshift of the galaxy by z_0, then an estimator of z_0, denoted by $\hat{z}$, will be a function of the observed wavelengths of the (n) identifying spectral features, i.e.

$$\hat{z} \quad = \quad \hat{z}(\lambda_1, ..., \lambda_n)$$

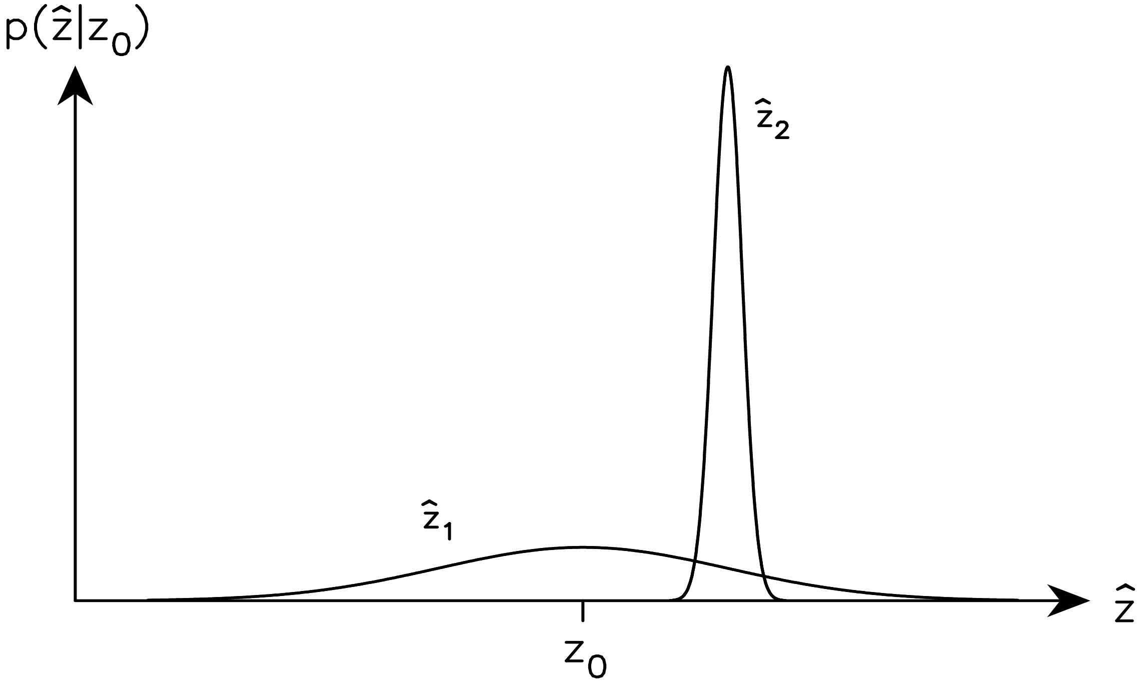
Since the sampling distribution of $\lambda_1, ..., \lambda_n$ depends on z_0, the pdf of $\hat{z}$ also depends on z_0, i.e. $p(\hat{z}) = p(\hat{z}; z_0)$.

We could measure the redshift using, e.g., a 1m-class ground based telescope with a low-resolution spectrograph, but with these data our determination of the redshift will be somewhat inaccurate (since our measured wavelengths of the identifying spectral features will be imprecise). Thus, if we were to repeat our observations with such a telescope a large number of times, a histogram of our estimated redshifts would tend in shape towards the pdf of $\hat{z}_1$, shown in Figure 9. In simple terms, we would say that our observation carried a large **statistical error** but small **systematic error**.

Suppose now we observe the same galaxy with e.g. the high-resolution spectrograph on

HST, and denote by $\hat{z}_2$ our HST estimator of z_0. With HST our wavelength measurements of the galaxy's spectral features will now be much more accurate, leading to a much narrower range of values (i.e. *realisations*) of $\hat{z}_2$, if we were to repeat our HST observations a large number of times. Suppose, however, that – for some reason – we **mis-identify** the features in the galaxy's spectrum, leading to a completely erroneous value of $\hat{z}_2$ in each of these realisations (although, of course, we would only know this if we knew the *true* value of z_0). In this case the pdf of $\hat{z}_2$ would be as shown in Figure 9, and in simple terms we would say that our observation carried a small **statistical error** but a large **systematic error**.

Figure 9: PDF of two estimators of the true redshift, z_0, of a galaxy.



We see from Figure 9 that $p(\hat{z}_1; z_0)$ is much broader than $p(\hat{z}_2; z_0)$, so there is a higher probability that $\hat{z}_1$ will differ considerably from z_0 than for $\hat{z}_2$. However, there **is**, at least, a non-negligible probability that $\hat{z}_1$ lies very close to z_0, whereas the 'narrowness' of $p(\hat{z}_2; z_0)$ means that $\hat{z}_2$ will almost always **systematically underestimate** the true redshift. These two extremes illustrate the essential difficulty in defining one single criterion which determines which estimator is 'best' in a given situation. If one wishes specifically to exclude large statistical errors, but is prepared to tolerate a small systematic 'offset' in the estimator of the parameter (particularly if it is possible to determine the size of that offset, perhaps from independent data, and thus correct for it), then $\hat{z}_2$ would be the better choice. If, on the other hand, even a small systematic error is unacceptable, then $\hat{z}_2$ would have to be regarded as a 'bad' estimator. In this example we have used the 'closeness' of $\hat{z}_1$ and $\hat{z}_2$ to z_0 as a measure of which estimator is better. We can formalise this idea of 'closeness' of an estimator to the true value of the parameter as follows:-

2.2.3 : Bias of an estimator

We define the **bias**, $B(\hat{\theta};\theta_0)$, of an estimator, $\hat{\theta}$, by

$$\begin{aligned} B(\hat{\theta};\theta_0) &= E(\hat{\theta};\theta_0) - \theta_0 \\ &= \int (\hat{\theta} - \theta_0)\, p(\hat{\theta};\theta)\, d\hat{\theta} \end{aligned}$$

where θ_0 is the true value of the parameter θ. Hence, when an estimator is **unbiased** its expected value is equal to the true value of the parameter.

2.2.4 : Risk of an estimator

We define the **risk**, $R(\hat{\theta};\theta_0)$, of an estimator, $\hat{\theta}$, by

$$\begin{aligned} R(\hat{\theta};\theta_0) &= E\left[(\hat{\theta} - \theta_0)^2;\theta_0)\right] \\ &= \int (\hat{\theta} - \theta_0)^2\, p(\hat{\theta};\theta)\, d\hat{\theta} \end{aligned}$$

The risk of an estimator is also known as the **mean squared error**. Note that when an estimator is unbiased then the risk is identically equal to the **variance** of the estimator.

In the example of Figure 9, $\hat{z}_1$ is an unbiased estimator with a large risk (and variance), whereas $\hat{z}_2$ is negatively biased, but has smaller risk (and very small variance).

Note that the bias of $\hat{z}_2$ is itself a function of z_0. This fact indicates two fundamental difficulties:-

- If we apply a correction to remove the bias of $\hat{z}_2$ at z_0 it does not follow in general that this correction will leave $\hat{z}_2$ unbiased *for all* true values of z; indeed the correction may *increase* the bias of the estimator for other true redshifts.

- In any case, to completely remove the bias of $\hat{z}$ at z_0 **strictly speaking** we need to already know the value of z_0 – if we knew that, then we would have no need to estimate the parameter!

Fortunately, in practice one can frequently define estimators which are unbiased for a wide range of, or indeed all, values of the unknown parameter, so that in particular we don't need to know the true value of the unknown parameter to know that its estimator is unbiased. The simplest example of such an estimator is the **sample mean**.

2.2.5 : The sample mean

Let $\{X_1, ..., X_n\}$ denote a random sample drawn from a population with pdf $p(x)$, mean value μ and finite variance σ^2. We define the **sample mean** as

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

Clearly $\hat{\mu}$ is an estimator. If each X_i is independently and identically distributed (iid), then $\hat{\mu}$ is an **unbiased** estimator of μ, for all values of μ. (For proof of this result see handout and lectures).

The **variance**, $\sigma^2_{\hat{\mu}}$, of the sample mean is given by

$$\sigma^2_{\hat{\mu}} = \sigma^2 / n$$

(For proof of this result see the handout: this proof is not examinable)

This result is extremely important in statistics, since it implies that, whatever the underlying population (provided it has finite variance) the distribution of the sample mean becomes increasingly concentrated near the population mean as the sample size increases. Thus, the larger the sample, the more sure we can be that $\hat{\mu}$ is a good estimator of μ. This idea is formalised quantitatively in the **law of large numbers**.

2.2.6 : The law of large numbers

Let $p(x; \mu, \sigma^2)$ be the pdf of a RV, X, with mean, μ, and finite variance, σ^2. Let $\hat{\mu}$ be the sample mean of a random sample of size n drawn from $p(x; \mu, \sigma^2)$. Let ϵ and δ be two specified small numbers such that $\epsilon > 0$ and $0 < \delta < 1$. If n is any integer such that $n > \sigma^2 / \epsilon^2 \delta$, then

$$\text{Prob}[\, |\hat{\mu} - \mu| < \epsilon \,] \quad \geq \quad 1 - \delta$$

Thus, we can make the probability that $\hat{\mu}$ lies within ϵ of μ arbitrarily close to unity, simply by taking a large enough sample of data.The proof of this theorem is, again, non-examinable, but is provided on a handout for completeness.

What is striking about the law of large numbers is the fact that we made no assumptions about the form of the pdf of X (apart from its finite variance), and yet we can *still* make precise statements about the probable 'closeness' of $\hat{\mu}$ and μ for a given sample size.

In fact, we can go much further than this in determining the properties of the sample mean, by using one of the most important theorems in statistics.

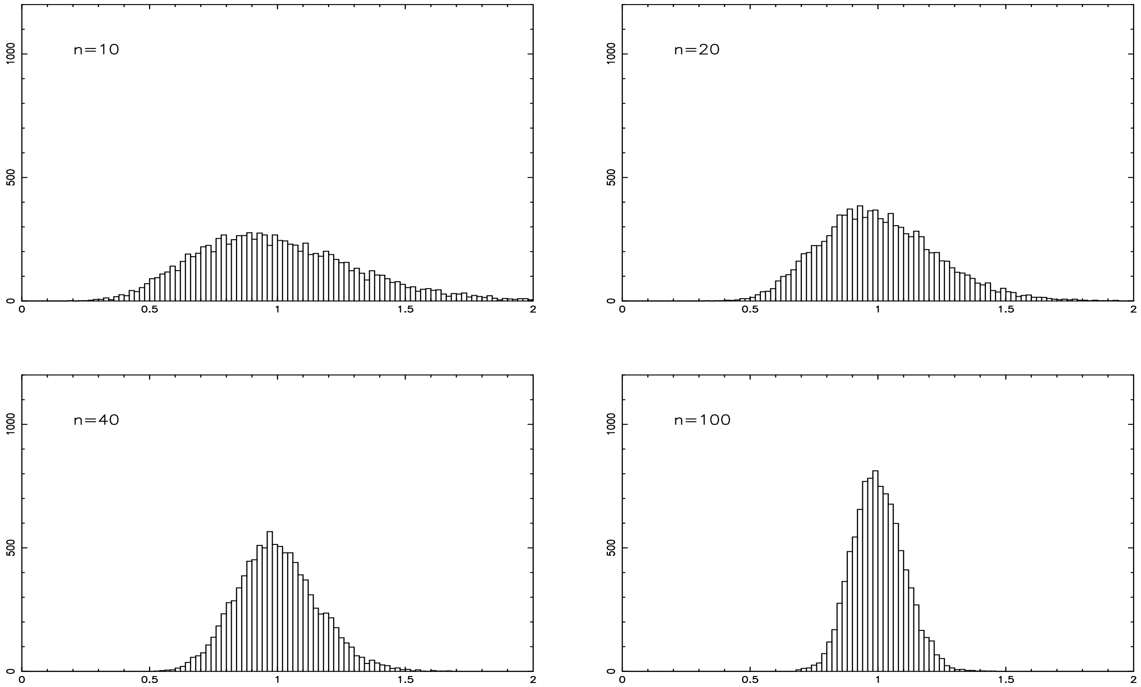
2.2.7 : The Central Limit Theorem

Let $p(x; \mu, \sigma^2)$ denote a pdf with mean μ and finite variance σ^2. Let $\hat{\mu}$ denote the sample mean of the iid random sample $\{X_1, ..., X_n\}$, of size n. Then as $n \to \infty$, the pdf of $\hat{\mu}$ approaches a normal pdf with mean value μ and variance σ^2/n.

We will not prove the central limit theorem in this course. We do, however, highlight its importance. The CLT states that, no matter what pdf our random sample is drawn from, the sample mean will have an approximately normal distribution as the sample size increases. The CLT justifies the importance of the normal distribution – in applied statistics in general, and in astronomy in particular. Astronomy is filled with situations where one 'bins' or groups sets of observational data. The CLT tells us that, when we bin data with a sufficiently large sample, the fluctuations in the average of the binned data will look approximately normally distributed. Figure 10 illustrates this, for random samples drawn from an exponential distribution – i.e. the underlying pdf is very different from a normal pdf, and yet the distribution of the sample mean very closely approximates a normal pdf as the sample size increases.

The sample mean is, thus, defined according to an intuitively simple expression, is unbiased, and has very special asymptotic properties which are almost independent of the pdf of the underlying population. This is rarely the case with other parameters of a pdf, however, and we require in statistics more general methods for finding estimators – methods which take account of the form of the underlying pdf from which our sample data are drawn.

Figure 10: Histograms of the sample mean of sample of size n drawn from an exponential distribution, for $n = 10$, $n = 20$, $n = 40$ and $n = 100$. Note the increasingly close approximation to a normal distribution.



2.3 Principle of Maximum Likelihood

Suppose we have a random sample, $\{X_1, ..., X_n\}$, drawn from the population with pdf $p(x;\theta)$. We define the **likelihood function**, $L(\theta)$, as the sampling distribution, $g(x_1, ..., x_n;\theta)$, of $\{X_1, ..., X_n\}$, but now considered as a function of θ. In other words we are now thinking of θ not as a fixed parameter, but as a *variable*. Thus,

$$L(\theta) \quad = \quad g(x_1, ..., x_n;\theta)$$

The principle of maximum likelihood essentially states that forming the likelihood function is a useful way to define a 'good' estimator of the parameter θ. The **maximum likelihood** estimator of θ, denoted by $\hat{\theta}_{\mathrm{ML}}$, is the value of θ which maximises $L(\theta)$. Thus, $\hat{\theta}_{\mathrm{ML}}$ satisfies

$$\frac{\partial L}{\partial \theta} = 0 \qquad \text{when} \quad \theta = \hat{\theta}_{\mathrm{ML}}$$

We can think of this definition in the following way. Suppose the particular values observed in our random sample are $\{x_1, ..., x_n\}$. If we were to vary the parameter, θ, we would generate a family of different pdfs. $\hat{\theta}_{\mathrm{ML}}$ is the value of θ corresponding to the pdf from which it is 'most likely' that the actual sample was drawn.

Note that if the $\{X_i\}$ are iid, then

$$L(\theta) \quad = \quad p(x_1;\theta)\,p(x_2;\theta)\,...\,p(x_n;\theta)$$

We extend to the case where the pdf is a function of several unknown parameters in the obvious way

$$\frac{\partial L(\theta_1, ..., \theta_k)}{\partial \theta_j} = 0 \qquad \text{when} \quad \theta_j = \hat{\theta}_j \quad (j = 1, ..., k)$$

For an iid random sample, $\{X_1, ..., X_n\}$, from a normal pdf, the maximum likelihood estimators of the mean, μ, and variance, σ^2, are

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}[x_i - \hat{\mu}_{\mathrm{ML}}]^2$$

i.e. simply the sample mean and variance. These results are derived on the accompanying handout and in the lectures. We already know from the preceding section that the sample mean is an unbiased estimator. What about $\hat{\sigma}^2_{\mathrm{ML}}$? After a great deal of rather tedious (and non-examinable! But see the handout anyway if you want to follow the details of the derivation) algebra, we can show that

$$E[\hat{\sigma}^2_{\mathrm{ML}}] = \frac{n-1}{n}\sigma^2$$

i.e. the sample variance is a **biased** estimator of σ^2. In fact, for *any* pdf with finite variance, we have:-

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2\right]=\frac{n-1}{n}\sigma^2$$

but we can easily define an unbiased estimator of σ^2 by multiplying the sample variance by $n/(n-1)$, i.e.

$$\hat{\sigma}^2_{\text{corr}}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu}_{\text{ML}})^2 \quad = \quad \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 \;-\; n\hat{\mu}^2_{\text{ML}}\right]$$

satisfies

$$E\left[\hat{\sigma}^2_{\text{corr}}\right] \quad = \quad \sigma^2$$

Why is $\hat{\sigma}^2_{\text{corr}}$ biased? If the mean, μ, were known *a priori*, then one can show that

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right]=\sigma^2$$

i.e. in this case the sample variance *is* unbiased. It is because in practice we also have to estimate μ that principle of maximum likelihood gives a biased estimator of σ^2.

2.4 : Least Squares Estimators

We now turn to another useful method for estimating parameters – the principle of least squares – which is particularly useful in astronomy where we often try to fit a simple functional relationship between two or more sets of observational data. To fix our ideas we will develop the theory of least squares in the context of a specific astronomical example: the period-luminosity (PL) relation for Cepheid variables.

2.4.1 : Preamble – The Cepheid PL relation

Cepheids are highly luminous pulsating stars whose pulsation period has been found to be related to their luminosity by a power law, i.e.

$$L \quad = \quad A\,P^b$$

where A and b are constants. The relation is usually considered in terms of magnitudes, i.e.

$$M \quad = \quad a \,+\, b\log P$$

The usefulness of Cepheids derives from the fact that their periods can be measured directly, thus allowing us to infer their absolute magnitude, and hence their distance via the familiar equation:-

$$m \quad = \quad M \,+\, 5\log r \,+\, 25$$

It is the step of inferring the absolute magnitude from the measured period which concerns us here, and which requires the application of statistical techniques. This is because, in practice, any group of Cepheids will not satisfy exactly the above linear relation between M and $\log P$. If we plotted the 'observed' values of $\{M_i, \log P_i\}$ for a sample of Cepheids we would expect the points to be scattered on the plane, due to a combination of observational errors in the measurement of M_i (which is, in any case, not measured directly but would itself have to be inferred from the measured *apparent* magnitude of the Cepheid combined with some independent estimate of its distance) and $\log P_i$, and intrinsic errors due to the inadequacy of the linear relation which we are assuming holds between these two quantities (recall the discussion in the introduction). Figure 11 shows the Cepheid PL relations derived for calibrating data in the LMC and SMC at a series of wavelengths from B to K. (In fact these plots show the apparent magnitude of the Cepheids, which **is** directly observed, but since the LMC and SMC Cepheids can all be assumed equidistant these apparent magnitudes are equivalent to absolute magnitudes, as is easily seen from the distance modulus formula above.)

As can be seen, these data clearly display a linear relationship but there is indeed a non-negligible scatter in the relation, so that – at a given period, there is a range, or distribution, of absolute magnitudes consistent with that period. But in order to use the PL relation to estimate the distance of a more remote Cepheid, we want to assign a **single** value of M to the star. In other words we want to fit a straight line (or more generally, a curve) through the $\{M_i, \log P_i\}$ scatterplot so that we have a one-to-one relationship between the observed (log) period and the inferred absolute magnitude.

We want this straight line to be the one which, in some sense, is the 'best fit' to the data – i.e. we want the observed data points (which we refer to as our 'calibrating data') to lie 'closest' to the best fit line. The principle of least squares provides us with a definition of what we mean by 'closest' in this context. We also want a means of quantifying whether the scatter of the data about this best fit straight line (what we call the **residuals** of the best fit) is consistent with our assumption of a straight line model in the first place. If a plot of our PL calibrating data looked like Figure 12, for example, then common sense would tell us that a straight line model was inappropriate. Statistics provides us with a means of quantifying this degree of 'inappropriateness' – what we call the **goodness of fit**.

Figure 11: PL relations for Cepheids in the LMC and SMC.

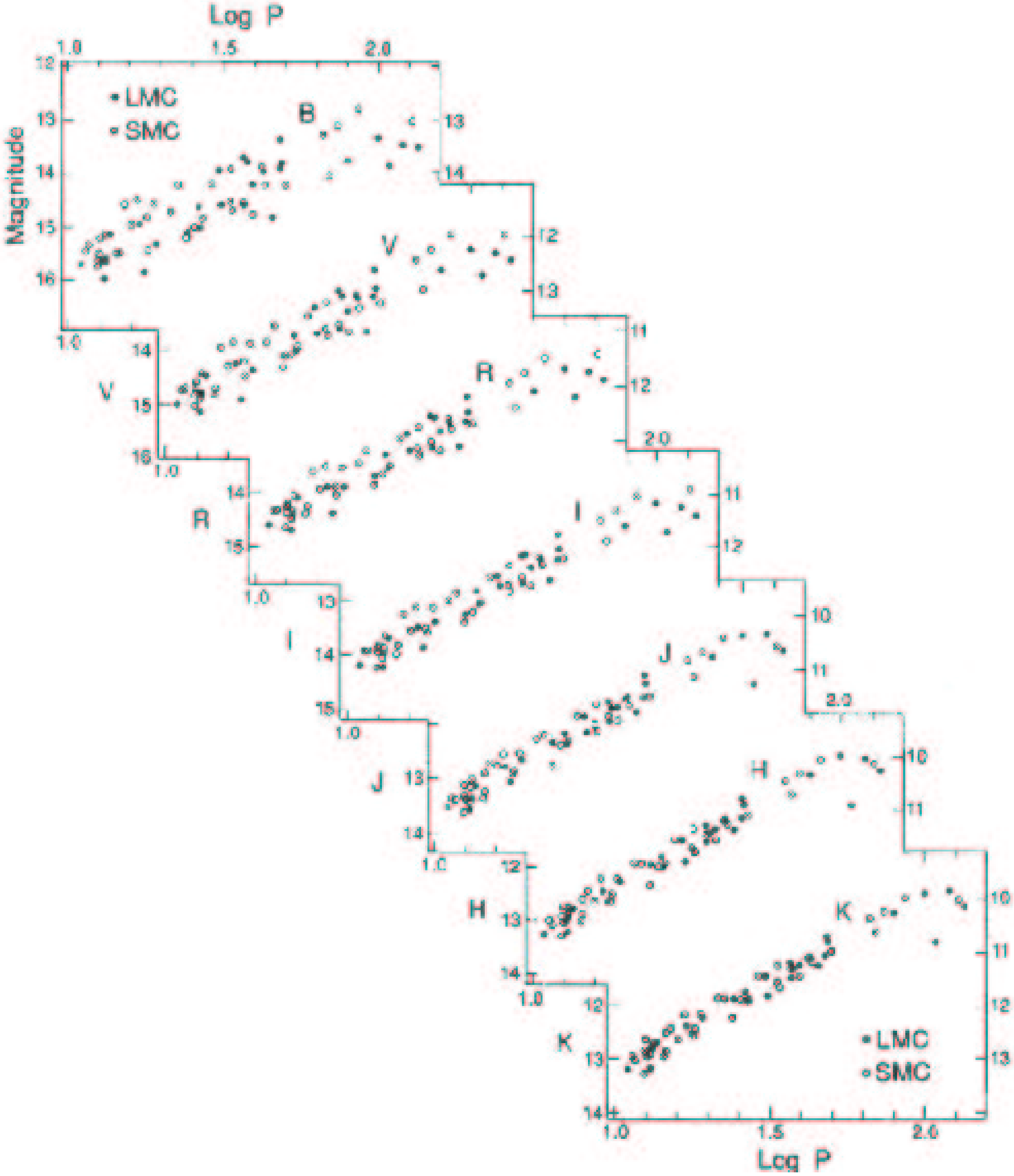
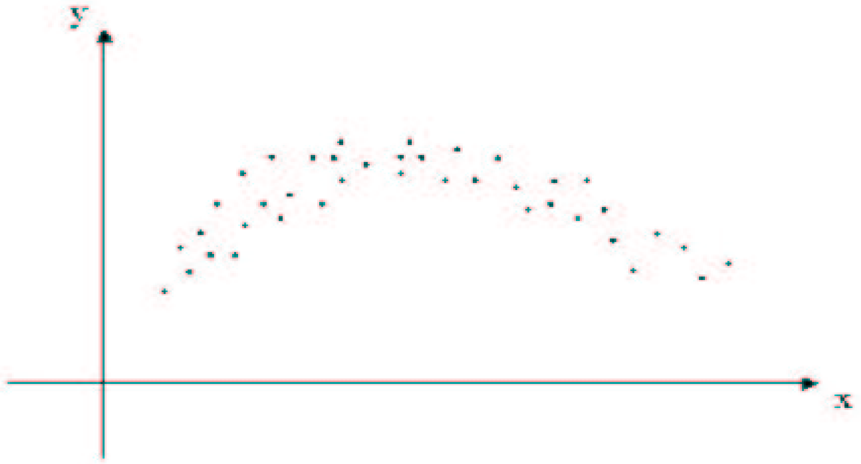


Figure 12: Data for which a straight line model is not appropriate.



2.4.2 : Ordinary Linear Least Squares

Suppose that the scatter in our plot of $\{M_i, \log P_i\}$ is assumed to arise from errors in only one of the two variables. This case is called **Ordinary Least Squares**. In the context of the PL relation, it is probably reasonable to assume that there is no error on the measured period of a Cepheid, or at least that this error is very small compared with the uncertainty on the absolute magnitude. We then call log period the **independent variable**, and absolute magnitude the **dependent variable**. Thus we suppose that we can write, for each Cepheid:-

$$M_i \quad = \quad a + b \log P_i + \epsilon_i$$

where ϵ_i is known as the **residual** of the i^{th} Cepheid – i.e. the difference between the observed value of M_i, and the value predicted by the best-fit straight line (see Figure 13).

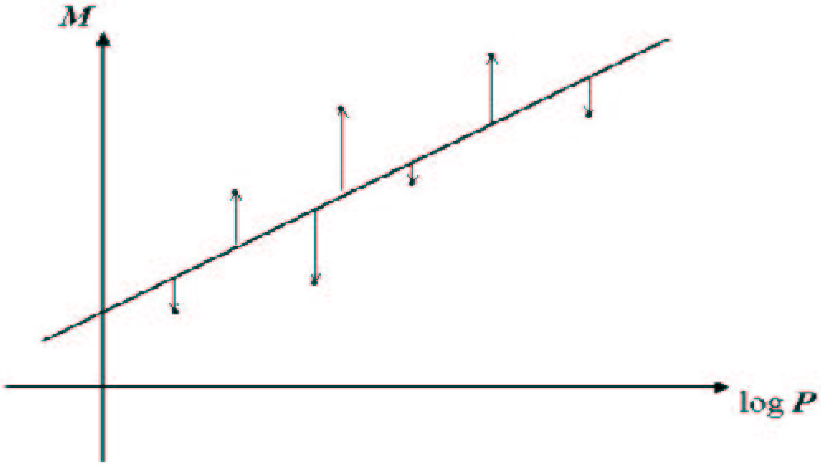
We assume that the $\{\epsilon_i\}$ are an iid random sample from some underlying pdf with mean zero and variance σ^2 – i.e. the residuals are equally likely to be positive or negative and all have equal variance.

The **principle of least squares** says that one should adopt as the best fit estimators of a and b the values which minimise the sum of the squared residuals, $S = \sum \epsilon^2$. Thus

$$S \quad = \quad \sum_{i=1}^{n} \left[M_i - (a + b \log P_i) \right]^2$$

and $\hat{a}$ and $\hat{b}$ are obtained by differentiating S with respect to a and b, setting the resulting equations (called the **normal equations**) equal to zero, and solving for a and b.

Figure 13: Schematic diagram indicating residuals of data points in the $\{M_i, \log P_i\}$ plane.



In general, if we write the linear relation as

$$Y_i \quad = \quad a + bX_i + \epsilon_i$$

where X_i is the independent variable and Y_i as the dependent variable, the **least squares estimators** of a and b minimise

$$S \quad = \quad \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

and $\hat{a}_{\rm LS}$ and $\hat{b}_{\rm LS}$ satisfy

$$\frac{\partial S}{\partial a} = 0 \quad \text{when} \quad a = \hat{a}_{\rm LS} \qquad \frac{\partial S}{\partial b} = 0 \quad \text{when} \quad b = \hat{b}_{\rm LS}$$

Solving these equations, $\hat{a}_{\rm LS}$ and $\hat{b}_{\rm LS}$ are given by

$$\hat{a}_{\rm LS} \quad = \quad \frac{\sum y_i \sum x_i^2 \; - \; \sum y_i x_i \sum x_i}{n \sum x_i^2 \; - \; (\sum x_i)^2}$$

$$\hat{b}_{\rm LS} \quad = \quad \frac{n \sum y_i x_i \; - \; \sum y_i \sum x_i}{n \sum x_i^2 \; - \; (\sum x_i)^2}$$

where n denotes the sample size and all summations are for $i = 1, ..., n$. If the residuals are drawn from a normal pdf then it is straightforward to show that the least squares estimators are also maximum likelihood estimators (see lectures).

It can also be shown that $\hat{a}_{\rm LS}$ and $\hat{a}_{\rm LS}$ are **unbiased** estimators of a and b respectively. The variance of $\hat{a}_{\rm LS}$ and $\hat{b}_{\rm LS}$ is given by

$$
\begin{aligned}
\mathrm{var}(\hat{a}_{\rm LS}) &= \frac{\sigma^2 \sum x_i^2}{n \sum x_i^2 \; - \; \left(\sum x_i\right)^2} \\
\mathrm{var}(\hat{b}_{\rm LS}) &= \frac{\sigma^2 \, n}{n \sum x_i^2 \; - \; \left(\sum x_i\right)^2}
\end{aligned}
$$

We can use these formulae to assign an error (i.e. by taking the square root of the variance) to the least squares fitted slope and intercept. In general, $\hat{a}_{\rm LS}$ and $\hat{b}_{\rm LS}$ will not be statistically independent. This means that they have non-zero **covariance**. (Recall that we defined in Section 1.9 the covariance of two random variables, X and Y, as $\mathrm{cov}(X, Y) = E[(X - \overline{x})(Y - \overline{y})]$, and it follows that $\mathrm{cov}(X, Y) = 0$ if X and Y are independent). In fact,

$$
\mathrm{cov}(\hat{a}_{\rm LS}, \hat{b}_{\rm LS}) = \frac{-\sigma^2 \sum x_i}{n \sum x_i^2 \; - \; \left(\sum x_i\right)^2}
$$

2.4.3 : Weighted Least Squares

A common situation met in astronomy (and indeed in all the physical sciences) is where one can model the relationship between bivariate data as a straight line, but it is **not** reasonable to assume that the residuals are all drawn from the same pdf. In particular, it is often the case that the residuals each have a different variance. For example, in the case of the Cepheid PL relation, shorter period Cepheids are – on average – less luminous, which could mean that the uncertainty on the measured apparent magnitude would be larger than that for longer period Cepheids. Equally, it could be the case that the *intrinsic scatter* (as opposed to the scatter due to observational errors) about the assumed straight line relation is a function of the independent variable; this situation has recently been suggested for the Tully-Fisher relation, which is a straight line relationship between the absolute magnitude (dependent variable) and log rotation velocity (independent variable) for spiral galaxies. Thus, in such cases, the i^{th} residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance σ_i^2, where the variance is allowed to be different for each residual.

If the residuals are **not** identically distributed, this will affect the best-fit straight line relation derived for a given set of data. One must 'weight' the least squares solution to take account of the different variance on each residual, since the residuals with large variance should have less influence on determining the best-fit parameters. We call such a procedure **weighted least squares**. We can find weighted least squares estimators of a and b in a similar fashion to that for ordinary least squares, but with a modified sum of squares function, S, given by

$$S = \sum_{i=1}^{n} \left[\frac{y_i - (a + bx_i)}{\sigma_i} \right]^2$$

which yields the solution

$$\begin{aligned} \hat{a}_{\rm WLS} &= \frac{\sum \frac{y_i}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \sum \frac{y_i x_i}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2} \\ \hat{b}_{\rm WLS} &= \frac{\sum \frac{1}{\sigma_i^2} \sum \frac{y_i x_i}{\sigma_i^2} - \sum \frac{y_i}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2} \end{aligned} \qquad (2)$$

where as before all summations are for $i = 1, ..., n$. The variance of $\hat{a}_{\rm WLS}$ and $\hat{b}_{\rm WLS}$ is given by

$$\begin{aligned} \mathrm{var}(\hat{a}_{\rm WLS}) &= \frac{\sum \frac{x_i^2}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2} \\ \mathrm{var}(\hat{b}_{\rm WLS}) &= \frac{\sum \frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2} \end{aligned} \qquad (3)$$

In the case where σ_i^2 is constant, for all i, these formulae reduce to those given in Section 2.4.2 for the unweighted case.

2.4.4 : Least Squares and Linear Regression

In the case of a bivariate normal distribution we saw in section 1.9 that the conditional distribution of Y given x, denoted $p(y|x)$, was a normal distribution with mean value which was a linear function of x. In other words if we consider the **conditional expectation value** of Y given x, denoted by $E(Y|x)$, as we vary x, this conditional expectation defines a straight line in the $\{x, y\}$ plane. We call this straight line the **linear regression** or **regression line** of Y on X. It can be shown that this regression line is identical to the best-fit straight line obtained by an ordinary least squares fit to the $\{x_i, y_i\}$ data, so that in this sense least squares and linear regression are equivalent. In fact, this equivalence holds not only for a bivariate normal distribution, but any bivariate distribution for which the conditional expectation of Y given x is a linear function of x.

2.4.5 : Extending Ordinary Least Squares

The simple formulation of ordinary least squares considered in this course can be extended in several different ways. For example, one can express the dependent variable as a linear function of two or more independent variables (e.g. for the Cepheid PL relation we can include a term which depends on the colour of a Cepheid; we call this the PLC relation). This extension is known as **multilinear least squares** or **multilinear regression** and can be formulated quite neatly – and completely generally – in terms of vectors and matrices. We do not consider multilinear least squares in this course, however.

One can also modify the assumptions of ordinary least squares by accounting for errors, or residuals, on *both* variables (e.g. for the Cepheid PL relation one could allow for an uncertainty on the measured period). This means that one has to modify the form of the sum of squares function S, which has to be minimised with respect to the unknown parameters of the best-fit straight line. The details of this generalisation to errors on both variables are quite straightforward in principle, but are algebraically rather messy and we do not attempt them here.

2.5 : Goodness of Fit

We have shown how to obtain (ordinary) least squares estimators of the slope and intercept of the best-fit straight line. We must still ask how good is our linear model in the first place; i.e. we can obtain the best-fit straight line but this may still be a very poor fit to the data.

How can we test if our model is a good one? Answering this question is tantamount to determining whether the residuals of the data are, indeed, drawn from the assumed distribution – i.e. a pdf with mean zero and variance σ^2 (or σ_i^2 in the case of weighted least squares). The **true** residuals are, in fact, unknown, since they are given by

$$\epsilon_i \quad = \quad y_i - a - bx_i$$

and the true values of the parameters a and b are, of course, unknown. We can *estimate* the residuals, however, in the obvious way simply by replacing a and b in the above formula by their least squares estimators, i.e.

$$\hat{\epsilon}_i \quad = \quad y_i - \hat{a}_{\rm LS} - \hat{b}_{\rm LS} x_i$$

and ask whether these estimated residuals are consistent with our model assumptions. This provides us with a means of assessing whether the linear model is a good one in the first place.

Our assumptions are known as our **hypothesis**, and we **test** this hypothesis when we test how well our data fit our model. We call such a hypothesis test a **goodness of fit test**. (We will consider more general hypothesis tests in the next section).

2.5.1 : The χ^2 statistic

We can test how well the data fits the linear model using the χ^2 statistic. For the simple case of one independent variable this is defined as

$$\chi^2 \quad = \quad \sum_{i=1}^{n} \left[\frac{y_i - \hat{a}_{\rm LS} - \hat{b}_{\rm LS} x_i}{\sigma_i} \right]^2$$

where σ_i^2 is the variance of the i^{th} residual and is assumed known *a priori*. In other words, χ^2 is the sum of the squared residuals, weighted by their variance.

Note that we **must** know the σ_i^2 *a priori*; if we don't then we can say nothing about the goodness of fit of the data to the model (see Figure 14 below).

If the residuals are distributed as $N(0, \sigma_i)$ then the statistic given above has the χ^2 pdf, given by

$$p(\chi^2) \quad = \quad p_0 \left(\chi^2\right)^{\nu/2} e^{-\chi^2/2} \quad \chi^2 \geq 0$$

Here ν is known as the number of **degrees of freedom** of the pdf. The mean value of the pdf is ν and the variance is 2ν.

For a sample size of n, the χ^2 statistic has $\nu = n - 2$ degrees of freedom: the number of degrees of freedom is smaller than n because the statistic is formed not from the true (and unknown) residuals, but from their *estimates* – i.e. we do not know the true values of a and b and must replace them by their least squares estimators when forming the χ^2 statistic.

2.5.2 : Using χ^2 to measure goodness of fit

We use tables of the cumulative distribution function of the χ^2 RV in order to determine whether the hypothesis that the data are well described by the linear model is justified[2]. If the value of the χ^2 statistic is found to be excessively large, or excessively small, compared to its expected value, then we reject our hypothesis and seek a better model.

How does this work in practice? Tables tell us the value of the χ^2 RV for which a certain percentage of the pdf lies to the **left** of that value (we call this a **percentile** of the CDF). For example:-

$$\chi^2_{0.995} \quad = \quad \text{value of } t \text{ for which Prob}(\chi^2 < t) = 0.995 \quad = \quad 32.8 \text{ for } \nu = 15$$

$$\chi^2_{0.90} \quad = \quad \text{value of } t \text{ for which Prob}(\chi^2 < t) = 0.90 \quad = \quad 9.24 \text{ for } \nu = 5$$

Thus we require to carry out the following steps to determine the goodness of fit for our linear model, $Y = a + bX$.

1. Using the real data and the formulae of Section 2.4, determine the least squares estimators of a and b.

2. Using these estimators of a and b, and the (assumed known) variance, σ_i^2, of each residual, calculated the observed value of the χ^2 statistic, i.e.

$$\chi^2_{\text{obs}} \quad = \quad \sum_{i=1}^{n} \left[\frac{y_i \; - \; \hat{a}_{\text{LS}} \; - \; \hat{b}_{\text{LS}} x_i}{\sigma_i} \right]^2$$

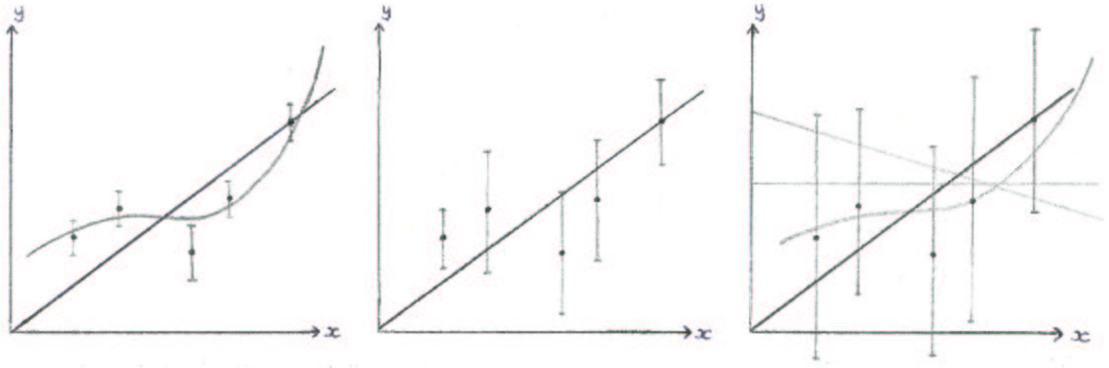
[2] Nowadays it is also common to use statistical packages on computer to determine the percentiles, rather than consulting tables

3. For the appropriate number of degrees of freedom (in this case $\nu = n - 2$, since we have two unknown parameters that we must replace with their estimators), compare $\chi^2_{\rm obs}$ with various percentiles of the χ^2 CDF, in order to determine how likely it is that one would obtain as large a value of $\chi^2_{\rm obs}$ (or indeed larger) if the linear model were correct.

4. Make a **decision** to **accept** or **reject** the hypothesis of a linear model, based on how likely $\chi^2_{\rm obs}$ is found to be.

Figure 14 shows how changing the size of the σ_i, and hence changing the value of $\chi^2_{\rm obs}$, changes our interpretation of the goodness of fit of the *same* observed data to a linear model. In the left hand panel, the errors are sufficiently small (and hence the value of $\chi^2_{\rm obs}$ sufficiently large) to indicate that a linear model is a **poor** model for the data – i.e. we need to consider a **curve**. In the right hand panel, conversely, the errors are so large (and hence the value of $\chi^2_{\rm obs}$ so small) that the best-fit straight line is **consistent** with the data, but so too are many other straight line fits, and indeed other model curves. We need more or better data to tell if the linear model is the most appropriate. In the central panel, the errors are of a size consistent with our hypothesis, as borne out by a value of $\chi^2_{\rm obs}$ which is close to the expected value for that number of degrees of freedom, and so we conclude that the linear model is a good one.

Figure 14: Best linear fits, with different σ_i and different χ^2, for the same data.



2.6 : Fitting General Models

We can apply the χ^2 goodness of fit test more generally than just to fit straight line relations. Suppose we have a physical model for the functional relationship between some variable, Y and another variable, X, i.e.

$$y_i^{\mathrm{model}} \equiv y^{\mathrm{model}}(x_i; \theta_1, ..., \theta_k)$$

where the θ_j are unknown parameters of the model. Suppose we now observe $\{y_i^{\mathrm{obs}}; i = 1, ..., n\}$, where we suppose that

$$y_i^{\mathrm{obs}} = y_i^{\mathrm{model}} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_i)$, for all i, and the ϵ_i are mutually independent.

Suppose we obtain least squares estimators, $\hat{\theta}_1, ..., \hat{\theta}_k$, of the parameters of the model[3]. Then, if our model is correct, it follows that

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i^{\mathrm{obs}} - y_i^{\mathrm{model}}}{\sigma_i} \right]^2$$

has a χ^2 distribution with $n - k$ degrees of freedom.

If the residuals are *not* normally distributed then we can still construct a statistic with an approximately χ^2 distribution by first binning the data. The average value of y^{obs} in each bin, when compared to y^{model} for that bin, will have a residual which is approximately normal due to the Central Limit Theorem. The 'closeness' to normality depends on the original pdf of the residuals and the number of points in each of the bins. Usually if this number exceeds about 15 then the approximation to normality is quite adequate.

[3] the details of how we do this in practice need not concern us in this course. For a general model it is often not possible to write down the analytic expression for the least squares estimators, in the same way as for the linear model, but there exist computer packages for determining the least squares estimators numerically in the general *non-linear* case.

SECTION 3 : Hypothesis Tests

The goodness of fit tests which we introduced in the previous section using the χ^2 statistic were an example of a **hypothesis test**. In this section we now consider hypothesis tests more generally.

3.1 : Simple Hypothesis Tests

A **simple hypothesis test** is one where we test a **null hypothesis**, denoted by H_1 (say), against an **alternative hypothesis**, denoted by H_2 – i.e. the test consists of only **two** competing hypotheses. We construct a **test statistic**, t, and based on the value of t observed for our real data we make one of the following two decisions:-

1. accept H_1, and reject H_2
2. accept H_2, and reject H_1

As an example of a simple hypothesis test, let X be a RV drawn from a normal pdf with variance equal to unity and mean value equal to μ, where it is known that either $\mu = 2$ or $\mu = -2$. Let our test statistic be simply $t = x$, the observed value of X in a random sample of size one. Let our null and alternative hypotheses be:-

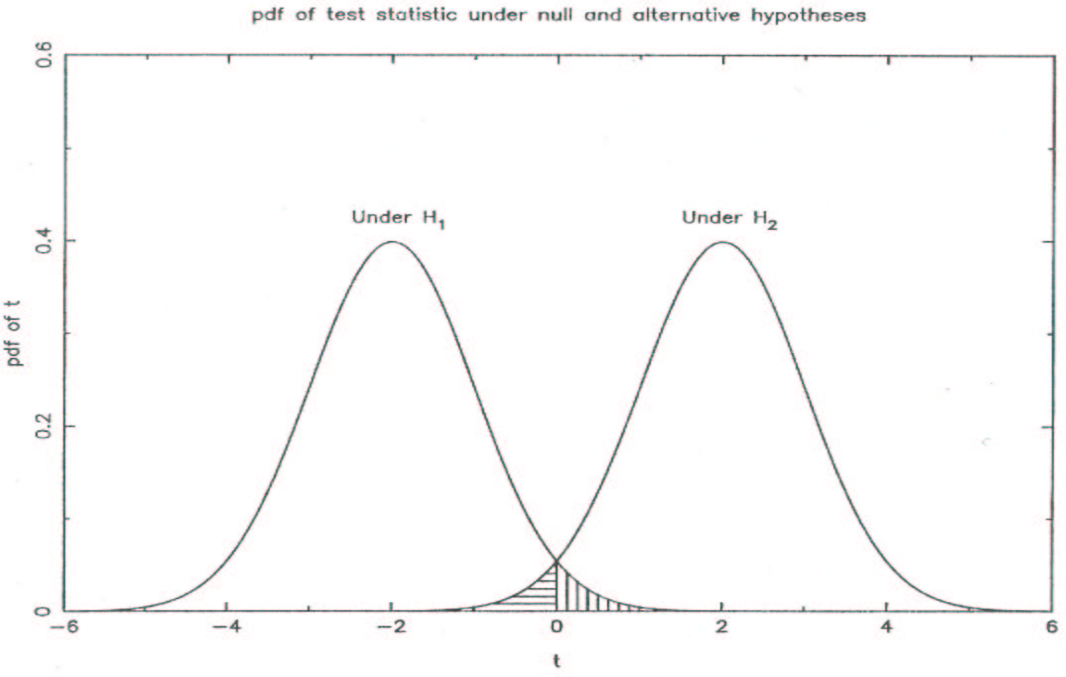
$$H_1 \; : \; \mu \; = \; -2 \qquad H_2 \; : \; \mu \; = \; 2$$

(Note that we could equal have chosen the null hypothesis to be $H_1 \; : \; \mu \; = \; 2$. The choice of which is the null and which is the alternative hypothesis – abbreviated as NH and AH – is basically up to the experimenter). Figure 15 shows the distribution of the test statistic, t, under the NH and AH specified above.

To carry out the hypothesis test we choose the **critical region** for the test statistic, t. This is the set of values of t for which we will choose to **reject** the null hypothesis and accept the alternative hypothesis. The region for which we accept the null hypothesis is known as the **acceptance region**. Note that we must choose the critical region and acceptance region ourselves. For example we might choose the critical region as the set of values of t for which $t > 0$. In other words, if our sampled value of x is found to be positive then we accept the alternative hypothesis that x was sampled from a normal pdf with mean $\mu = 2$, whereas if our sampled value of x is found to be negative, or equal to zero, then we accept the null hypothesis that x was sampled from a normal pdf with mean $\mu = -2$. In Figure 15, this particular choice of acceptance region and critical region is shown as the horizontally and vertically striped area respectively under the normal pdfs.

Note that our decision about whether the accept or reject the NH depends on the critical region, which has to be chosen by the observer. A different choice of critical region might lead to a different decision. This might seem to make the business of hypothesis testing a little subjective, but in some ways this subjectivity is inevitable. Statistical theory can never **absolutely** determine which of two competing hypotheses is correct – all it can do is tell us, provided certain assumptions are valid, how **probable** the two competing hypotheses are. Whether one (or indeed both!) of the hypotheses is then deemed to be *too* improbable to be accepted is – in the final analysis – up to the observer to decide. Very often this decision will depend on whether one is trying to prove one's own theory or model (i.e. by finding observational evidence to back it up), or disprove someone else's theory! We will return to this point shortly when we discuss significance.

Figure 15: PDF of test statistic under NH and AH for a simple hypothesis test.



While the choice of critical region may be subjective, once we have specified our choice of critical region we can objectively quantify what is the probability of making an **incorrect decision**.

3.2 : Incorrect Decisions

We can make an incorrect decision in one of two ways

3.2.1 : Type I error

A **type I error** occurs when we **reject** the null hypothesis when it is **TRUE** – i.e. when we should have accepted it. P(I) often denotes the probability of incurring a type I error.

3.2.2 : Type II error

A **type II error** occurs when we **accept** the null hypothesis when it is **FALSE** – i.e. when we should have rejected it. P(II) often denotes the probability of incurring a type II error.

We can calculate P(I) and P(II) in the simple example introduced above (see lectures), where the areas under the appropriate normal pdf can be found by consulting the tables provided.

A good hypothesis test should have small P(I) and P(II). Broadly speaking, this means that the distributions of the test statistic under H_1 and H_2 should have little overlap. We can always reduce P(I) by suitable choice of critical region, but this is inevitably at the cost of increasing P(II). It is often useful to choose the critical region which minimises some weighted combination of P(I) and P(II), but there is no general strategy suitable for all situations.

One frequently adopted criterion is the **power** of a hypothesis test, defined as the probability of rejecting H_1 when it is false, i.e. power = 1 - P(II). Choosing a critical region which maximises the power for a given alternative hypothesis is generally a useful strategy for defining a good hypothesis test.

3.3 : Level of Significance

The **level of significance** of a hypothesis test is the maximum probability of incurring a type I error which we are willing to risk when making our decision. In practice a level of significance of 5% or 1% is common. If a level of significance of 5% is adopted, for example, then we choose our critical region so that the probability of rejecting the null hypothesis when it is *true* is no more than 0.05

If the test statistic *is* found to lie in the critical region then we say that the null hypothesis is rejected at the 5% level, or equivalently that our rejection of the null hypothesis is **significant** at the 5% level. This means that, **if** the null hypothesis **is** true, and we were to repeat our experiment or observation a large number of times, then we would expect to obtain – by chance – a value of the test statistic which lies in the critical region (thus leading us to reject the NH) in no more than 5% of the repeated trials. In other words, we expect our rejection of the null hypothesis to be the *wrong* decision in no more than 5 times out of every 100 experiments. Yet another way to express this is to say that we are '95 % confident' that we have made the correct decision in rejecting the null hypothesis.

As mentioned above, the choice of significance level is somewhat subjective. Suppose, for example, that one is comparing the model prediction of another astronomer's favourite theory (here the NH) to the prediction of one's own pet theory (here the AH). In this case one might regard rejection at the 10% significance level to be sufficient grounds for ruling out the other astronomer's theory. Why? Because **if** the other astronomer's theory is true, there is at most a one in ten chance of the test statistic falling in the critical region (i.e. a one in ten chance of obtaining data similar to – as 'bad' as, if you like – the **actual** data which we do obtain). If, on the other hand, one were seeking support for one's own theory as the NH, then rejection at the 10% significance level might not be sufficient grounds to give up on one's theory, since one can argue that the **actual** data obtained happens to be one of those one in ten data sets which, by chance (or 'bad luck', if you like), yield a test statistic lying in the critical region – even when the NH is true.

How can we get around this? As remarked in section 3.1, we can always choose a more stringent critical region. For example, if we could reject the NH at, say, the 1% or 0.1% level, then we can be much more sure that the test statistic obtained for our real data does not lie in our critical region by chance, even though the NH is true. In other words, we reduce the probability of a type I error. But recall from section 3.1 that this will inevitably increase the chances of accepting the alternative hypothesis when it is false – i.e. making a type II error. Again, the key here is for the distribution of the test statistic to be so nearly disjoint under the null and alternative hypotheses (i.e. having so little overlap) that we can afford to adopt a 'tough' critical region without increasing P(II) too much. Clearly one effective way to reduce the overlap between the pdf of test statistics is to acquire more, and better, data, but in astronomy this is often a painful – and expensive – solution!

3.4 : Two Tailed Tests

It is common for the critical region to be defined as both the upper *and* lower tails of the distribution of the test statistic under H_1. For example, consider the random variable $X \sim N(\mu, 1)$ and the test statistic $t = x$. Consider the null and alternative hypotheses

$$H_1 \quad : \quad \mu = 0 \qquad\qquad H_2 \quad : \quad \mu \neq 0$$

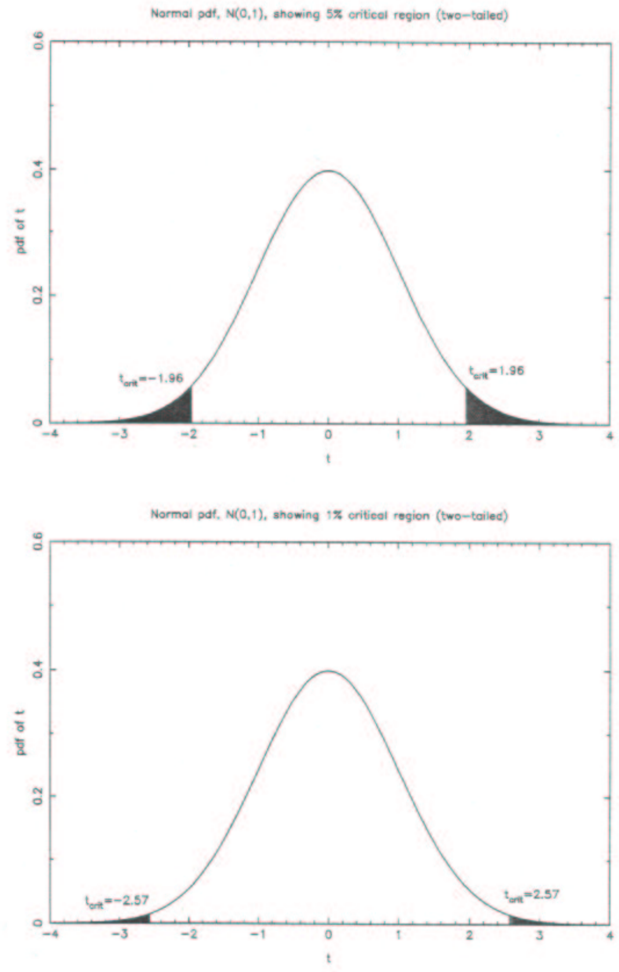
Then a value of t either much larger or smaller than zero might lead us to reject H_1 and accept H_2, since H_2 only states that the mean value is *different* from zero. In this example, adopting a 5% level of significance with a two tailed test would give as the critical region for t

$$\{t : |t| \geq 1.96\}$$

while for a 1% level of significance with a two tailed test, the critical region for t would be

$$\{t : |t| \geq 2.57\}$$

Figure 16: Two-tailed critical regions



In these examples $t = \pm 1.96$ and $t = \pm 2.57$ are the **critical values** for the test statistic; i.e. they indicate the boundary between the critical region and acceptance region. These two-tailed critical regions are shown in Figure 16.

3.5 : Goodness of Fit for Discrete Distributions

We can illustrate some of the important ideas of hypothesis testing by considering how we test the goodness of fit of data to **discrete** distributions. We do this using the χ^2 statistic.

Suppose we carry out n observations and obtain as our results k different discrete outcomes, $E_1, ..., E_k$ which occur with frequencies $o_1, ..., o_k$ ('o' for 'observed'). An example of such observations might be the number of meteors observed on n different nights, or the number of photons counted in n different pixels of a CCD.

Consider the null hypothesis that the observed outcomes are a sample from some model discrete distribution (e.g. a Poisson distribution). Suppose, under this null hypothesis, that the k outcomes, $E_1, ..., E_k$, are expected to occur with frequencies $e_1, ..., e_k$ ('e' for 'expected'). We can test our null hypothesis by comparing the observed and expected frequencies and determining if they differ significantly. We construct the following χ^2 test statistic.

$$\chi^2 \quad = \quad \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

where $\sum o_i = \sum e_i = n$. Under the null hypothesis this test statistic has approximately a χ^2 pdf with $\nu = k - 1 - m$ degrees of freedom. Here m denotes the number of parameters (possibly zero) of the model discrete distribution which one needs to estimate before one can compute the expected frequencies, and ν is reduced by one further degree of freedom because of the constraint that $\sum e_i = n$. In other words, once we have computed the first $k-1$ expected frequencies, the k^{th} value is uniquely determined by the sample size n.

This χ^2 goodness of fit test need not be restricted *only* to discrete random variables, since we can effectively produce discrete data from a sample drawn from a continuous pdf by binning the data. Indeed, as we remarked in Section 2.2.7 the Central Limit Theorem will ensure that such binned data are approximately normally distributed, which means that the sum of their squares will be approximately distributed as a χ^2 random variable. The approximation to a χ^2 pdf is very good provided $e_i \geq 10$, and is reasonable for $5 \leq e_i \leq 10$.

3.5.1 : Example 1

A list of 1000 'random' digits – integers from 0 to 9 – are generated by a computer. Can this list of digits be regarded as uniformly distributed?

Suppose the integers appear in the list with the following frequencies:-

r	0	1	2	3	4	5	6	7	8	9
o_r	106	88	97	101	92	103	96	112	114	91

Let our NH be that the digits are drawn from a uniform distribution. This means that each digit is expected to occur with equal frequency – i.e. $e_r = 100$, for all r. Thus:-

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = 7.00$$

Suppose we adopt a 5% level of significance. The number of degrees of freedom, $\nu = 9$; hence the critical value of $\chi^2 = 16.9$ for a one-tailed test. Thus, at the 5% significance level we **accept** the NH that the digits are uniformly distributed.

3.5.2 : Example 2

A coin is tossed 200 times, and 115 heads and 85 tails are recorded. Test the null hypothesis that the coin is fair, using a 5% level of significance.

Under the NH of a fair coin we have $e_1 = e_2 = 100$. Thus:-

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = 4.5$$

Here, the number of degrees of freedom, $\nu = 1$, for which we have a critical value of $\chi^2 = 3.84$. Hence we **reject** the NH at the 5% significance level – i.e. the coin is *not* fair.

3.5.3 : Example 3

The table below shows the number of nights during a 50 night observing run when r hours of observing time were 'clouded out'. Fit a Poisson distribution to these data for the pdf of r and determine if the fit is acceptable at the 5% significance level.

r	0	1	2	3	4	> 4
No. of nights	21	18	7	3	1	0

Of course one might ask whether a Poisson distribution is a sensible model for the pdf of r since a Poisson RV is defined for any non-negative integer, whereas r is clearly at most 12 hours. However, as we saw in Section 1.3.2, the shape of the Poisson pdf is sensitive to the value of the mean, μ, and in particular for small values of μ the value of the pdf will be negligible for all but the first few integers, and so we neglect all larger integers as possible outcomes. Hence, in fitting a Poisson model we also need to estimate the value of μ. We take as our estimator of μ the **sample mean**, i.e.

$$\hat{\mu} \quad = \quad \frac{21 \times 0 \, + \, 18 \times 1 \, + \, +7 \times 2 \, + \, 3 \times 3 \, + \, 1 \times 4}{50} \quad = \quad 0.90$$

Substituting this value into the Poisson pdf we can compute the *expected* outcomes, $e_r = 50\,p(r;\hat{\mu})$, where

$$p(0;0.90) = 0.4066 \qquad p(1;0.90) = 0.3659 \qquad p(2;0.90) = 0.1647$$
$$p(3;0.90) = 0.0494 \qquad p(4;0.90) = 0.0111 \qquad p(5;0.90) = 3.3 \times 10^{-5}$$

If we consider only five outcomes, i.e. $r \leq 4$, since the value of the pdf is negligible for $r > 4$, then the number of degrees of freedom, $\nu = 3$ (remember that we had to estimate the mean, μ). The value of the test statistic is $\chi^2 = 0.68$, which is smaller than the critical value. Hence we **accept** the NH at the 5% level – i.e. the data are well fitted by a Poisson distribution.

3.5.4 : The Binomial Distribution

In Section 3.5.3 we could have fitted the data with another discrete model – the **binomial distribution**. Suppose there are a total of n hours in each observing night (e.g. $n = 8$ or $n = 12$). Let θ denote the probability of any single hour being 'clouded out'. The binomial distribution gives the probability of getting r out of n hours clouded out $(r = 0, 1, ..., n)$, viz:-

$$p(r;\theta) = \frac{n!}{r!(n-r)!}\theta^r(1-\theta)^{n-r}$$

$p(r;\theta)$ is the binomial pdf. It is quite straightforward to show (see handout) that the binomial distribution has mean, $E(r) = n\theta$ and variance, $\text{var}(r) = n\theta(1-\theta)$.

As in Section 3.5.3, we have to estimate a single parameter – in this case θ (assuming that the number of observing hours, n, is known) – in fitting the data to a binomial model. We do this by equating the sample mean, $\hat{\mu}$, with the expected value of r, i.e. $n\theta$. We can then construct a χ^2 statistic exactly as in 3.5.3. (remembering to reduce the number of degrees of freedom by one because we need to estimate θ).

3.6 : The Kolmogorov-Smirnov Test

Suppose we want to test the hypothesis that a sample of data is drawn from the underlying population with some given pdf. We could do this by binning the data and comparing with the model pdf using the χ^2 test statistic. This approach might be suitable, for example, for comparing the number counts of photons in the pixels (i.e. the bins) of a CCD array with a bivariate normal model for the 'point spread function' of the telescope optics, where the centre of the bivariate normal defines the position of a star.

For small samples this does not work well, however, as we cannot bin the data finely enough to usefully constrain the underlying pdf – particularly if our pdf is *multivariate*, as in the case of the bivariate normal example above, and requires several parameters to define it.

A more useful approach in this situation is to compare the sample **cumulative distribution function** with a theoretical model. We can do this using the **Kolmogorov-Smirnov** (KS) test statistic.

Let $\{x_1, ..., x_n\}$ be an iid random sample from the unknown population. Suppose the $\{x_i\}$ have been arranged in ascending order. The sample cdf, $S_n(x)$, of X is defined as:-

$$S_n(x) = \begin{cases} 0 & x < x_1 \\ \frac{i}{n} & x_i \leq x < x_{i+1}, \qquad 1 \leq i \leq n-1 \\ 1 & x \geq x_n \end{cases}$$

i.e. $S_n(x)$ is a step function which increments by $1/n$ at each sampled value of x.

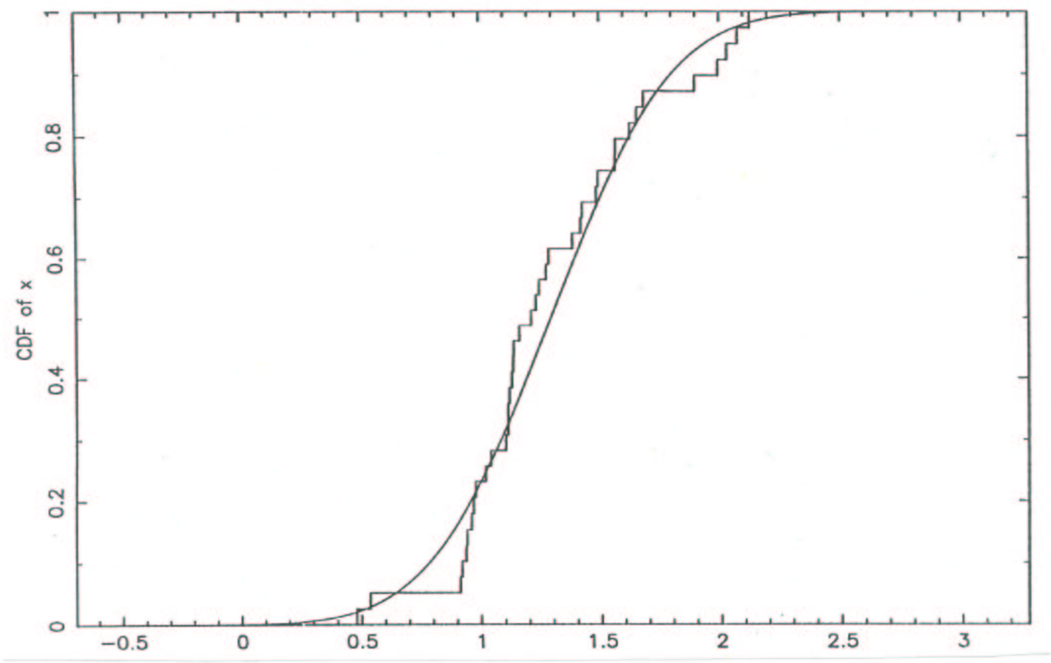
Let the model cdf be $P(x)$, corresponding to pdf $p(x)$, and let the null hypothesis be that our random sample is drawn from $p(x)$. The KS test statistic is

$$D_n \;=\; \max |P(x) - S_n(x)|$$

It is easy to show that D_n always occurs at one of the sampled values of x. The remarkable fact about the KS test is that the distribution of D_n under the null hypothesis is **independent of the functional form** of $P(x)$. In other words, whatever the form of the model cdf, $P(x)$, we can determine how likely it is that our *actual* sample data was drawn from the corresponding pdf. Critical values for the KS statistic are tabulated or can be obtained from numerical algorithms.

Figure 17 shows the KS test applied to the log period distribution in a sample of LMC Cepheids. Shown is the sample cdf of the 39 stars, together with the model cdf with which they are being compared: a normal distribution with mean and variance equal to the sample mean and variance of the real data. The observed value of the test statistic for these data, $D_{\mathrm{obs}} = 0.124$. Comparison with the critical values of the distribution show that $\mathrm{Prob}(D_n > D_{\mathrm{obs}}) = 0.562$. Thus, if the NH is true, there is a more than 50% chance that one would obtain as large, or indeed larger, a value of D_n for a randomly chosen sample of 39 Cepheids drawn from the model normal pdf. This clearly suggests that we should accept the null hypothesis for these data – i.e. the distribution of log periods is adequately described by a normal pdf.

Figure 17: Example KS Test



There is also a two-sample version of the KS test, where one tests the null hypothesis that the two samples are drawn from the same underlying population. The test statistic is now

$$D_{m,n} \quad = \quad \max|S_m(x) - S_n(x)|$$

where S_m and S_n denote the sample CDFs of two samples of size m and n respectively. Again the distribution of $D_{m,n}$ under the null hypothesis is independent of the underlying pdf. This is especially useful, because it means that we can test whether two samples are drawn from the same underlying population **without having to assume anything about the form of that population**.

The KS test is an example of a **robust**, or **nonparametric**, test since one can apply the test with minimal assumption of a parametric form for the underlying pdf. The price for this robustness is that the **power** of the KS test is lower than other, parametric, tests. In other words there is a higher probability of accepting a false null hypothesis – that two samples *are* drawn from the same pdf – because we are making no assumptions about the parametric form of that pdf.

3.7 : The Student's t Test

In Section 3.1 we introduced the notion of a hypothesis test, and gave some important definitions, by considering a hypothesis test to determine the mean, μ, of a normal pdf with known variance, $\sigma^2 = 1$, based on a single]sampled value, x. Effectively we constructed the test statistic

$$z = \frac{x - \mu}{\sigma}$$

where $\mu = -2$ under H_1 and $\mu = 2$ under H_2 and z was a RV drawn from a standard normal pdf, $N(0,1)$, with mean zero and unit variance. (N.B. recall that a statistic cannot depend on any *unknown* parameters, but here σ is assumed known and μ is specified exactly under either H_1 and H_2, so it makes sense to regard z as a statistic).

The more realistic situation, on the other hand, is where σ is *not* known *a priori*. In this case we can infer nothing about μ from a single observation since we have no idea of how 'broad' the pdf is. If $n \geq 2$, however, then we can construct a hypothesis test for the value of the true mean, μ, by first determining the sample mean and variance of our random sample.

Suppose we want to test the hypothesis that the true mean takes some specific value, μ_0, i.e. we take as our null and alternative hypotheses:-

$$\text{NH} : \mu = \mu_0 \qquad \text{AH} : \mu \neq \mu_0$$

We construct the following test statistic

$$t = \frac{\hat{\mu} - \mu_0}{\sigma_{\hat{\mu}}}$$

where $\hat{\mu}$ is the sample mean, i.e. $\hat{\mu} = \frac{1}{n}\sum x_i$, and $\sigma_{\hat{\mu}}$ is the **standard error on the mean** (see handout), i.e.

$$\sigma_{\hat{\mu}} = \left[\frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \right]^{\frac{1}{2}}$$

Thus

$$t = \frac{\sqrt{n(n-1)}(\hat{\mu} - \mu_0)}{[\sum (x_i - \hat{\mu})^2]^{\frac{1}{2}}}$$

t has a pdf known as the student's t distribution. It is similar in shape to a standard normal pdf, $N(0,1)$, but with wider 'wings' (i.e. positive kurtosis) and its shape also depends on n – see the figure in the statistical tables. The pdf of t has $\nu = n - 1$ degrees of freedom. Thus, for a sample of size n, to carry out our hypothesis test we determine the value of the student's t statistic under the NH that $\mu = \mu_0$ and compare this value with the critical values of the pdf, for the appropriate number of degrees of freedom.

Note that the hypothesis test given above, where our AH is $\mu \neq \mu_0$, calls for a **two-tailed test**, since a value of t significantly larger *or* smaller than zero would argue in favour of the AH. If, on the other hand, we want only to test if $\mu > \mu_0$ ($\mu < \mu_0$) then we would take as our critical region the upper (lower) tail of the student's t distribution.

3.8 : Difference of Means

Let $\{x_1, ..., x_{n_1}\}$ and $\{y_1, ..., y_{n_2}\}$ be iid random samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively, where $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Suppose we wish to test the NH that $\mu_1 - \mu_2 = \mu_0$, i.e. the means of the pdfs from which the two samples are drawn differ by a fixed amount. Under the NH then the difference of the sample means, $\hat{\mu_1} - \hat{\mu_2}$, is a normal pdf with mean μ_0 and variance $\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$. If σ^2 were known then the appropriate test statistic to test the NH would be

$$z = \frac{(\hat{\mu_1} - \hat{\mu_2}) - \mu_0}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which under the NH would have pdf $N(0,1)$. If, as in Section 3.7, σ^2 is *not* known *a priori*, then we use the test statistic

$$t = \frac{(\hat{\mu_1} - \hat{\mu_2}) - \mu_0}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$\hat{\sigma} = \left[\frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (x_i - \hat{\mu_1})^2 + \sum_{i=1}^{n_2} (y_i - \hat{\mu_2})^2 \right) \right]^{\frac{1}{2}}$$

(i.e. $\hat{\sigma}^2$ is the weighted mean of the unbiased estimators, from the first and second samples, of the variance on a single observation – see lectures)

Under the NH t has the student's t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. Clearly to test if $\mu_1 = \mu_2$ we simply set $\mu_0 = 0$.

3.9 : F Test for the Ratio of Variances

Let $\{x_1, ..., x_{n_1}\}$ and $\{y_1, ..., y_{n_2}\}$ be iid random samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively. Unlike Section 3.8, we do **not** now assume that $\sigma_1^2 = \sigma_2^2$. In fact we specifically want a simple hypothesis test of:-

$$\text{NH} : \quad \sigma_1^2 = \sigma_2^2 \qquad \text{AH} : \quad \sigma_1^2 \neq \sigma_2^2$$

We use the test statistic, f, defined by:-

$$f \quad = \quad \frac{\hat{\sigma}_1^2}{\hat{\sigma}_1^2}$$

where

$$\hat{\sigma}_1^2 \; = \; \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \hat{\mu_1})^2 \quad \text{and} \quad \hat{\sigma}_2^2 \; = \; \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \hat{\mu_2})^2$$

Under the NH, that the two distributions have equal variance, this test statistic has a pdf known as the F distribution, with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom. (We usually write this as F_{ν_1,ν_2}). The pdf of F has a complicated analytic form which need not concern us here. The essential idea of how it is used in practice is that, if the NH is true, then values of the f statistic significantly larger or smaller than unity are unlikely. Hence, by computing the f statistic and comparing the observed value with tabulated critical values, we can make a decision whether to accept or reject the NH that the pdfs from which the two samples were drawn have equal variance.

Typical astrophysical problems to which the F test can be applied include comparing the luminosity function of stars of different spectral types, or galaxies of different morphological type. One can also test, for example, whether the spatial distribution of galaxies of different morphological types is significantly different – e.g. do spirals and ellipticals have the same spatial distribution in galaxy clusters, or are ellipticals found preferentially in the cores of clusters. (Primordial spirals which originally formed in cluster cores are thought to have been torn apart and 'cannibalised' by ellipticals because of the strong tidal forces in the cluster core, so that they are not found in the cores of clusters today). See example sheet 4 for some similar applications of the F test.

3.10 : Hypothesis Tests on the Sample Correlation Coefficient

The final type of hypothesis test which we consider is associated with testing whether two variables are statistically independent, which we can do by considering the value of the **sample correlation coefficient**. In Section 1.9 we defined the covariance of two RVs, X and Y, as

$$\text{cov}(X,Y) \quad = \quad E[(X - \mu_{\text{x}})(Y - \mu_{\text{y}})]$$

and the correlation coefficient, ρ, as

$$\rho \quad = \quad \frac{\text{cov}(X,Y)}{\sigma_{\text{x}}\sigma_{\text{y}}}$$

While we defined ρ in Section 1.9 in the context of its role as a parameter of the bivariate normal distribution, one can define the covariance and correlation coefficient of *any* two RVs (i.e. with *any* bivariate distribution) using the above formulae. As in the case of a bivariate normal pdf, it follows that

$$X \text{ and } Y \text{ are independent} \quad \leftrightarrow \quad \text{cov}(X,Y) = 0 \quad \leftrightarrow \quad \rho = 0$$

We estimate ρ by the **sample correlation coefficient**, $\hat{\rho}$, defined by:-

$$\hat{\rho} \quad = \quad \frac{\sum(x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)}{\sqrt{[\sum(x_i - \hat{\mu}_x)^2]\,[\sum(y_i - \hat{\mu}_y)^2]}}$$

where, as usual, $\hat{\mu}_x$ and $\hat{\mu}_y$ denote the sample means of X and Y respectively, and all sums are over $1, ..., n$, for sample size, n. $\hat{\rho}$ is also often denoted by r, and is referred to as 'Pearson's correlation coefficient'.

Of course, if X and Y *do* have a bivariate normal pdf, then ρ corresponds precisely to the parameter defined in Section 1.9. To test hypotheses about ρ we need to know the sampling distribution of $\hat{\rho}$. We consider two special cases, both of which are when X and Y have a bivariate normal pdf.

3.10.1 : $\rho = 0$ (i.e. X and Y are independent)

If $\rho = 0$, then the statistic

$$t \quad = \quad \frac{\hat{\rho}\sqrt{n-2}}{1 - \hat{\rho}^2}$$

has a student's t distribution, with $\nu = n - 2$ degrees of freedom. Hence, we can use t to test the hypothesis that X and Y are independent. (See example sheets and lectures).

3.10.2 : $\rho = \rho_0 \neq 0$

In this case, then **for large samples**, the statistic

$$z \quad = \quad \frac{1}{2} \log_e \left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}} \right)$$

has an approximately normal pdf with mean, μ_z and variance σ_z^2 given by

$$\mu_z \;=\; \frac{1}{2} \log_e \left(\frac{1 + \rho_0}{1 - \rho_0} \right) \qquad \sigma_z^2 \;=\; \frac{1}{n-3}$$

SECTION 4 : Point and Interval Estimation

In the previous sections we have discussed how to derive a **point estimator** of a parameter – i.e. a single number, $\hat{\theta}$, which we associate with the *true* (but unknown) value of some parameter, θ. We derived $\hat{\theta}$ by applying e.g. the principle of maximum likelihood or the principle of least squares. In order to assess the likely *range* of true values of of θ, we can derive the dispersion, $\sigma_{\hat{\theta}}$, of the estimator $\hat{\theta}$ (equal to the square root of the variance of the estimator) – or more generally the *covariance* in the multivariate case, where we are simultaneously estimating several parameters and our estimates may not be independent. Thus, we can adopt

$$\hat{\theta} \quad \pm \quad \sigma_{\hat{\theta}}$$

as a suitable measure of the range of likely true values of θ.

We can approach the problem in a complementary fashion, by deriving an **interval estimate** for θ.

4.1 : Defining Confidence Intervals

We illustrate the construction of confidence intervals for a specific example. Consider an iid random sample of size n from the normal distribution $N(\mu, \sigma^2)$, and suppose that the dispersion, σ is known a priori. If we define the variable

$$z \quad = \quad \frac{\sqrt{n}\,(\hat{\mu} - \mu)}{\sigma}$$

then it follows that $z \sim N(0,1)$. We therefore know that

$$\text{Prob}\,[-1.96 \leq z \leq 1.96] \quad = \quad 0.95$$

After some algebra one may easily show that this probability statement is precisely equivalent to

$$\text{Prob}\left[\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} \;\leq\; \mu \;\leq\; \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right] \quad = \quad 0.95$$

We call the interval

$$\left[\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}}\;,\;\hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

a **95% confidence interval** for the true mean, μ. We refer to the two ends of the confidence interval as **95% confidence limits** for μ.

Confidence intervals involving the standard normal pdf, $N(0,1)$, are particularly common. Thus

$$\hat{\mu} \pm 2.58\frac{\sigma}{\sqrt{n}} \qquad \hat{\mu} \pm 2.64\frac{\sigma}{\sqrt{n}} \qquad \hat{\mu} \pm \frac{\sigma}{\sqrt{n}}$$

are 99%, 90% and 68% confidence intervals for μ respectively.

4.2 : Interpreting Confidence Intervals

We must be careful in interpreting the meaning of probability statements concerning confidence intervals.

In the above example of sampling from $N(\mu, \sigma^2)$, for a *given* sample of actual values from the pdf $\hat{\mu}$ is a unique number, so the probability that the true mean value, μ, lies within the chosen confidence interval is either zero or unity.

The meaning of a confidence interval requires one to think in terms of repeating the process of random sampling from the pdf a large number of times – each time obtaining a different value of of the sample mean, and hence different confidence limits for μ, for the *same* fixed (but unknown) true mean, μ. The probability statement

$$\text{Prob}\left[\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right] \quad = \quad 0.95$$

means that we would expect μ to lie within the confidence limits in 95% of the large number of different samples. We are, thus, **95% confident** that μ lies within in the interval that we obtain with our actual, observed, value of $\hat{\mu}$.

4.3 : Shortest Confidence Intervals

Finally, note that confidence intervals of a given percentage level are *not* unique. One can prove, however, that the **shortest** confidence interval for the mean of a normal pdf corresponds to the case where the sample mean is taken to lie precisely in the centre of the confidence interval.

Thus, the shortest $100(1-\alpha)\%$ confidence interval for μ in the above example, is

$$\hat{\mu} \quad \pm \quad z_\alpha \frac{\sigma}{\sqrt{n}}$$

where z_α satisfies

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_\alpha} e^{-t^2/2} dt \quad = \quad \frac{\alpha}{2}$$