Statistical Astronomy

Introduction

- This is a course on how to make general statements about the Universe based on specific examples of its behaviour – fundamental to all physical sciences.
- Astronomy/astrophysics is unusual because it is generally not an experimental science (you can't add a little more carbon to a star to see what happens!)
- We proceed by plausible reasoning, improving our state of knowledge by incorporating new information (as and when it becomes available) into our physical theories.

But what exactly is 'plausible reasoning'?

Types of reasoning

- All stars in the Pleiades are within 200 pc of Earth (STATEMENT C)
- Alcyone is a member of the Pleiades (STATEMENT B)
- Alcyone is within 200pc of Earth (STATEMENT A)

These three statements are related.

Let us assume C is true. It defines the hypothesis space in which we are reasoning.

Then:

If B is true, then A is true

If A is false, then B is false

This is deductive reasoning. A is a logical consequence of B (and C)

But what can we say about B if A (and C) are true? We are not told 'all stars within 200pc of Earth are in the Pleiades'!)

But we can say:

If A is true, then B is more plausible ⊛

If B is false, then A is less plausible ⊛⊛

This is an example of plausible reasoning and is the basis of physical model-building. Many such statements combine to make our degree of belief in B very high indeed.

Boolean Algebra

Both deductive and plausible reasoning follow the rules of logic notated by George Boole (1854):

Denote a set of statements by A, B, C...

The negation of a statement is written:

$\overline{A} \equiv$ 'A is false'

The 'logical product' written:

$AB \equiv$ 'both A and B are true'

The 'logical sum' written:

$A + B \equiv$ 'at least one of A or B is true'

Further properties:

$$A(B+C) = AB + AC$$

$$A + AB = A$$

$$A + BC = (A+B)(A+C)$$

$$A + \bar{A} = \text{true}$$

$$A\bar{A} = \text{false}$$

$$\overline{(\bar{A})} = A$$

$$A + \bar{A}B = A + B$$

$$\left.\begin{aligned} \overline{(A+B)} &= \bar{A}\,\bar{B} \\ \overline{AB} &= \bar{A} + \bar{B} \end{aligned}\right\} \text{De Morgan's Theorem}$$

We can extend this to include the idea of plausible reasoning. Write

$A|B \equiv$ "the conditional plausibility that A is true given that B is true".

We could assign a real number to this, so that a larger number represents a greater plausibility.

In our Pleiades example we could write

$$B|AC > B|C \quad (*)$$

$$A|\bar{B}C < A|C \quad (**)$$

- It can be shown that these degrees of plausibility can be consistently mapped onto a real function, $p(x)$ which is continuous and monotonically increasing in the range $0 \leq p(x) \leq 1$.

- We can write our sum rule as

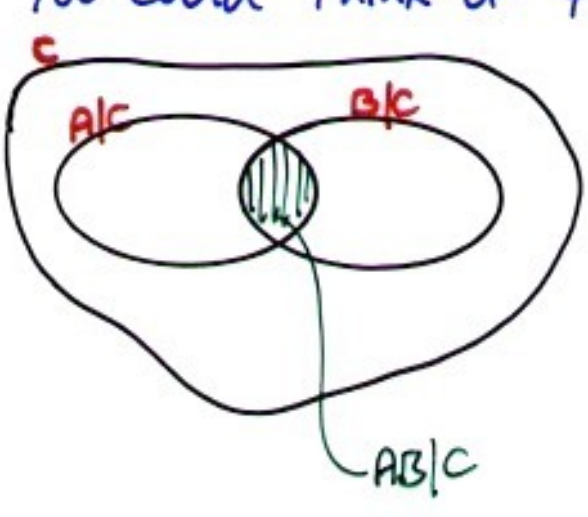
$$p(A|B) + p(\bar{A}|B) = 1$$

("the plausibility of a statement A, given B, plus the plausibility of its negation, given B, must equal 1")

- The product rule (relating AB to A + B) is

$$p(AB|C) = p(A|C)\,p(B|AC) = p(B|C)\,p(A|BC)$$

You could think of this in terms of a Venn Diagram:



If A and C are true, only a fraction $\frac{AB|C}{A|C}$ of this certainty corresponds to B also being true.

ie $\frac{p(AB|C)}{p(A|C)} = p(B|AC)$

Bayes' Theorem

We can write the product rule as

$$p(B|Ac) = \frac{p(B|C)\, p(A|BC)}{p(A|C)}$$

This is known as Bayes' Theorem (Rev. Thomas Bayes, 1763)

Bayes used it to assess hypotheses in the light of fresh evidence.

Historically, p has been called PROBABILITY, and the process of logical reasoning that uses it is Bayesian Probability Theory (BPT)

- The interpretation of 'probability' as 'degree of belief' is both the oldest and the youngest interpretation.
  - oldest because it was the original idea introduced by Laplace, Bernoulli and Bayes
  - youngest because it was totally eclipsed by the 'frequentist' interpretation of probability until it was put on firmer footing by Jeffreys (1939) and Jaynes (1950s →)

Summary

Bayesian Probability Theory (BPT) regards 'probability' as a real number between 0 and 1 measuring the plausibility of a proposition when incomplete knowledge means we cannot know its truth or falsehood with certainty.

With background assumptions C, propositions A and B have probabilities that obey

$$p(A|C) + p(\bar{A}|C) = 1 \qquad \text{(sum rule)}$$

$$p(AB|C) = p(A|BC)\,p(B|C) \qquad \text{(product rule)}$$

- Frequentist Probability Theory (FPT) is considered in the next lecture. Later, we will return to BPT to see how to assign probabilities and interpret its results in detail.

But first an example:

Example of Bayes Theorem in action

- Harold may have a rare disease, present in 1 out of 10000 people of his age and background.

He takes a blood test which is known to show positive in 95% of cases when the person has the disease, an 1% of cases when they don't.

Harold tests positive! What is the probability that he has the disease?

Ans: Denote H = healthy, D = diseased
+ = positive result, − = negative result.

Let C represent the assumption that Harold is typical of the population from which these results are drawn. Then:

$$p(D|C) = 0.0001; \quad p(H|C) = 0.9999$$
$$p(+|DC) = 0.95$$
$$p(+|HC) = 0.01$$

we want to know $p(D|+C)$, the probability that he has the disease, given he tested +ve.

You may think he's in trouble. Lets see:

By Bayes theorem

$$p(D|+C) = \frac{p(D|C) \cdot p(+|DC)}{p(+|C)}$$

($p(D|C)$: 0.0001; $p(+|DC)$: 0.95)

We know everything on the right except $p(+|C)$, the probability of a positive result for a group member. But each group member is either healthy or diseased,

so

$$p(+|C) = \overbrace{p(H|C) \cdot p(+|HC)}^{\text{healthy}} + \overbrace{p(D|C) \cdot p(+|DC)}^{\text{diseased}}$$

$p(H|C)$: 0.9999; $p(+|HC)$: 0.01; $p(D|C)$: 0.0001; $p(+|DC)$: 0.95

[proof - EFS]

so

$$p(D|+C) = \frac{0.0001 \times 0.95}{0.9999 \times 0.01 + 0.0001 \times 0.95}$$

$$\simeq 0.01$$

ie, there is only a 1% probability he has the disease! Most positive results are from healthy people, because there are so many healthy people.

MORAL:

$$p(+|DC) \neq p(D|+C)$$

$p(+|DC)$: 0.95 (test effectiveness); $p(D|+C)$: 0.01 (Harold having disease)

Example #2

What is $p(A+B|C)$ in terms of $p(A|C) + p(B|C)$?

We can evaluate this by multiple applications of the product and sum rules:

Remember $A+B = \overline{\bar{A}\,\bar{B}}$ (Boolean algebra)

so using $p(X|Y) + p(\bar{X}|Y) = 1$ (sum rule)

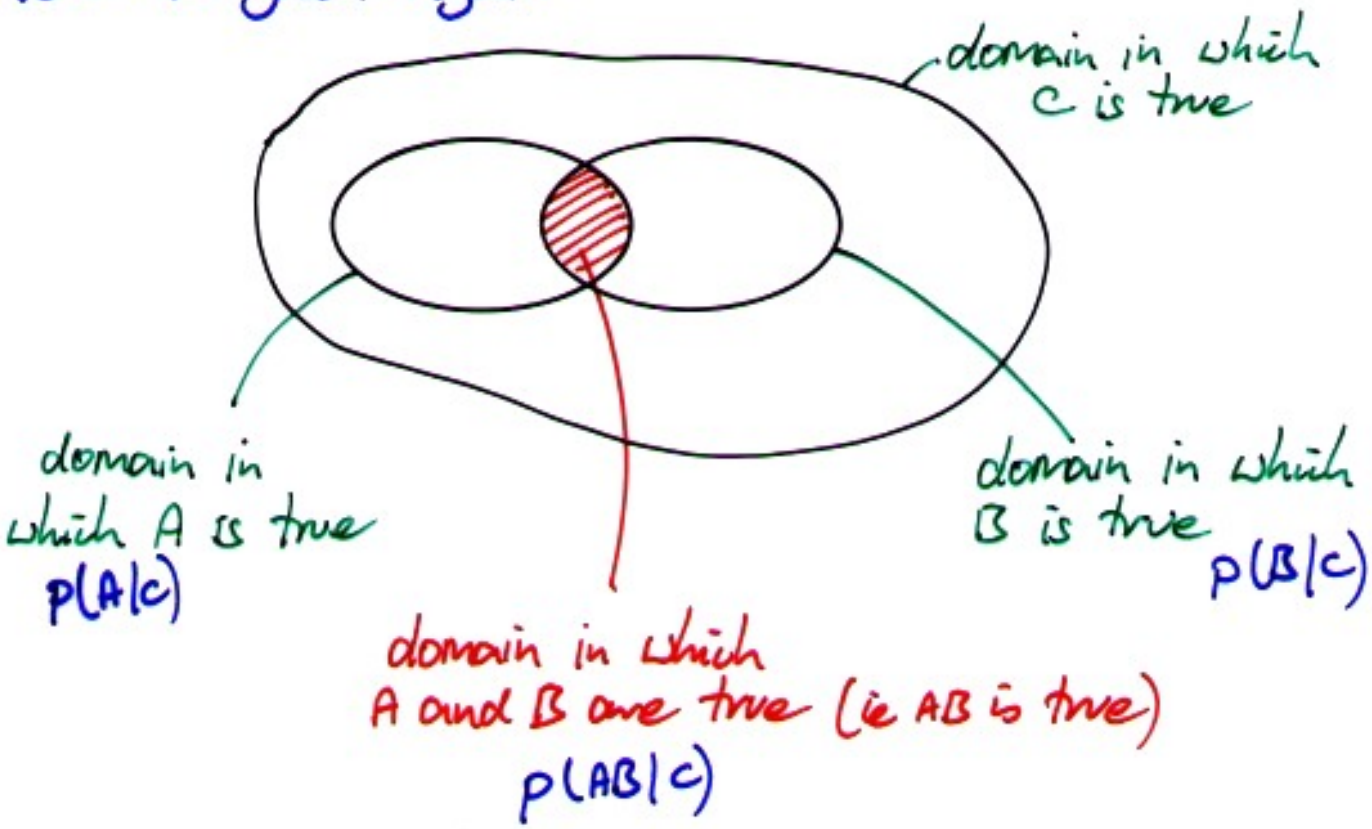
$$\Rightarrow \quad p(A+B|C) + p(\bar{A}\,\bar{B}|C) = 1$$

so
$$\begin{aligned} p(A+B|C) &= 1 - p(\bar{A}\,\bar{B}|C) \\ &= 1 - p(\bar{A}|C)\,p(\bar{B}|\bar{A}C) \quad \text{(prod. rule)} \\ &= 1 - p(\bar{A}|C)[1 - p(B|\bar{A}C)] \quad \text{(sum)} \\ &= p(A|C) + p(\bar{A}|C)p(B|\bar{A}C) \quad \text{(sum)} \\ &= p(A|C) + p(B|C)\,p(\bar{A}|BC) \quad \text{(prod)} \\ &= p(A|C) + p(B|C)[1 - p(A|BC)] \quad \text{(sum)} \end{aligned}$$

$$\Rightarrow \quad p(A+B|C) = p(A|C) + p(B|C) - p(AB|C)$$

This is called the Extended Sum Rule

You can remember this by thinking of a simple Venn Diagram again:



$$p(A+B|C) = p(A|C) + p(B|C) - p(AB|C)$$

'A or B'

to avoid counting the hatched area twice.

Frequentist Probability Theory

Bayesian probability is overtly subjective. It works with states of belief, so that different people will assert different probabilities to the same statement if their state of knowledge is different.

- 'Frequentist probability' is one attempt to remove this subjectivity. Many would say it fails to do this (and shouldn't try in the first place!) but it <u>is</u> popular and remains the basis of much analyses.

- It's important to understand both approaches ...

Frequentist definition of Probability

Suppose we perform an experiment (such as tossing a coin) N times. We define the relative frequency of an outcome with attribute A_i (eg 'heads') as

$$\text{relative freq}(A_i) = \frac{\text{N° of outcomes with attribute } A_i}{\text{total N° of experiments}}$$

$$= \frac{n(A_i)}{N}$$

Now (crucially) we define the probability of outcome A_i as

$$P_f(A_i) \equiv \lim_{N \to \infty} \frac{n(A_i)}{N}$$

- a 'frequentist' equates probability to a limiting relative frequency.

Assumptions:

- All experiments are done under the 'same conditions' ['same' meaning delivering the same probability !]
- The limit converges
- Past frequencies predict future frequencies

Combinatorial definition of probability

Suppose we toss a dice. We believe it to be 'fair' so the combinatorial probability of tossing a 6 is

$$P_c = \frac{\text{N° of ways to toss 6}}{\text{N° of possible outcomes}} = \frac{1}{6}$$

Note there is still some circularity in this ('fair' means equally probable), but there is no need for an infinite number of hypothetical trials. We do however need to be able to enumerate the total number of possible outcomes.

Laplace & Bernoulli showed that, as $N \to \infty$ there is a good probability that the relative frequency of an event is close to its combinatorial P_c.

Points to note:

- With the appropriate interpretation of 'p', both P_f and P_c obey the sum & product rules already derived
- Both P_f and P_c are valid starting points for assigning 'degrees of belief' in Bayesian probability theory.

Example

There are two urns, each containing red & white balls.

- urn 1 contains 3 red & 7 white
- urn 2 contains 6 red & 4 white

Blindfold, you take a ball from an urn.

What is the probability: that it is red?

: that it came from urn 1, given it was red?

Ans

We can use Bayes' theorem, assigning probabilities combinatorially:

$$\underbrace{p(R|1) = \frac{3}{10}}_{\text{probability of drawing red, given it's from urn 1}} \; ; \; p(R|2) = \frac{6}{10} \; ; \; \underbrace{p(1) = p(2) = \frac{1}{2}}_{\text{we don't know which urn}}$$

$$p(R) = p(R|1)p(1) + p(R|2)p(2) = \frac{3}{10} \cdot \frac{1}{2} + \frac{6}{10} \cdot \frac{1}{2} = \underline{0.45}$$

From Bayes Theorem

$$p(1|R) = \frac{p(1)\,p(R|1)}{p(R)} = \frac{\frac{1}{2} \cdot \frac{3}{10}}{0.45} = \underline{0.3}$$

Probability Distributions

So far we have looked at situations that are true or false, with the probability distributed between the two options (BPT) or with a frequency of heads/tails (FPT). What if there are many outcomes to consider? (eg, the number of photons arriving in a time τ from a source)?

- We distribute the probability around the options, defining $\underbrace{p(0)}_{\text{prob of zero photons}}$, $p(1)$, $p(2)$ .... $p(\infty)$, with $\sum_{i=0}^{\infty} p(i) = 1$ ... etc

- In frequentist probability theory (FPT) an observed event with several possible outcomes is called a random event. If the outcome is a measurable quantity [eg length, flux, counts etc] it is called a random variable. (RV)

[In BPT there is no such idea. Randomness ≡ lack of knowledge]

- If an RV can take only a finite number of values it is a "discrete random variable". We can associate a probability $p(r)$ with each outcome r.

The set of all $p(r)$ is called the probability distribution of the discrete random variable r.

Poisson Distribution

A Poisson RV obeys the following:

1) The probability of an event occuring in a time interval τ is independent of any past events.
2) Events occur at an intrinsic rate, μ, such that the probability of a single event in time δt is $\mu \delta t$
3) The probability of two (or more) events happening at the same time is zero.

It is a discrete distribution because we deal with counts.

$$p(N|\tau,\mu) = \frac{(\mu\tau)^N}{N!} e^{-\mu\tau}$$

probability of seeing N events in time τ, given μ

See handout for the proof.

Note that the random variable could be counts over space (rather than time), such as the probability of finding N galaxies in a particular patch of sky. In that situation, μ would have dimensions of $1/\text{solid angle}$

Example

On an image of the sky, stars are distributed randomly with each star, on average, per solid angle Ω.
Drawing a circle on the image of solid angle 6Ω what is the probability:

a) that it contains exactly 6 stars?

b) that it contains 10 stars or fewer?

Ans:

Use Poisson dist^n with 'μ' $= \frac{1}{\Omega}$, 'τ' $= 6\Omega$

$$p(N) = \frac{\left(\frac{1}{\Omega} \cdot 6\Omega\right)^N \; e^{-\frac{1}{\Omega} \cdot 6\Omega}}{N!}$$

$$= \frac{6^N}{N!} e^{-6}$$

a) $$p(6) = \frac{6^6}{6!} e^{-6} = 0.16$$

b) This is an example of a cumulative probability.

$$p(N \leq 10) = p(0) + p(1) + \dots + p(9) + p(10)$$

(simply derived from the 'extended sum rule' for mutually exclusive propositions)

N	$p(N) = \frac{6^N}{N!} e^{-6}$
0	0.0024
1	0.0148
2	0.0446
3	0.0892
4	0.1338
5	0.1606
6	0.1606
7	0.1377
8	0.1033
9	0.0688
10	0.0413
	0.8883

$$\Rightarrow p(N \leq 10) \approx 0.89$$

(and $p(N \geq 11) \approx 1 - 0.89 = 0.11$)

Continuous Distributions

What if the random variable can have any real value (in some interval), such as mass, redshift, angle etc? If the random variable is X we define the probability density function, $p(x)$, so that

$$\text{prob}(X \text{ is in the range } x \to x+dx) = p(x)\,dx$$

- Note that $p(x)$ is not a probability. $p(x)dx$ is. We will denote probability by P, pdf by p.

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

- The probability that X lies between a & b is
$$P(a<X<b) = \int_a^b p(x)\,dx$$

- The cumulative probability that $X<a$ is
$$P(X<a) = \int_{-\infty}^{a} p(x)\,dx$$

This is the cumulative distribution function (CDF) of X.

- Note that Bayes' Theorem looks the same when using pdfs:

$$p(x|y)\,dx = \frac{p(x)\,dx\;\; p(y|x)\,dy}{p(y)\,dy}$$

$$\Rightarrow \quad p(x|y) = \frac{p(x)\;p(y|x)}{p(y)}$$

we will use this a lot later...

The Uniform Distribution

- the simplest example of a continuous distribution
- describes a random variable with a uniform probability in some interval $a < x < b$, ie

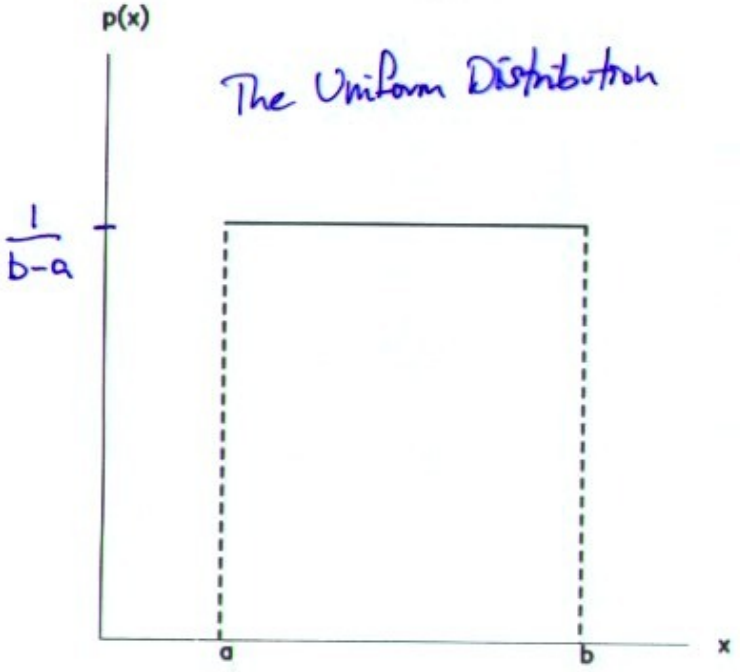
$$p(x) = \begin{cases} 1/(b-a) & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

[note that this is correctly normalised: $\int_{-\infty}^{\infty} p(x)\,dx = 1$]

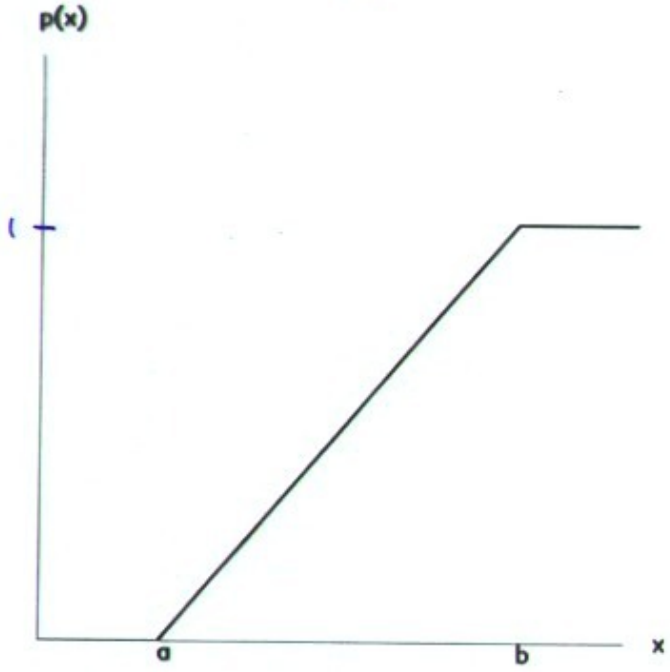
The CDF is

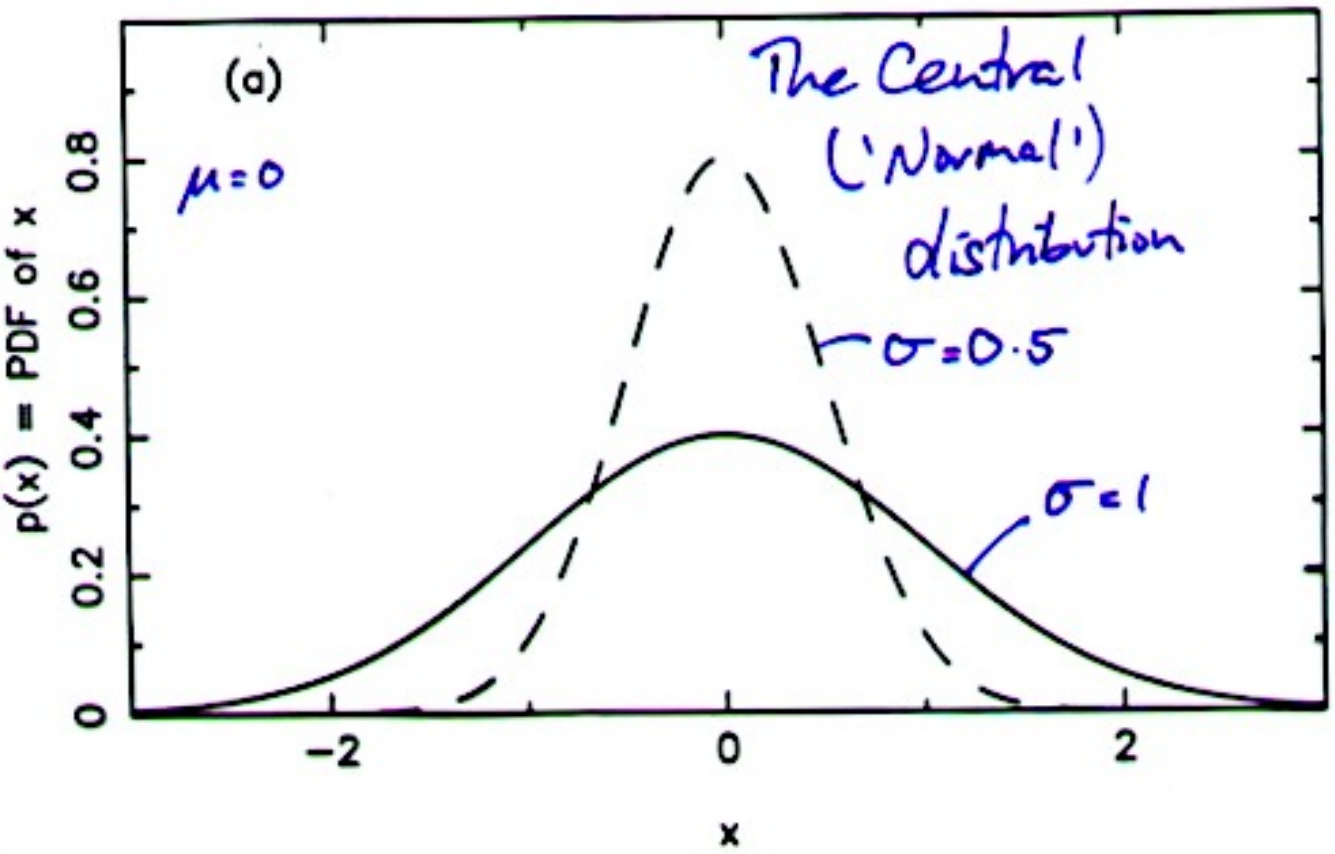
$$P(X < x') = \begin{cases} 0 & x' < a \\ (x'-a)/(b-a) & a < x' < b \\ 1 & x' \geq b \end{cases}$$

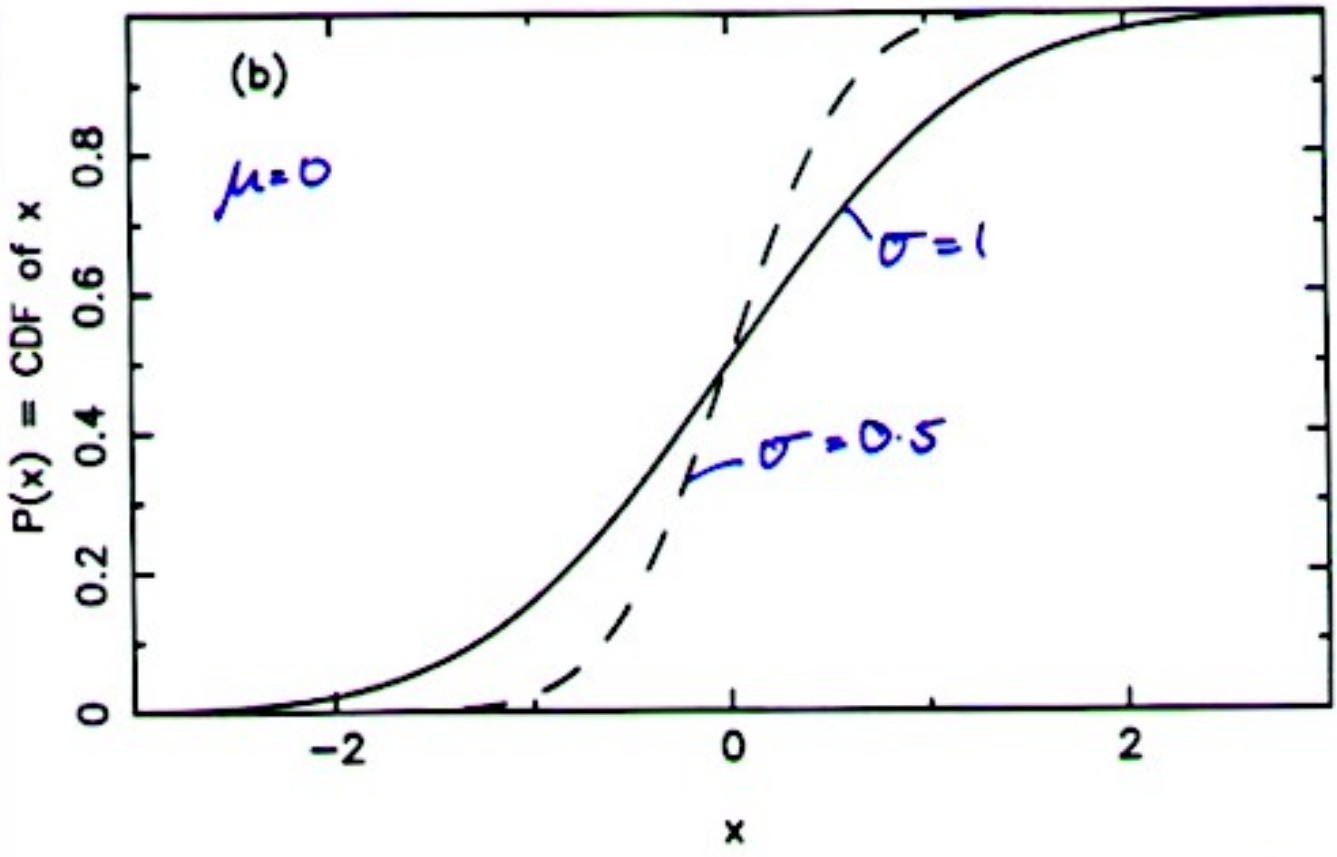
PDF



CDF




(a)
μ=0
The Central ('Normal') distribution
σ=0.5
σ=1
p(x) ≡ PDF of x
0
0.2
0.4
0.6
0.8
−2
0
2
x

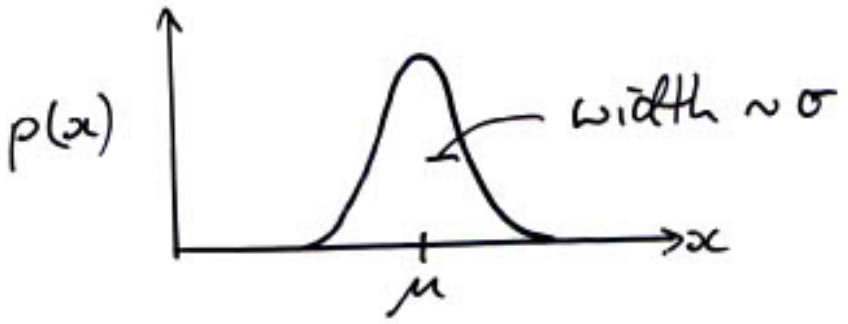

(b)
μ=0
σ=1
σ=0.5
P(x) ≡ CDF of x
0
0.2
0.4
0.6
0.8
−2
0
2
x

The Central Distribution

Also called (for no good reason), the Normal or Gaussian distribution.

It has the pdf:

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



- This is a 'bell-shaped' curve, symmetrical about $x = \mu$ and with a width $\sim \sigma$

[strictly, when $(x-\mu) = \sigma$, $p = p_{max} \cdot e^{-1/2} \simeq 0.6\, p_{max}$]

- There is no analytic form of the CDF, $\Phi(t)$:

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

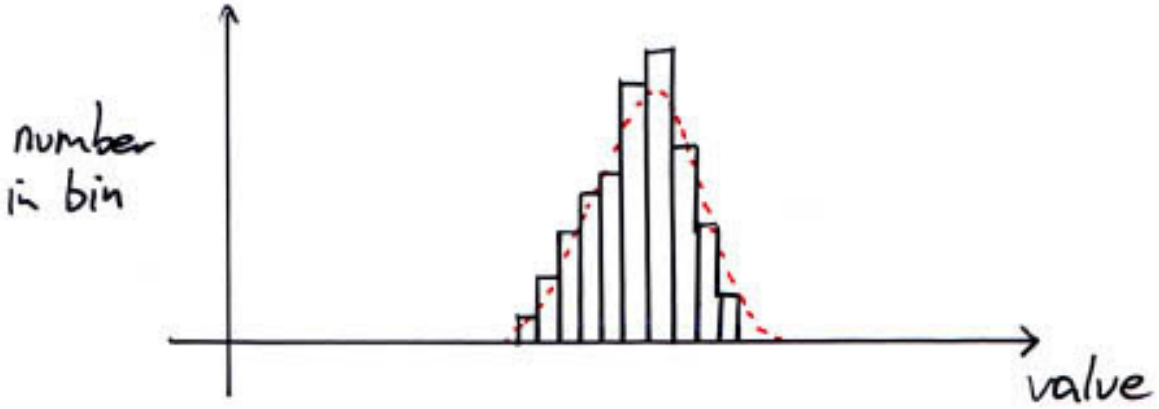
but it is usefully expressed in terms of the error function:

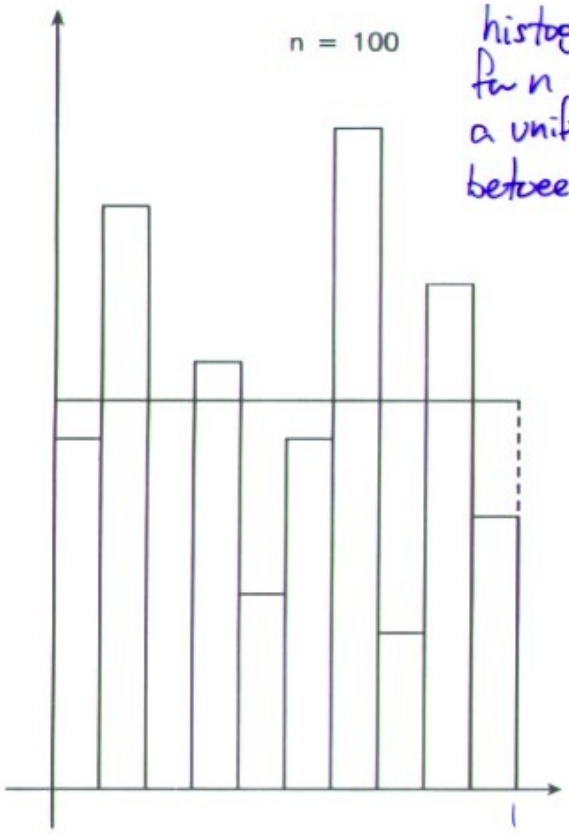
$$\mathrm{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-x^2}\, dx$$

- The Central distribution is very important
  1) because of the Central Limit theorem (see later)
  2) because it represents the correct Bayesian pdf when we know only the variance (& mean) of the noise in our measurement (see late too!)

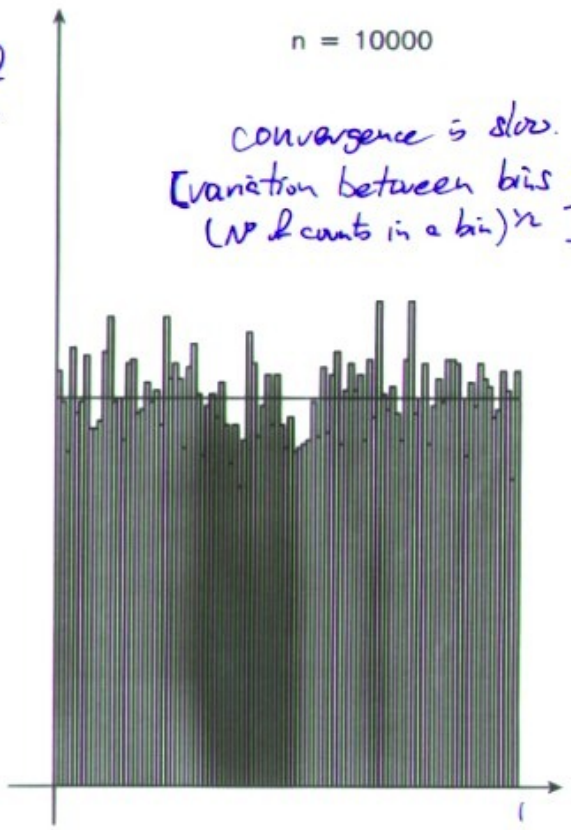
How do we determine pdfs?

- The most intuitive approach is to make repeated measurements of a R.V. and produce a histogram of the results.
  1) Divide the range of results into 'bins' of equal width
  2) Count the number of results that fall into each bin
  3) Plot as a graph, normalising so the total area = 1.





histograms for n trials of a uniform RV between 0, 1.



convergence is slow.
[variation between bins ≃ (№ of counts in a bin)$^{1/2}$]

- For a sufficiently large number of trials (+ bins), the histogram will approach the (frequentist) pdf for the R.V.

- From a Bayesian point of view, the pdf represents our state of knowledge about the 'noise': a uniform pdf represents no preference for one value over another. A Central pdf represents knowledge of only its mean + variance, etc...

Central distribution as a limiting distribution

If we take our Poisson distribution:

$$p(N) = \exp(-\lambda) \cdot \frac{\lambda^N}{N!} \qquad (\lambda = \mu\tau)$$

then for $N \gg 1$, it can be shown that

$$p(N) \simeq \frac{1}{\sqrt{2\pi\lambda}} \exp\left[-\frac{(N-\lambda)^2}{2\lambda}\right]$$

which is Central, with $\mu = \lambda$ and $\sigma = \sqrt{\lambda}$

Measures & 'Moments' of a distribution

The r^{th} moment of a pdf $p(x)$ is defined as:

$$\langle x^r \rangle = \int_{-\infty}^{\infty} x^r p(x)\, dx$$

- The 1st moment ($r=1$) is $\langle x \rangle = \int_{-\infty}^{\infty} x\, p(x)\, dx$.
  It is called the mean of the R.V. or sometimes the expectation value of the R.V. X, $E(X)$.

- The 2nd moment ($r=2$) is $\langle x^2 \rangle = \int_{-\infty}^{\infty} x^2 p(x)\, dx$.
  It is called the mean square of the R.V.

- The variance of the R.V. is defined as

$$\text{var}[X] = \int_{-\infty}^{\infty} (x - \langle x \rangle)^2 p(x)\, dx$$

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

$$= \langle x^2 \rangle - \langle 2x \langle x \rangle \rangle + \langle x \rangle^2$$

$$= \langle x^2 \rangle - 2\langle x \rangle \langle x \rangle + \langle x \rangle^2$$

$$= \langle x^2 \rangle - \langle x \rangle^2$$

and is the 'mean square deviation from the mean'.

- The variance is often written σ^2, and $\sigma = \sqrt{\sigma^2}$ is called the standard deviation

Examples

Poisson: $p(x) = \frac{e^{-\lambda}(\lambda)^x}{x!}$: mean $= \lambda$, var $= \lambda$

Uniform: $p(x) = \frac{1}{b-a}$, $a < x < b$: mean $= \frac{(a+b)}{2}$

var $= \frac{(b-a)^2}{12}$

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

mean $= \mu$

var $= \sigma^2$

- Note that not all pdfs have defined moments, eg the Cauchy distribution,

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

does not have a variance, as the integral $\int_{-\infty}^{\infty} p(x) \cdot x^2 \, dx$ diverges.

- We can define other measures :

The median of the R.V. divides the cdf into 2 equal halves, so that

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

[ie the R.V. is equally likely to be $> x_{med}$ as $< x_{med}$.]

The mode of the R.V. corresponds to its most probable value, ie $p_{max} = p(x_{mode})$, if such a single value exists.

The skewness and kurtosis of X can also be defined

↑ 'lobsidedness' ↑ 'lingyness'

(but are not much use!)

Changing variable

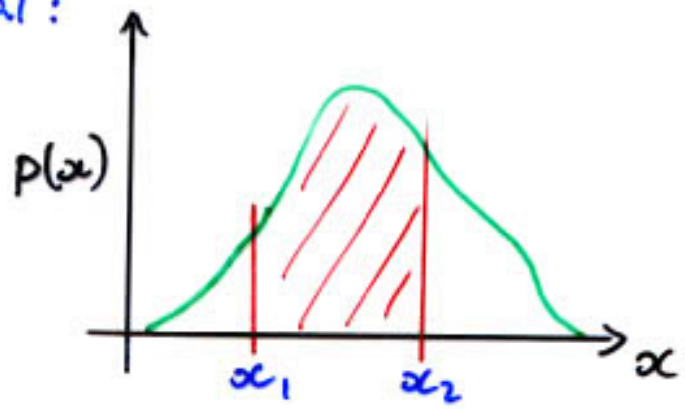
Imagine we know the pdf of some variable, x. What is the pdf of a function of x (say z)?

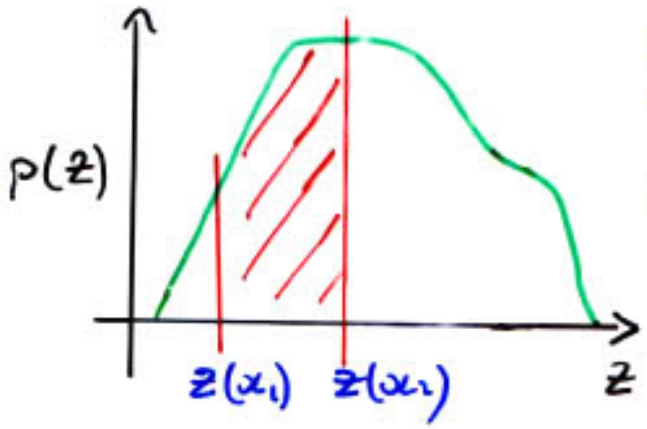
ie $p(x)$ known - the pdf

$z(x)$ known - the new variable

what is $p(z)$?

- The trick is to realise that the probability contained in equivalent intervals must be identical:





The hatched areas must be the same (if x maps uniquely to z)

ie $$\left| \int_{x_1}^{x_2} p(x)\,dx \right| = \left| \int_{z(x_1)}^{z(x_2)} p(z)\,dz \right|$$

[modulus to ensure integrals are +ve]

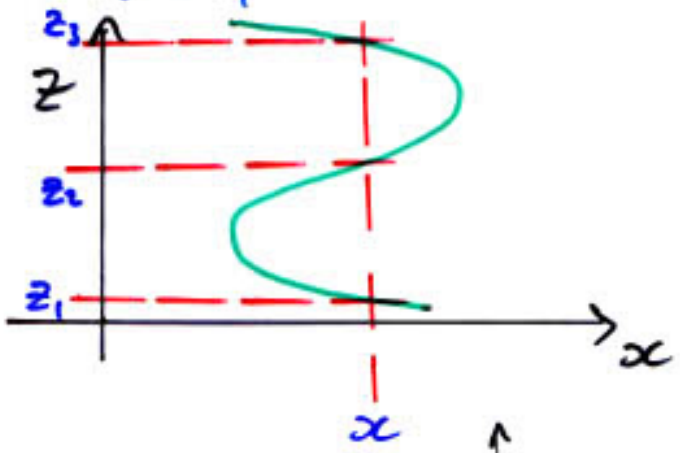
- For a small interval in x δx, with corresponding interval in z δz, we can write:

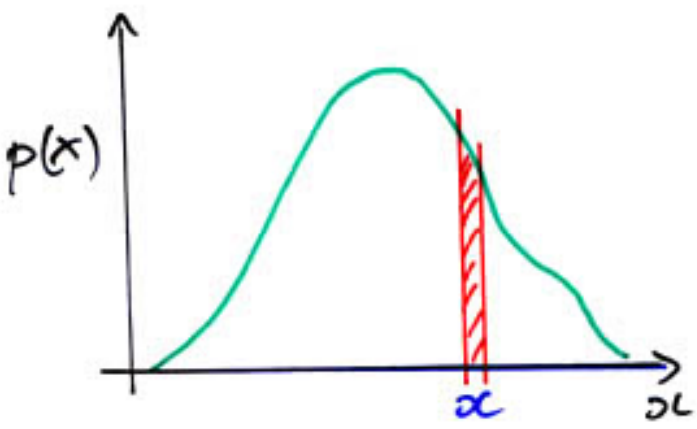
$$p(x)\,|\delta x| = p(z)\,|\delta z|$$

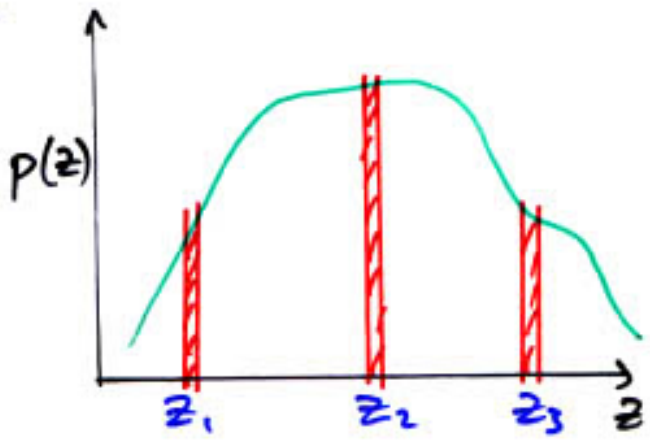
so in the limit

$$\boxed{p(z) = p(x)\left|\frac{dx}{dz}\right|}$$

- Note that things get trickier if x does not map uniquely to z







Extension to many dimensions

- Say you have a probability distribution in many dimensions

$$p(x_1, x_2, \ldots, x_n)$$

and you need to change parameters to $(y_1, y_2 \ldots, y_n)$.
If the mapping $x \to y$ is unique then

$$p(y_1, y_2, \ldots, y_n) = p(x_1, x_2, \ldots, x_n)\,|J|$$

$|J|$: "Jacobian determinant"

where

$$J = \frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(y_1, y_2, \ldots, y_n)}$$

is a matrix with elements $J_{ij} = \dfrac{\partial x_i}{\partial y_i}$

eg: change co-ordinates from (x, y, z) to (r, θ, ϕ) (spherical polars)

$$y_1 = r \qquad x_1 = x = r\cos\phi\sin\theta$$
$$y_2 = \theta \quad ; \quad x_2 = y = r\sin\phi\sin\theta$$
$$y_3 = \phi \qquad x_3 = z = r\cos\theta$$

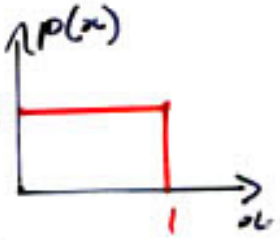
$$J = \begin{pmatrix} c\phi\, s\theta & r c\phi c\theta & -r s\phi s\theta \\ s\theta s\phi & r s\phi c\theta & r c\phi s\theta \\ c\theta & -r s\theta & 0 \end{pmatrix}$$

and $|J| = r^2 \sin\theta$, so $\underline{p(r,\theta,\phi) = p(x,y,z) r^2 \sin\theta}$

Example: The Exponential distribution.

- Computer random number generators usually produce uniform distributions:

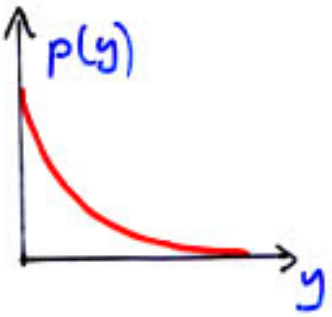
$$p(\alpha)\, d\alpha = \begin{cases} d\alpha & 0 < \alpha < 1 \\ 0 & \text{otherwise} \end{cases}$$



- Choose $y(\alpha) \equiv -\ln\alpha$, $\Rightarrow \alpha = e^{-y(\alpha)}$

then $p(y)\, dy = \underbrace{p(\alpha)}_{=1} \underbrace{\left|\frac{d\alpha}{dy}\right|}_{=e^{-y}} dy$

so $p(y) = e^{-y}$



- so if X is uniformally distributed, $Y = -\ln X$ will have an exponential distribution.

Multivariate distributions

The idea of a pdf extends smoothly into more than one variable - multivariate distributions.

Joint pdfs

The joint pdf of two quantities x_1 and x_2 is defined as

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

and extensions to higher dimensions follow similarly.

Marginal distributions

We can recover the pdf of a single variable by **marginalising** (integrating) over the others

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

This is an important part of Bayesian analysis - see later...

Statistical Independence

- X_1 and X_2 are independent RVs if, and only if, their joint pdf can be written as the product of their marginal pdfs. ie

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

Put another way, if x_1 & x_2 are independent:

$$p_1(x_1) = p(x_1/x_2)$$ [knowing x_2 does not affect $p(x_1)$]

But $p(x_1, x_2) = p_2(x_2)\, p(x_1/x_2)$ [product rule]

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

For n independent RVs

$$p(x_1, x_2 \ldots, x_n) = \prod_{i=1}^{n} p_i(x_i)$$

[note, the subscript in p_i is often dropped]

The Bivariate Normal Distribution

We have seen that the Normal (central) dist^{n} has the form:

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

where μ is the mean, & σ^2 the variance of the dist^{n}.

- For two <u>independent</u> quantities x_1 & x_2 :

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2} \right]$$

putting $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$; $\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$; $\underline{\underline{A}} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix}$

we can write this as

$$p(\underline{x}) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{1}{2} (\underline{x}-\underline{\mu})\, \underline{\underline{A}}\, (\underline{x}-\underline{\mu})^T \right]$$

(^{T} = transpose : $\underline{x}^T = (x_1, x_2)$)

Note also that $\underline{\underline{A}}^{-1} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$ (check: $\underline{\underline{A}}\,\underline{\underline{A}}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ ✓)

So if we put $\underline{\underline{C}} = \underline{\underline{A}}^{-1}$:

Then $p(\underline{x}) = \frac{1}{2\pi \left(\det(\underline{\underline{C}})\right)^{1/2}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu})\underline{\underline{C}}^{-1}(\underline{x}-\underline{\mu})^T\right]$

This can be easily extended to n independent quantities with $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ etc. Note the 2π becomes $(2\pi)^{n/2}$.

If $n=2$ we have the bivariate Normal distribution

" $n>2$ " " " multivariate " "

- $\underline{\underline{C}}$ is called the covariance matrix of the variables,

If the variables are independent, it is diagonal, but our definition can be extended to accommodate variables that are not independent (ie, that are correlated)

$$\underline{\underline{C}} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \quad \text{where } \sigma_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$$

clearly, $\sigma_{11} = \langle x_1^2 \rangle - \langle x_1 \rangle^2 = \sigma_1^2$ (sim $\sigma_{22} = \sigma_2^2$)

$\sigma_{12} = \sigma_{21}$ = the covariance of x_1 and x_2

We can also define the correlation coefficient between x_1 and x_2 as

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

We can now write the general bivariate Normal distribution as:

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

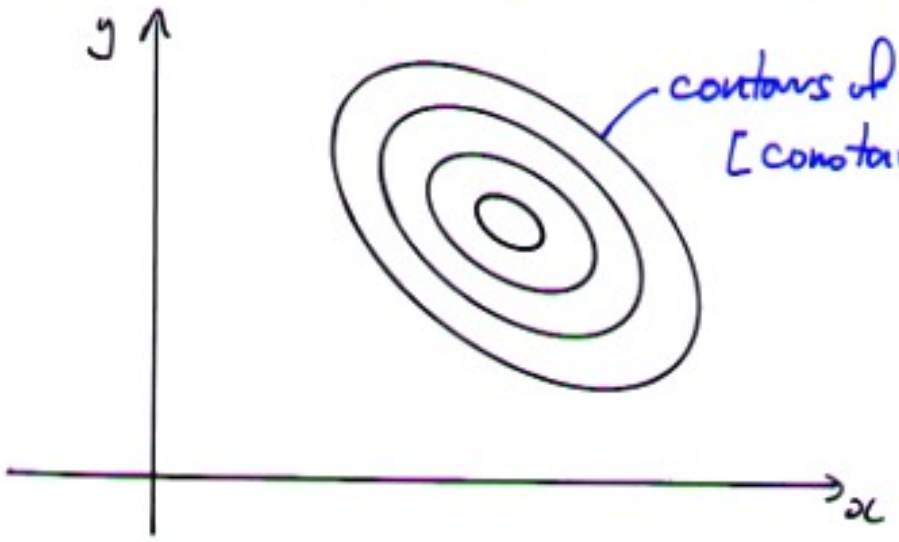
where

$$Q(x,y) = \left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2$$

- The dist^n is defined by five parameters, $\mu_x, \sigma_x, \mu_y, \sigma_y, \rho$.
- The marginal pdfs of X and Y are just univariate Normal:

eg $$p_x(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right] \qquad \left(= \int p(x,y)\,dy\right)$$

- We can plot these joint pdfs as contour plots:



Here, X and Y are anticorrelated ($\rho < 0$)

Samples and Parents

- We have used terms like 'mean', 'variance' and 'covariance' to parameterise pdfs. These parameters define the pdf of a particular quantity.

- We can also define approximations to these parameters, as determined from a finite set of data, ie

  Let there be n measurements of RV x:

  "mean of x" $= \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$

  "variance of x" $= \frac{1}{n}\sum x_i^2 - \left(\frac{1}{n}\sum x_i\right)^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$

- To prevent confusion, we call these the **sample mean** and the **sample variance** (etc), as they are computed just from a sample of data.

- The 'true' parameters of the pdf are called the **distribution mean** (etc) or the **parent mean** (etc.)

  We would expect $\bar{x}$ to approach μ as $n \to \infty$ (etc).

Parameter Estimation

Often (very often!) we are presented with the task of estimating something using data that contain random errors & other uncertainties. The rest of the course is on how to do this well!

Eg: "A2 statistics"

Imagine we make n independent measurements of some quantity under identical conditions (we believe). We can compute the sample mean, $\bar{x}$, and variance, σ^2, of our measurements:

$$\bar{x} = \frac{1}{n}\sum x_i \quad ; \quad \sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

Here, σ is our estimate of the 'error' on each measurement and when we average n of them we would expect the sample mean to deviate from the true mean only by about $\frac{\sigma}{\sqrt{n}}$, ie

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{n}}$$

μ: 'true' (parent) mean – the quantity we wish to estimate.

Most people would interpret this as meaning that μ lies between $\left(\bar{x} - \frac{\sigma}{\sqrt{n}}\right)$ and $\left(\bar{x} + \frac{\sigma}{\sqrt{n}}\right)$ 68% of the time.

ie $$P\left(\bar{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}}\right) = 68\% \qquad \text{BPT}$$

Such people are natural "Bayesians" – they talk about the probability of a parameter (μ) as a description of our state of knowledge of its value.

- But in frequentist statistics, probabilities can only be assigned to random variables, and μ is definitely NOT a random variable (it is a constant). Instead we think in terms of the random variable that is our sample mean, $\bar{X}$. ($\bar{x}$ is one value of $\bar{X}$) and say

$$P\left(\mu - \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + \frac{\sigma}{\sqrt{n}}\right) = 68\% \qquad \text{FPT}$$

which is a probabilistic statement about $\bar{X}$, given μ, σ and n! Despite the fact this is not the statement we intuitively want, it is all 'conventional' (non-Bayesian) statistics can give us, and much complication can follow from this.

Bayes' Theorem Revisited

- We have already met Bayes' Theorem:

$$P(X|YI) = \frac{P(Y|XI) \cdot P(X|I)}{P(Y|I)}$$

How does it help us in parameter estimation?

- In BPT we assign probabilities to statements, reflecting our degree of belief in the statements. We can therefore assign probabilities to hypotheses, such as

"The Moon has a mass of 6×10^{22} kg"

or

"The Moon has a mass of 7×10^{22} kg"

So replace:

X with a hypothesis

Y with some data

I with other relevant background information

we now have

$$P(\text{hypothesis}|\text{data}, I) = \frac{P(\text{data}|\text{hypothesis}, I) \cdot P(\text{hypothesis}|I)}{P(\text{data}|I)}$$

PONDER THIS EXPRESSION

- If we have some new data, we are interested in how it affects our degree of belief in a particular hypothesis (eg about the Moon's mass)

ie we are interested in $P(\text{hypothesis} \mid \text{data}, I)$

["the probability of the hypothesis given the data"]

- called the posterior probability.

- $P(\text{hypothesis} \mid I)$ is called the prior probability - our state of knowledge prior to acquiring the data.

- $P(\text{data} \mid \text{hypothesis}, I)$ is called the likelihood of the hypothesis ie, the probability we would have seen what we did see, assuming the validity of the hypothesis.

- $P(\text{data} \mid I)$ is called the evidence. It is a constant for all hypotheses, so does not help us in choosing between them. We can always determine it by normalising the posterior probability to 1.

So we have:

posterior probability $\propto$ prior probability $\times$ likelihood

"what we know now" — "what we did know" — "influence of the data"

So Bayes' Theorem encapsulates the concept of learning
- Our new state of knowledge is based on what we knew before, but modulated by new data

Q: Why do we need the prior probability?

A: Imagine a measurement of the Moon's mass that came in at 100kg, and consider the hypothesis "The Moon has a mass of 100 kg".

Here $P(\text{data} \mid \text{hypothesis}, I) \simeq 1$ — this measurement would be very probable if the hypothesis were true.

But we have a lot of previous data on the Moon indicating its mass is $\gg 100$kg, independent of this measurement: ie prior probability $P(\text{hypothesis} \mid I) \ll 1$. Hence posterior probability $P(\text{hypothesis} \mid \text{data}, I) \ll 1$ too, but will have grown a little in comparison to other hypotheses.

Bayes' Theorem is often described as 'common sense reduced to mathematics'

Example - Is it a fair coin? [From Sivia, 1996]

Suppose a coin could be made with a bias, H, so that when flipped the probability that it comes up heads is H.

$H = 0$ represents a two-tailed coin!
$H = \frac{1}{2}$ " " fair "
$H = 1$ " " two-headed coin,

and H can have any value between 0 & 1.

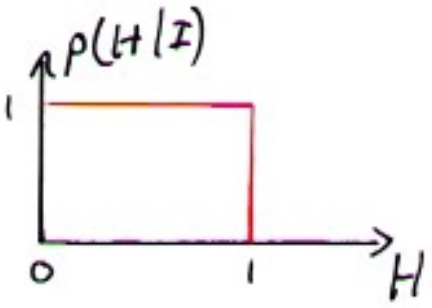
We are given a coin with an unknown bias, H, and flip it a few times. How does our state of knowledge of H evolve during this process?

Solution

We can express our state of knowledge of H as a probability distribution function $p(H \mid \underbrace{\text{data}}_{\text{results of flips}}, I)$

At the start we are totally ignorant about H, and cannot favour any one value, so the prior pdf for H is uniform:

ie $$p(H|I) = \begin{cases} 1 & 0 \le H \le 1 \\ 0 & \text{otherwise} \end{cases}$$



[this is correctly normalised: $\int_{-\infty}^{\infty} p(H|I)\, dH = 1$]

Suppose we know the bias, H, and flip the coin N times.
The probability of obtaining R heads with these flips is

$$p(R \text{ heads} \mid H, I) = \underbrace{\frac{N!}{R!(N-R)!}}_{\text{number of permutations}} \underbrace{H^R (1-H)^{N-R}}_{\substack{\text{prob of } R \text{ heads} \\ \text{followed by } N-R \text{ tails}}}$$

- the binomial distribution.

As we are only interested in terms containing H, write

$$p(\underbrace{R \text{ heads}}_{\substack{\text{experimental} \\ \text{data}}} \mid H, I) \propto H^R (1-H)^{N-R}$$

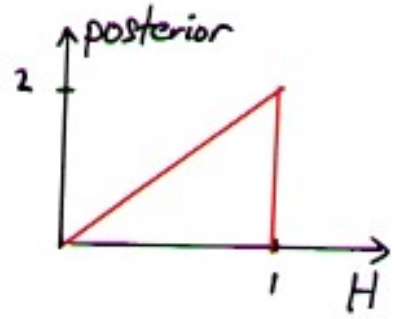
Now Bayes' Theorem tells us

$$\underbrace{p(H \mid \text{result of flip}, I)}_{\text{posterior}} \propto \underbrace{p(H|I)}_{\text{prior}} \cdot \underbrace{p(\text{result of flip} \mid H, I)}_{\text{likelihood}}$$

If the 1st flip comes up heads, then $N=1$, $R=1$ and the 'likelihood' is $\propto H^1(1-H)^0 = H$, and our posterior is

$$p(H \mid 1^{st}\ result, I) \propto 1 \cdot H$$

(1: uniform prior; H: likelihood)



We can infer that:

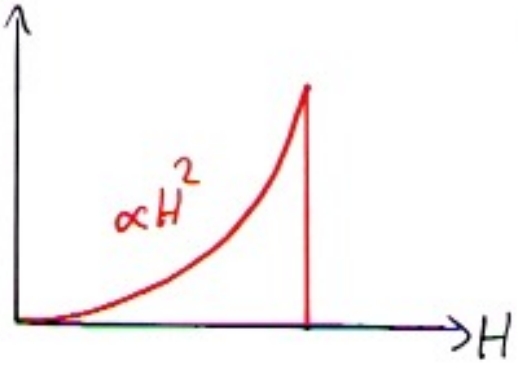
- the probability it has 2 tails is zero [$p(0)=0$]
- at this stage it's most probable the coin is 2-headed but our confidence rises steadily from $H=0$ to $H=1$.

We now flip the coin again. It comes up heads AGAIN! The likelihood of both these results, for a particular H, is $\propto H^2(1-H)^0 = H^2$

$$\text{so } p(H \mid 1^{st} \text{ and } 2^{nd}\ result, I) \propto 1 \cdot H^2$$

(1: uniform prior; H^2: likelihood)

posterior after the 2 flips



- we are more inclined than ever to believe this is a 2-headed coin!

But we could have got this result another way, by treating the 2nd flip on its own and using the posterior of the 1st experiment as the prior for the second:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

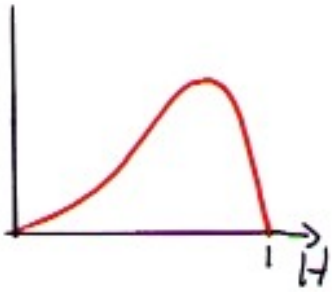
$$\propto \underbrace{H}_{\substack{\text{our state of knowledge} \\ \text{after the 1st flip}}} \times \underbrace{H}_{\substack{\text{new data from} \\ \text{2nd flip, treated} \\ \text{on its own.}}} \propto H^2 \text{ again}$$

- So we can do the analysis in one-step or treat it as a sequential learning process. We get the same answer.
- If the 3rd flip is a tail, this on its own has a likelihood

$$p(\text{tail} \mid H, I) \propto (1-H)$$

so the new posterior for H is

$$p(H \mid 3 \text{ flips}, I) \propto \underbrace{H^2}_{\substack{\text{prior, before} \\ \text{3rd flip}}} (1-H)$$



See handout for further trials.

Q: How is our posterior distribution for H, after many flips, dependent on our prior?

A: If our prior is not too insistant (ie if we are uncertain of the true value of H at the start), then our final pdf for H will be ~ independent of the prior. – the information in a large number of flips is greater than our prior knowledge of the value of H, and overwhelms it.

General Procedure for Parameter Estimation (Bayesian)

- The above is an example of a general type of problem:
  - We have a model for how the World works that depends on one (or more) parameters
    [our model was the binomial distribution, and our parameter was H]
  - The model tells us the likelihood of any given set of data $\{D_k\}$, given a value for the parameter

model independent assumptions.

$$\text{Likelihood} = p(\{D_k\} \mid m, I)$$

data — model parameter(s)

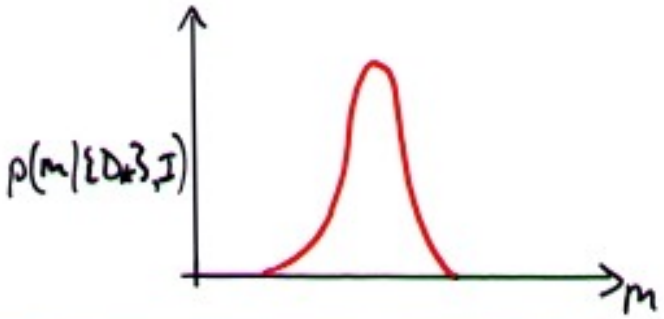
- We take account of the prior probability of our model based on previous experience:

$$\text{Prior} = p(m \mid I)$$

- Bayes' Theorem delivers the (posterior) probability of the model parameters:

$$\text{Posterior} = p(m \mid \{D_k\}, I) \propto p(m \mid I) \cdot p(\{D_k\} \mid m, I)$$

which we can normalise.

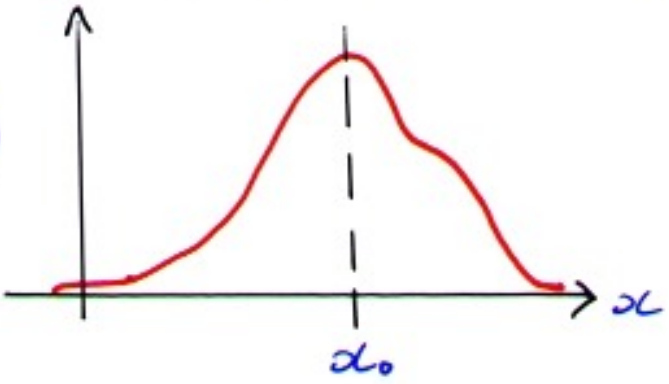


this pdf contains all that BPT has to say about the value of m.

Often we wish to summarise it with 2 numbers: the 'best estimate' and its 'uncertainty'....

Best Estimates and Error Bars

The posterior pdf for some model parameter x can have any shape, but is usually peaked.



The most probable value of x is x_0

$$p(x_0|\text{data}) = p_{max}$$

and $$\left.\frac{dp}{dx}\right|_{x_0} = 0 \quad ; \quad \left.\frac{d^2p}{dx^2}\right|_{x_0} < 0$$

- We could expand p in a Taylor series about x_0, but we usually need fewer terms for the same accuracy using

$$L = \ln[p(x|\text{data})]$$

- the Log-posterior probability. Note $\left.\frac{dL}{dx}\right|_{x_0} = 0$ too.

Expanding L about x_0:

$$L(x) = L(x_0) + \frac{1}{2}\left.\frac{d^2L}{dx^2}\right|_{x_0}(x-x_0)^2 + \dots$$

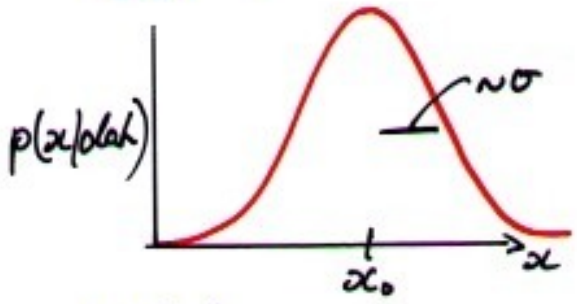
By stopping at the second term we are saying that

$$p(x|data) \simeq \underbrace{c}_{\text{constant}} \cdot \exp\left[\frac{1}{2}\left.\frac{d^2\mathcal{L}}{dx^2}\right|_{x_0}(x-x_0)^2\right]$$

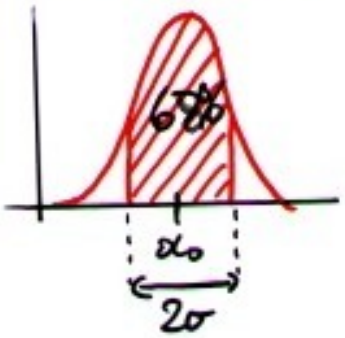
Noting that $\left.\frac{d^2\mathcal{L}}{dx^2}\right|_{x_0}$ is negative, this is a gaussian pdf

$$p(x|data) \simeq \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-x_0)^2}{2\sigma^2}\right]$$

where $\sigma = \left(-\left.\frac{d^2\mathcal{L}}{dx^2}\right|_{x_0}\right)^{-1/2}$



- In a gaussian, about 68% of the probability lies within σ of x_0,



so we can say

$$\text{prob}\left[x_0-\sigma \leq x \leq x_0+\sigma \,|\, data\right] = 0.68$$

σ represents our 68% error bar in this case.

Example

We can apply this gaussian approximation to our coin example:

$$p(H\,|\,\{data\}, I) \propto H^R (1-H)^{N-R}$$

[posterior probability of H, given R Heads in N flips, using a uniform prior for H]

$$L = \ln p = R \ln H + (N-R)\ln(1-H) + \text{const}$$

$$\frac{dL}{dH} = \frac{R}{H} - \frac{(N-R)}{(1-H)}$$

and $$\frac{d^2L}{dH^2} = -\frac{R}{H^2} - \frac{(N-R)}{(1-H)^2}$$

- Our most probable value of H, H_0, occurs when $\left.\frac{dL}{dH}\right|_{H_0} = 0$

i.e. $$0 = \frac{R}{H_0} - \frac{(N-R)}{(1-H_0)}$$

$$H_0 = \frac{R}{N}$$

- this is just the relative frequency of heads (very reasonable)

Also, we have

$$\sigma \simeq \left(- \frac{d^2 L}{dH^2} \Big|_{H_0} \right)^{-1/2}$$

ie $$\sigma = \left[\frac{H_0 (1 - H_0)}{N} \right]^{1/2}$$

Let's look at this:

- For fixed N we are most sure of our value of H when $H_0 \simeq 0$ and $\simeq 1$, ie when the coin is very biased.
  - biased coins are easy to spot.
- Also for fixed N, σ is greatest when $H_0 = 0.5$
  - it's hard to be confident a coin is fair.
- We know that the peak of the posterior pdf (H_0) does not shift much after a few flips, so from that point on:

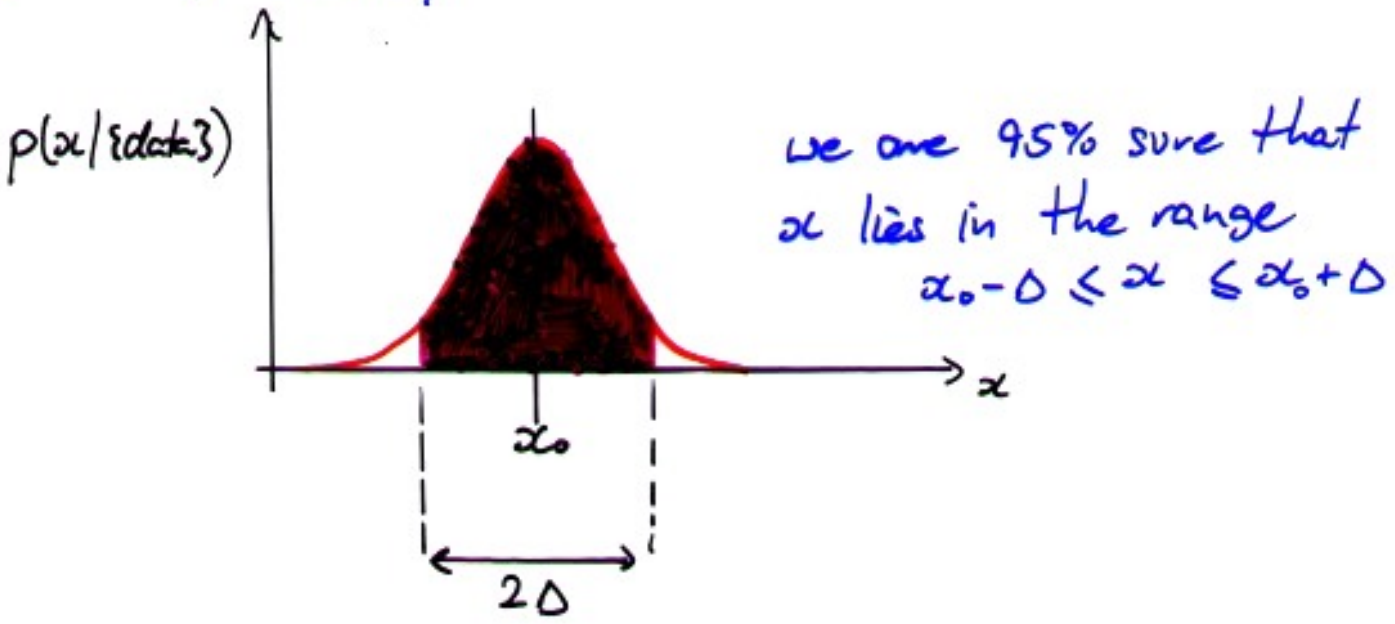
$$\sigma \propto \frac{1}{\sqrt{N}}$$

So our uncertainty in H drops as $1/\sqrt{N}$

- again, very reasonable.

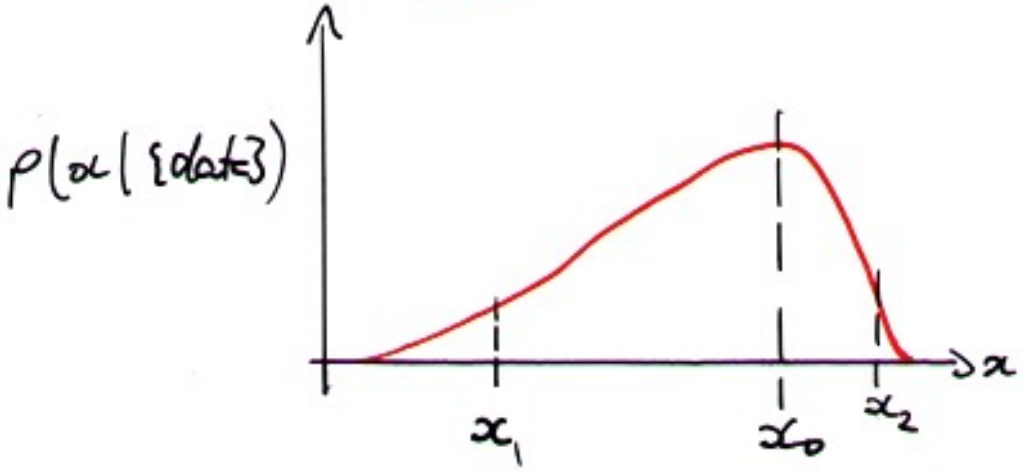
Shortest confidence intervals

- For a symmetric posterior pdf it is easy to define the 'confidence interval' around the most probable value of our parameter that contains, say, 95% of the probability:



[remember, for a gaussian the 68% interval is $x_0-\sigma \leq x \leq x_0+\sigma$]

- for an asymmetric pdf the choice is less clear



There is no obvious choice of x_1 to x_2 that contains 95% of the probability.

But we can define a shortest confidence interval. That is a pair of values x_1, x_2 for which

$$\text{prob}(x_1 \leq x \leq x_2 \mid \{data\}, I) = 0.95$$

and $|x_2 - x_1|$ is a minimum.

- Note that for an asymmetric pdf x_0 is not the mean (expectation) value for x, it is the most probable value.
  - again, only the full posterior pdf contains the whole story.

Gaussian Noise

It is common to be presented with a set of direct measurements of a parameter, contaminated by gaussian noise [why gaussian? - see later]

Eg:- measurements of the flux density of a radio source

- measurements of the mass of the Sun

⋮

and so on.

Let the true value of the parameter be μ, and our k^{th} measurement of it be

$$x_k = \mu + \text{noise}_k$$

We assume the noise is gaussian, with a mean of zero and a variance σ_0^2.

$$\text{noise}_k = x_k - \mu$$

↑ noise in the k^{th} measurement

- Given μ, the probability of making a measurement x_k is just the probability that the 'residual' $x_k - \mu$ can be attributed to the noise, ie

$$p(x_k | \mu, \sigma_0, I) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_k-\mu)^2}{2\sigma_0^2}\right]$$

- If our N measurements $\{x_k\}$ are independent, then their joint probability (the 'likelihood') is:

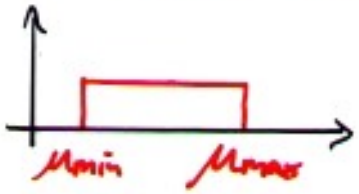
$$p(\{x_k\} | \mu, \sigma_0, I) = \prod_{k=1}^{N} p(x_k | \mu, \sigma_0, I)$$

$$= \frac{1}{\sigma_0^N (2\pi)^{N/2}} \exp\left[-\sum_{i=1}^{N} \frac{(x_k-\mu)^2}{2\sigma_0^2}\right]$$

To turn this into a posterior probability for μ, given the data and σ_0 we need Bayes' theorem and a prior pdf for μ.

Let us choose a uniform prior in the range μ_{min} to μ_{max}:

$$p(\mu|\sigma_0, I) = \begin{cases} \frac{1}{\mu_{max} - \mu_{min}} & \mu_{min} \leq \mu \leq \mu_{max} \\ 0 & \text{otherwise} \end{cases}$$



Now $p(\mu|\{x_k\}, \sigma_0, I) \propto p(\mu|\sigma_0, I) \cdot p(\{x_k\}|\mu, \sigma_0, I)$

and $\ln[p(\mu|\{x_k\}, \sigma_0 I)] = L = -\sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma_0^2} + \text{const}$

↑ terms independent of μ.

- The value of μ that maximises L is its most probable value (given the data & σ_0), μ_0. ie

$$\left.\frac{dL}{d\mu}\right|_{\mu_0} = \sum_{k=1}^{N} \frac{(x_k - \mu_0)}{\sigma_0^2} = 0$$

σ^2 is constant, so $\sum x_k - \sum \mu_0 = 0$

ie $\sum x_k = N\mu_0$

So the most probable value of our parameter is just the mean of the measurements:

$$\mu_0 = \frac{1}{N} \sum_{k=1}^{N} x_k$$

- No great surprise, but we have proved it.

[Note we have assumed gaussian noise, with known, constant, σ_0^2]

- Our standard confidence interval is defined by

$$\left(-\frac{d^2 L}{d\mu^2}\Big|_{\mu_0} \right)^{-1/2} = \left(\sum_{k=1}^{N} \frac{1}{\sigma_0^2} \right)^{-1/2} = \left(\frac{N}{\sigma_0^2} \right)^{-1/2} = \frac{\sigma_0}{\sqrt{N}}$$

So our posterior pdf for μ can be summarised as

$$\mu = \mu_0 \pm \frac{\sigma_0}{\sqrt{N}} \qquad \text{where } \mu_0 = \frac{1}{N} \sum x_k$$

- Again, no great surprise, but now we see where it comes from (and what assumptions it makes)

Data with different error bars

We have just considered the case of data with constant error bars:

What if the errors in the measurements are different?

If the errors are still gaussian, we proceed as before, but each measurement x_k has an associated σ_k:

Now, $$L = -\sum_{k=1}^{N} \frac{(x_k-\mu)^2}{2\sigma_k^2} + \text{const.}$$

The most probable μ (μ_0) satisfies

$$\left.\frac{dL}{d\mu}\right|_{\mu_0} = -\sum_{k=1}^{N} \frac{(x_k-\mu_0)}{\sigma_k^2} = 0$$

$$\text{So } \mu_0 = \frac{\sum_k x_k/\sigma_k^2}{\sum_k 1/\sigma_k^2}$$

This is a weighted mean. Each datum has been loaded by our confidence in it.

- The standard width of our posterior for μ can be calculated as:

$$\text{error in } \mu = \left(-\frac{d^2\mathcal{L}}{d\mu^2}\bigg|_{\mu_0} \right)^{-1/2} = \left(\sum_{k=1}^{N} \frac{1}{\sigma_k^2} \right)^{-1/2}$$

[clearly this reduces to $\frac{\sigma_0}{\sqrt{N}}$ if all the variances are σ_0^2.]

Model Fitting

This approach is easily extended to more sophisticated problems

Example: A neutron star - neutron star binary system can be expected to produce sinusoidal gravitational waves at twice the binary's orbital frequency:

We can model this wave as

$$h(t) = A \sin(\omega t + \phi)$$

$h(t)$: strain measurement; ω, ϕ: frequency & phase known from visual observation

Our task is to determine the amplitude of the waves, A, from the output of a GW detector, $d(t_k)$ [$\{t_k\}$ are the times of observation]

- Let us assume the noise in the detector is gaussian with a known variance σ_0^2.

We proceed as before, but now the likelihood of the data depends on how well it fits the model:

$$p(\{d_k\}|A,\omega,\phi) \propto \prod_k \exp\left[-\frac{(d_k - A\sin(\omega t_k+\phi))^2}{\sigma_0^2}\right]$$

we want the posterior for A:

$$p(A|\{d_k\},\omega,\phi) \propto \underbrace{p(A|\omega,\phi)}_{\text{prior on } A}\ \underbrace{p(\{d_k\}|A,\omega,\phi)}_{\text{likelihood of } \{d_k\}}$$

again, let's take the prior as uniform $\begin{matrix}\text{for } A\geq 0\\ 0 \quad \text{for } A<0\end{matrix}$

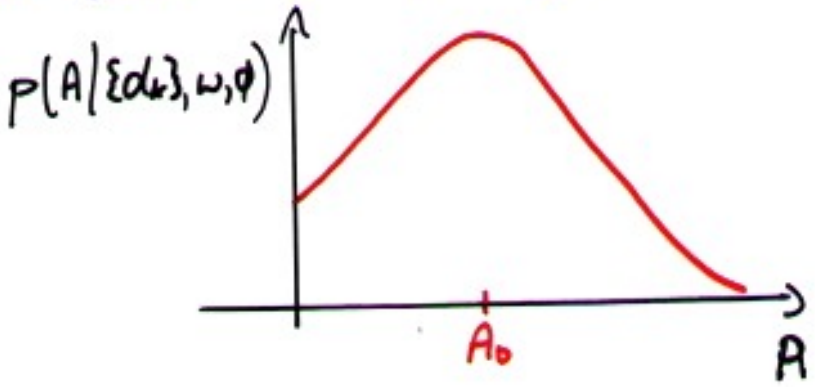
Now $$L = -\sum_{k=1}^{N}\frac{(d_k - A\sin[\omega t_k+\phi])^2}{2\sigma_0^2}$$

Our most probable A (A_0) satisfies

$$\left.\frac{dL}{dA}\right|_{A_0} = 0 \quad \Rightarrow \quad A_0 = \underbrace{\frac{\sum_k d_k \sin(\omega t_k + \phi)}{\sum_k \sin^2(\omega t_k + \phi)}}$$

depends on known quantities.

and our posterior for A may look like:



If the signal is weak.

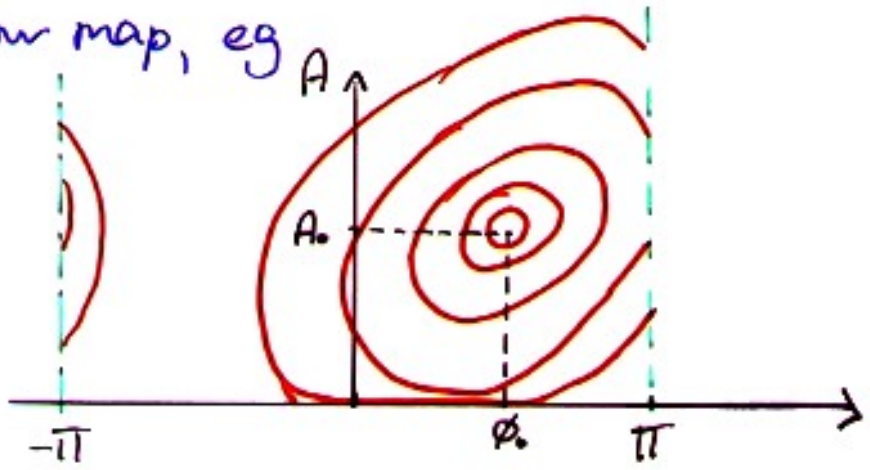
Handling more than one parameter:

What if, in the above, we didn't know A or ϕ?

We can determine a joint posterior for A and ϕ:

$$p(A, \phi \,|\, \{d_k\}, \omega) \propto \underbrace{p(A, \phi \,|\, \omega)}_{\text{joint prior for } A \text{ and } \phi} \underbrace{p(\{d_k\} \,|\, A, \omega, \phi)}_{\text{likelihood (unchanged)}}$$

This is a probability surface that we can represent as a contour map, eg



The most probable value of A is A_0 and of ϕ is ϕ_0.

Note we have used the prior:

$$p(A,\phi|\omega) = p(A|\omega)\,p(\phi|\omega) \quad [A,\phi \text{ independent}]$$

and $p(A|\omega)$ = constant $A>0$, zero otherwise

$p(\phi|\omega)$ = constant $-\pi<\phi<\pi$, zero otherwise.

Marginal Distributions

Suppose we are not interested in ϕ_0. We can determine the posterior pdf for A alone if we marginalise over ϕ.

Remember that $p(X|I) = \int_{-\infty}^{\infty} p(X,Y|I)\,dY$

so $p(A|\{d_k\},\omega) = \int_{-\infty}^{\infty} p(A,\phi|\{d_k\},\omega)\,d\phi$

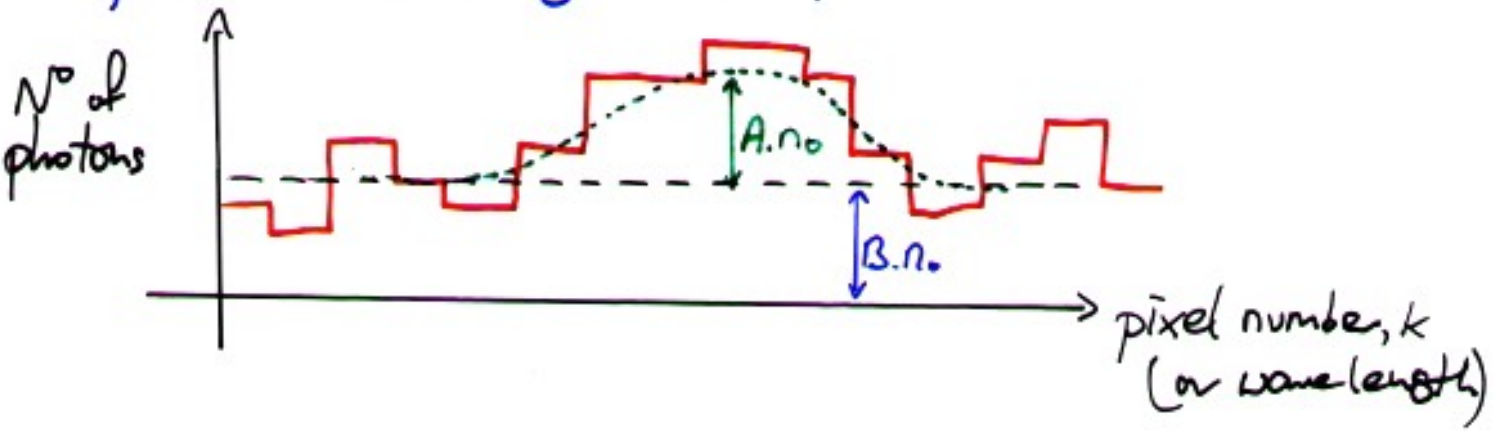
We have projected the 2-D posterior onto the A-axis, gathering together all the information favouring a particular A-value, independent of ø.

- Note that the value of A_0 derived from this marginalised pdf need not be exactly the same as the value derived from the full joint posterior, as they mean slightly different things.

Estimating 2 parameters simultaneously, example #2

[See Sivia p.37]

Imagine you have an observation of an optical spectral line, made with a good CCD.



- Within the data (red line) there is a spectral line (green line) the height of which you want to estimate, (A). n_0 is a constant that depends on the exposure time - the more time the more photons.

- There is also a background level (from atmospheric emission and the CCDs dark current), B.

Question:

Given the shape (but not height) of the spectral line, how do we optimally use the data to estimate A ?

Answer:

We do a fit, consistent with the noise, combining all the available information in all the data.

- We are told the line is thermally broadened, so has a Gaussian profile of width w

The k^{th} datum of the ideal (green) line is therefore

$$D_k = n_0 [A e^{-(x_k - x_0)^2/2w^2} + B]$$

where x_k is the position of the k^{th} datum, and x_0 the position of the centre of the line. These $\{D_k\}$ values comprise our model. We imagine we are told n_0, x_0 and w. Our job is to estimate A from the real data.

- We are counting photons, so if we are expecting D_k photons in the k^{th} pixel on average, the probability of getting N_k photons is

$$P(N_k | D_k) = \frac{D_k^{N_k} e^{-D_k}}{N!}$$

- the Poisson Distribution. [Note that N_k is an integer ≥ 0, whereas D_k is a real number ≥ 0]

- this is the probability of the data given our model
  - is the likelihood of the data.

Noting that D_k depends on our model parameters A & B, the joint likelihood of all the data, taken as independent, is

$$P(\{N_k\} | A, B, I) = \prod_{k=1}^{m} P(N_k | A, B, I)$$

(m ← m data points)

($\{N_k\}$: all the data; A, B: parameters)

- To get a posterior for A & B we need to set priors on A & B.

Lets choose $P(A, B | I) = \begin{cases} \text{constant for } A>0 \text{ and } B>0 \\ \text{zero otherwise} \end{cases}$

as A & B must be +ve.

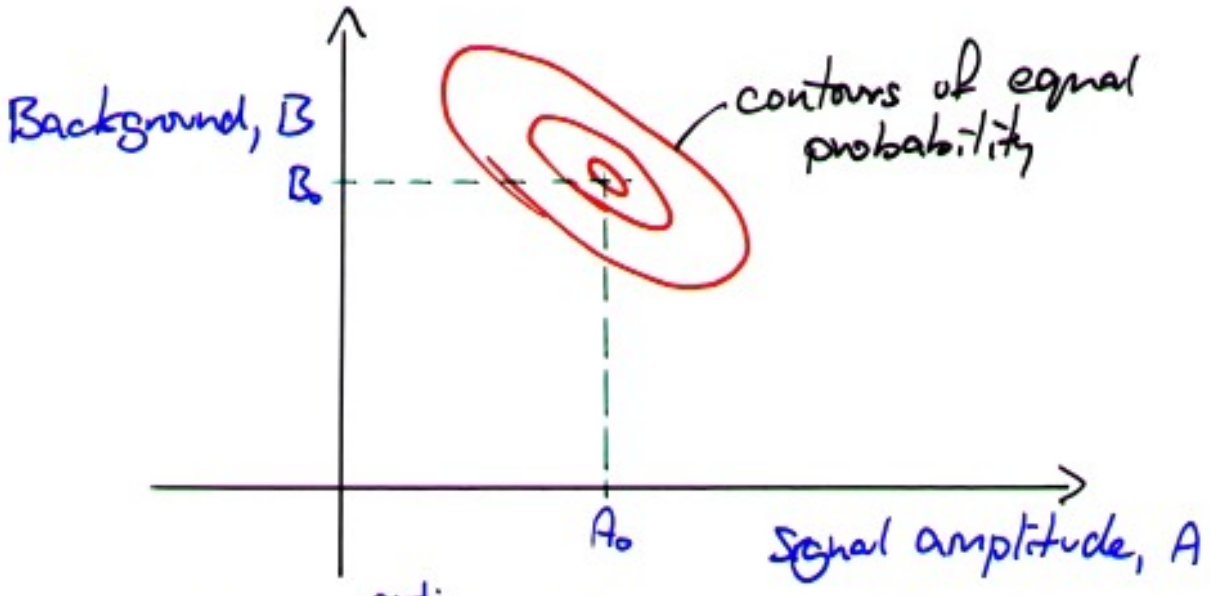
[It turns out that for multiplicative factors like A & B, a prior that most fully expresses our ignorance is of the form $P(A, B|I) \propto \frac{1}{AB}$, but we'll skip this at present]

- Our log-posterior for A & B is the log of prior × likelihood, ie

$$L = \ln[P(A, B|\{N_k\}, I)] = \text{const.} + \sum_{k=1}^{M} [N_k \ln D_k - D_k]$$

where D_k are functions of A & B, and $A \geq 0$, $B \geq 0$.

- We can now plot L as a function of A & B:



The parameters are anti-correlated - the data show that if A is small, B must be big, and visa versa.

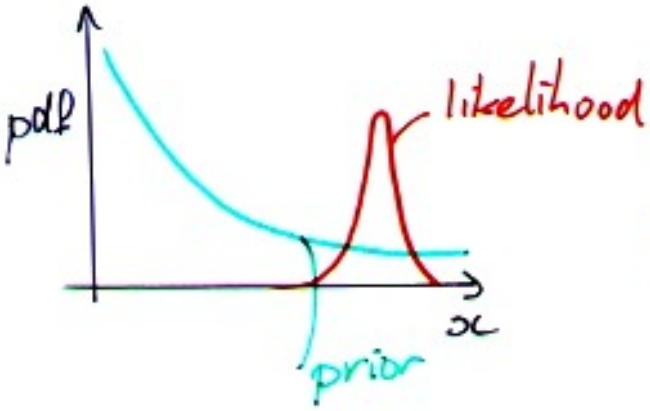
- see handout.

"Maximum Likelihood" – a shortcut

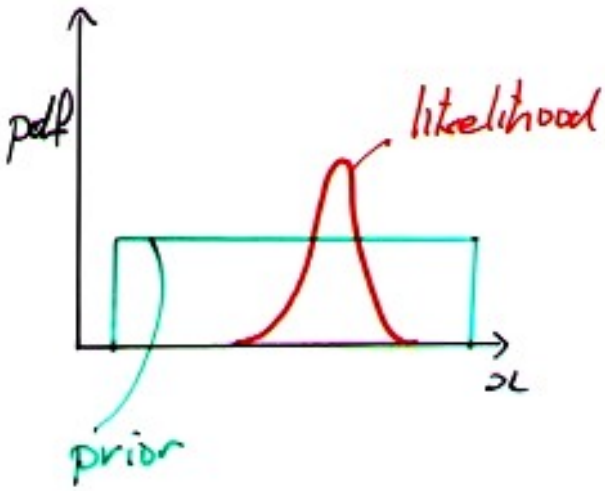
- The correct Bayesian approach is always to compute the posterior distribution of the parameter being estimated:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

- You will have noticed that often the prior is ~ constant over the region that contains most of the likelihood:



or even...



- when this is the case we can write:

$$\text{posterior} \propto \text{likelihood}$$

and the most probable x is the value that maximises the likelihood of the data.

Least squares

- Lets go further, and assume the data is affected by independent Gaussian errors, then (as before)

$$\text{likelihood} \propto \exp\left[-\sum_{k=1}^{N} \frac{(F_k - D_k)^2}{\sigma_k^2} \right]$$

where F_k is our model we are trying to fit

D_k is the data

σ_k^2 is the variance of the k^{th} datum

- The most probable value of our model parameter is therefore the one that minimises

$$\chi^2 = \sum_{k=1}^{N} \frac{(F_k - D_k)^2}{\sigma_k^2}$$

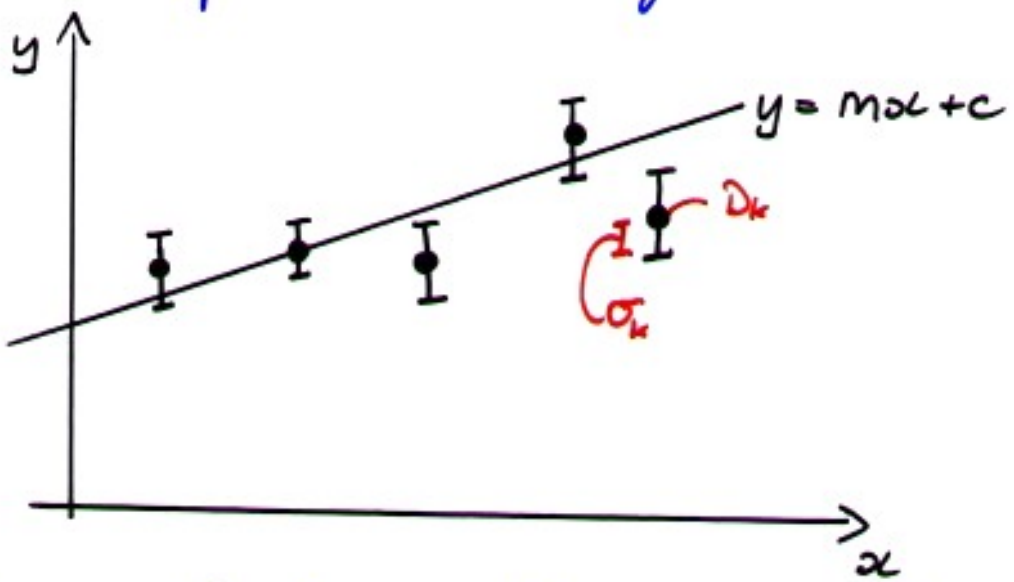
'chisquared' pronounced kiesquared

If the errors on all the measurements are the same, we need only minimise $\sum_{k=1}^{N} (F_k - D_k)^2$

i.e., find the least square difference between our model F and the data D, summed over all the data.

Example: straight-line fitting

- A classic example of least-squares: to fit m and c:



Errors in y only (σ_k is the error on the k^{th} datum)

The data are $\{D_k\}$ at x-positions $\{x_k\}$ so that

$$\chi^2 = \sum_{k=1}^{N} \frac{(mx_k + c - D_k)^2}{\sigma_k^2}$$

$$\text{Min } \chi^2 \Rightarrow \frac{\partial \chi^2}{\partial m} = 0 \quad ; \quad \frac{\partial \chi^2}{\partial c} = 0$$

ie $$m \sum \frac{x_k^2}{\sigma_k^2} + c \sum \frac{x_k}{\sigma_k^2} = \sum \frac{D_k x_k}{\sigma_k^2}$$

and $$m \sum \frac{x_k}{\sigma_k^2} + c \sum \frac{1}{\sigma_k^2} = \sum \frac{D_k}{\sigma_k^2}$$

$\Rightarrow$ solve for m and c ...

Evidence

We have been concentrating on parameter estimation (say, of a parameter "a"), based on data, d, using Bayes theorem:

$$p(a|d,I) = \frac{p(a,I) \cdot p(d|a,I)}{p(d|I)}$$

Because we have only been interested in "a" we have ignored the denominator on the right hand side: it does not depend on a, and ∴ only affects the normalisation of $p(a|d,I)$.

We can ensure normalisation by requiring that

$$\int_{-\infty}^{\infty} p(a|d,I)\, da = 1$$

- but what is $p(d|I)$?

- $p(d|I)$ is called the "evidence" (or "evidence for the model") or, more helpfully the "marginal likelihood":

$$p(d|I) = \int p(d,a|I)\, da$$

$$= \int \underbrace{p(a|I)\, p(d|a,I)}_{\text{note this is the numerator in Bayes Theorem.}}\, da$$

- It is the joint probability of the data and the parameter marginalised over the parameter — the probability of the data no matter what the parameter value is.

- It is therefore the likelihood of the model that parameterises the problem with the parameter a. (basically, I)

- Note that the probability of the model would be

$$p(I|d) = \frac{p(I)\overbrace{p(d|I)}^{\text{what we have been talking about}}}{p(d)}$$

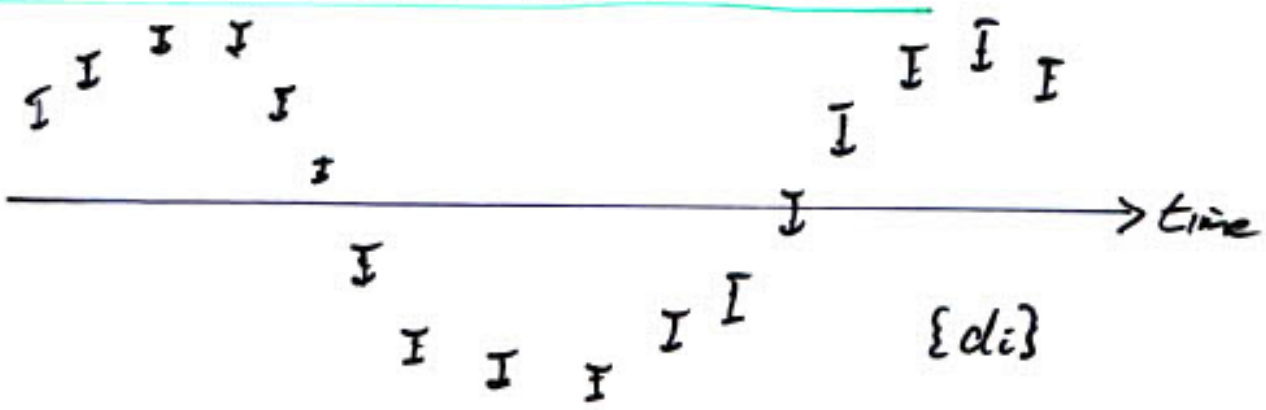
- We have now gone "one step back" and are considering the probabilities of models rather than model parameters. The basic idea is the same, but $p(d)$ is now rather vague – the probability of the data irrespective of an interpretive model.

  – see part #2 of Statistical Astronomy!

- However, we can say that a well-chosen model, capable of explaining the data should have a relatively high evidence value.

eg: Two different models fitted to data

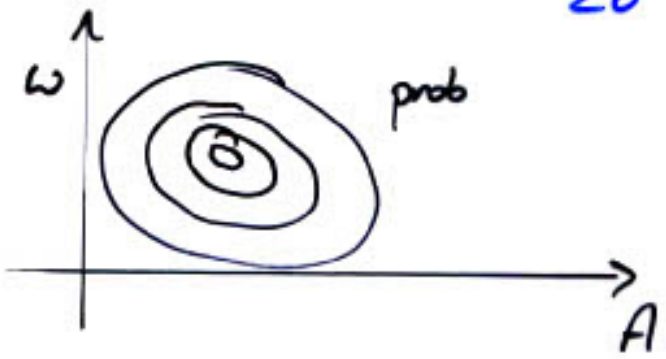
Data:



We could fit these to $A\sin\omega t_i$:

$$p(A,\omega|\{d_i\},I_1) \propto p(A,\omega)\prod_i \exp\left[-\frac{(d_i - A\sin\omega t_i)^2}{2\sigma^2}\right]$$

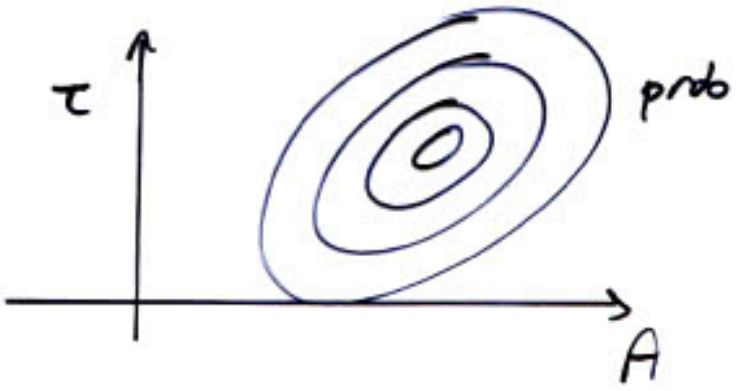
$\Rightarrow$



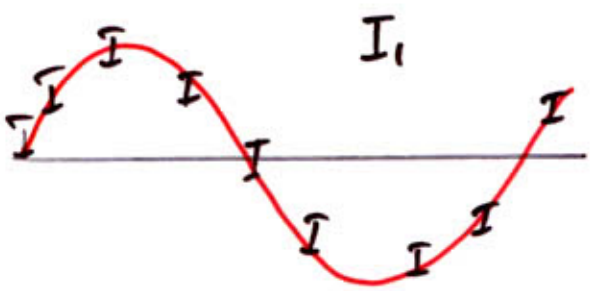
... but we could also fit them to $Ae^{-t/\tau}$!

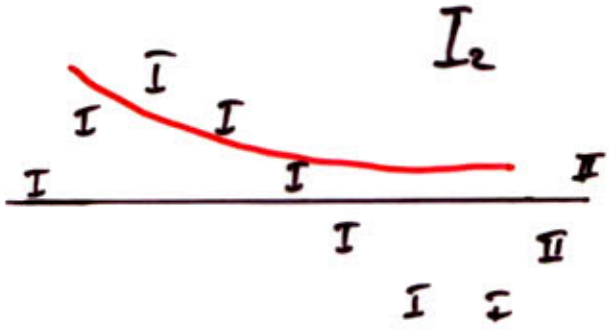
$$p(A,\tau|\{d_i\},I_2) \propto p(A,\tau)\prod_i \exp\left[-\frac{(d_i - Ae^{-t_i/\tau})^2}{2\sigma^2}\right]$$

$\Rightarrow$



- Both will produce a fit, like





but clearly I_1 (the sinusoid model) is better somehow than I_2 (the exponential model).

- This "goodness of fit" is reflected in the evidences:

$$p(\{d_i\}|I_1) \gg p(\{d_i\}|I_2)$$

so all other things being equal, I_1 should be preferred over I_2

- which brings us to the idea of hypothesis testing...

(see part #2!)