BERTINORO 2 (JVW)

Yet more probability
Bayes' Theorem*
¡Monte Carlo!

*The Reverend Thomas Bayes
1702-61

The Power-law (Scale-free) Distribution

$N(>L) = K L^{(\gamma+1)}$ (integral form)

$dN = (\gamma+1) K L^{\gamma} dL$ (differential form)

where N is the number of objects or events with a measured property (say luminosity) either >L (integral) or within the bin dL centered on L.

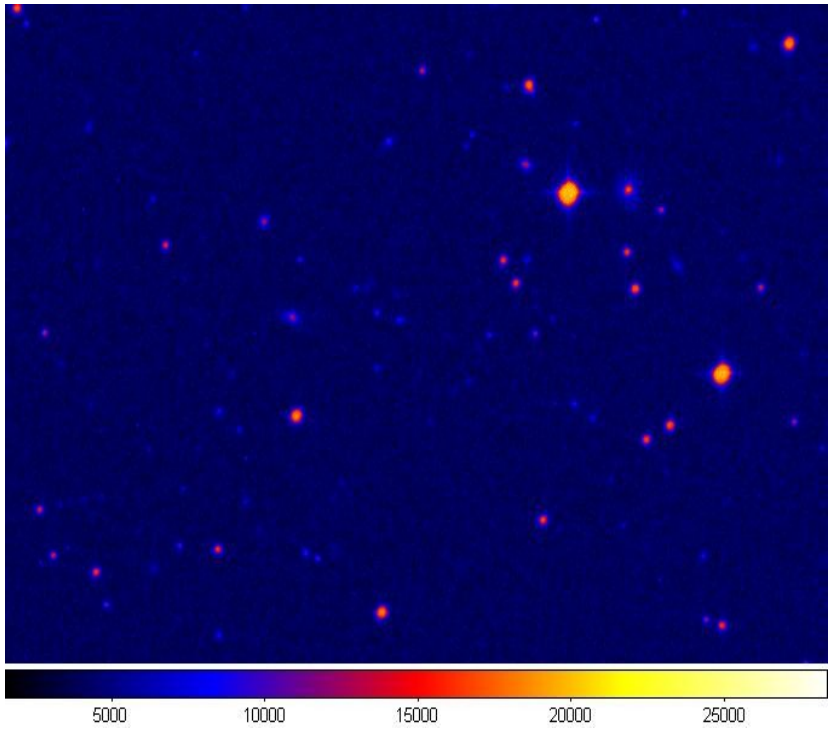
NB mean = variance = ∞, and γ is ~always negative (more little than big)

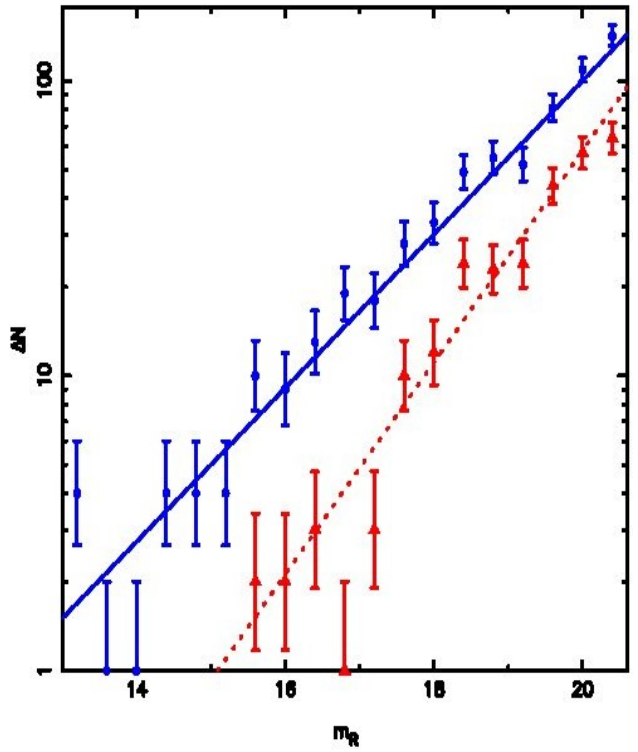
In real life: stock-market fluctuations, growth rates of companies, salaries, internet connectivity, critical exponents for fluids, earthquakes, avalanches, forest fires, sandpile-slips ....

In astronomy: initial (Salpeter) mass function, initial fluctuation spectrum, source counts, number-magnitude relations, luminosity functions ....

Criticality is the key

Power law example





A 15-arcmin-square of sky from the R-band UKSTU sky survey at RA 22^h, Dec -18°. The scanning process recognizes about 750 images in the area. Right: the number - magnitude count (a 'source count' at other wavelengths) for all objects (dots) and for objects classified as galaxies (triangles). Note that magnitude is an inverse logarithmic scale: $m_1 - m_2 = -2.5 \log(L_1/L_2)$, where L is luminosity.

Power-Law Pitfalls

- index of the exponent

- integral vs differential

- ΔL vs $\Delta \log L$

- choice of interval; where to plot values of L

- choice of range; finite bounds?

- sampling errors, rms, mean, median..

This is the distribution from the devil - we must learn to live with it.

Probability: summary + new things

Laplace: Principle of Indifference: assign equal probabilities unless we know better.

If can identify equal cases, calculating prob ≡ counting.

Can estimate from data – `frequentist'; fraught with risk.

Probability is a **numerical formalization of our degree or intensity of belief**.

The common language from formalizing this: Cox's rule, Kolmogorov axioms, (1) random event: 0<prob (A)<1, (2) sure event prob(A)=1, (3) exclusive events prob(A or B) = prob(A) + prob(B).

From these (a) independence: prob(A and B) = prob(A) . prob(B)
(b) conditional probability: prob(A|B) = prob(A and B)/prob(B)

(nb suppose A and B independent: prob(A|B) = prob(A), and thus prob(A and B) = prob(A).prob(B) again)

Bayes' Theorem

$$prob(B|A) = prob(A|B) \cdot prob(B) / prob(A)$$

..is really an identity, from prob(*A* and B) = prob(B and *A*)

The event **A** (the data), follow **B**
Prob(**A**) is the **normalizing** factor
Prob(**B**) is the **prior probability**, to be modified by experience (namely the data **A**)
Prob(**A**|**B**) is the **likelihood**
Prob(**B**|**A**) is the **posterior probability**, the answer, the subsequent state of belief

An innocent mathematical identity – but ***its interpretation or application has momentous consequences*** for analysis of data, experimentation.

Notice also the affinity with **maximum likelihood** analysis.

What does this mean?

Example: the marbles-in-the-pot calculation, **M** white marbles, **N** red marbles. What's the probability of drawing 3 red and 2 white out of the **M**+**N** in 5 tries? This is a counting problem, **PoI**, etc, and we can count.

But this is not what we want to know! This is the wrong sum!

We do *not* want to know the prob of drawing a certain number of each colour. What we want is the *inverse probability* calculation: we have data, *ie* we have a certain number drawn, say 2 white, 3 red – and we want to infer the population properties of what's in the pot.

Example: N red, M white in a pot, total N+M=10. If I make 5 draws (T=5) and get 3 red and 2 white, how many reds are in the pot?

So: from Bayes

$$\text{prob(contents of pot | data)} \;\alpha\; \text{prob(data | contents of pot)}$$

We can deal with the term on the RHS. We take as a model for the probability of red, pr = N/(N+M) , assuming no funny business

What does this mean? 2

The likelihood term = $\binom{T}{R} \cdot pr^{R} \cdot pr^{(T-R)}$

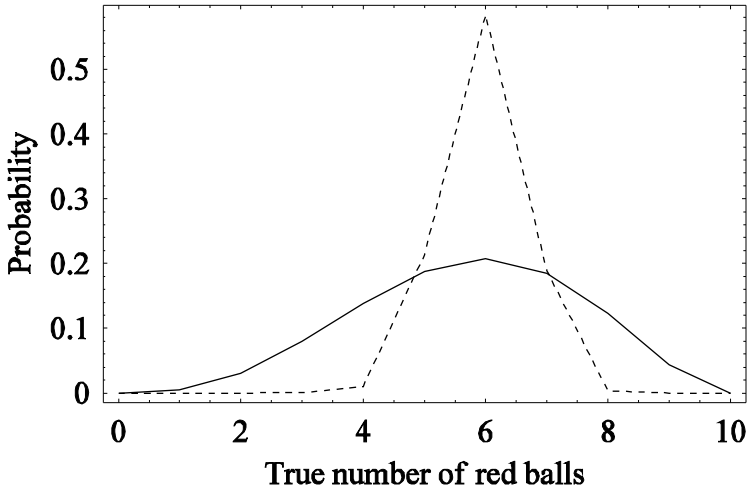
Binomial coefficient

$$\binom{x}{y} \equiv \frac{x!}{y!\,(x-y)!}$$

This is just the binomial distribution, which we met already – the old question of **n** successes out of **N** trials.

What about the prior? Let's take it as uniformly likely between 0 and N+M.

So we can calculate the (un-normalized) posterior probability:



Here's the results for our original 5 tries (3 reds) and 50 tries (30 reds).

What does this mean? 3

Unsurprising? Common sense? Maybe….but consider….

1. We now can describe our state of belief about the contents of the pot in physical or mathematical terms. We believe on the basis of data that there are 6 reds, but – in the case of the 5 tries, there could be as few as 2 and as many as 10. The probability of the pot containing 3 reds or less is 11 per cent, etc.

2. We have answered our scientific question: we have made an inference about the contents on the basis of data.

Bayes' theorem allows us to make inferences from the data, rather than compute the data we would get if we happened to know all relevant information.

Example: data from 2 populations – different means? Most books: here's the data you get if you have populations with different means. That's **not** what we asked! We want to know, given the data, what is the probability/belief state of our model.

3. Note use of prior information – we assigned probabilities to **N** to reflect what we know. `Prior' suggests `before' but means **`what we know apart from the data'.** In the current example we used a uniform prior – and got a not unexpected result. Priors can **change anticipated results** in violent and dramatic ways.

How to describe the posterior distribution?

The **peak* of the posterior probability distribution** is one way amongst many of characterizing the distribution by a single number.

The **posterior mean** is another choice, defined by

$$< \rho > = \int \rho \, \text{prob}(\rho \mid \text{data}) \, d\rho$$

If we have had **N** successes and **M** failures, the posterior mean is given by a famous result called **Laplace's Rule of Succession**:

$$< \rho > = (N+1) / (N+M+2)$$

Unless posterior distributions are very narrow, attempting to characterize them by a single number is misleading.

Depends on what is to be done with the answer, which in turn depends on having a carefully-posed question in the first place.

* = Maximum Likelihood if prior = 1 (`flat prior')

Inferences with Probability - Methodology

1. Parameter estimation. This is closely related to the field of data-modelling. We have a probability distribution **f (data | ά)** and we wish to know the parameter vector **ά**.

The Bayesian route: **compute the posterior distribution of ά.**

2. This method is nearly the classical technique of **maximum likelihood (ML)**. If the prior is ``diffuse'' then posterior probability is proportional to likelihood term **f (data | ά)** . **ML** picks out the **mode** of the posterior, the value of **ά** which maximizes the likelihood. Characterizing the posterior by one number is useful because of powerful theorems on **ML.**

3. Often knowing the **posterior distribution** of the parameter of interest is enough for e.g. a comparison with an exactly-known quantity, perhaps derived from some theory.

4. We may wish to compare with an experimental determination of some other parameter vector **β**, e.g. for scalar parameters **α** and **β**, we might ask for the probability that **α > β**……

Monte Carlo: why generate random numbers?

On many occasions in hypothesis-testing and model-fitting we must have a set of numbers distributed how we might guess the data to be.

We may wish

☺ to check error propagation

☺ to test a test to see if it works as advertised;

☺ to test efficiency of tests;

☺ to find how many iterations we require to reach a given level of significance;

☺ to test our code.

We gotta have these random numbers, of two kinds:
- **uniformly** distributed,
- drawn randomly from a parent population of **known** frequency distribution.

The pitfalls of random-number generators

Usual form: x=ran1(idum)

No excuse for using bad random data.
EXAMPLE: RANDU, the infamous IBM random-number generator.

Cycle length - how long is it before the pseudo-random cycle is repeated?

Important to understand the characteristics of the generator.

Essential to follow the prescribed procedure.

Never forget that the routines generate pseudo-random numbers.

Numerical Recipes presents a number of methods, from single expressions to powerful routines.

Random numbers for a frequency distribution

How do we draw a set of random numbers following a **given** frequency distribution?

Suppose we have a way of producing random numbers that are uniformly distributed, in say the variable α; and we have a functional form for our frequency distribution $dn/dx = f(x)$. We need a transformation $x = x(\alpha)$ to distort the uniformity of α to follow $f(x)$. But we know that

$$\frac{dn}{dx} = \frac{dn}{d\alpha}\frac{d\alpha}{dx}$$

and as $dn/d\alpha$ is uniform, thus

$$\frac{dn}{dx} = \frac{d\alpha}{dx},$$

and

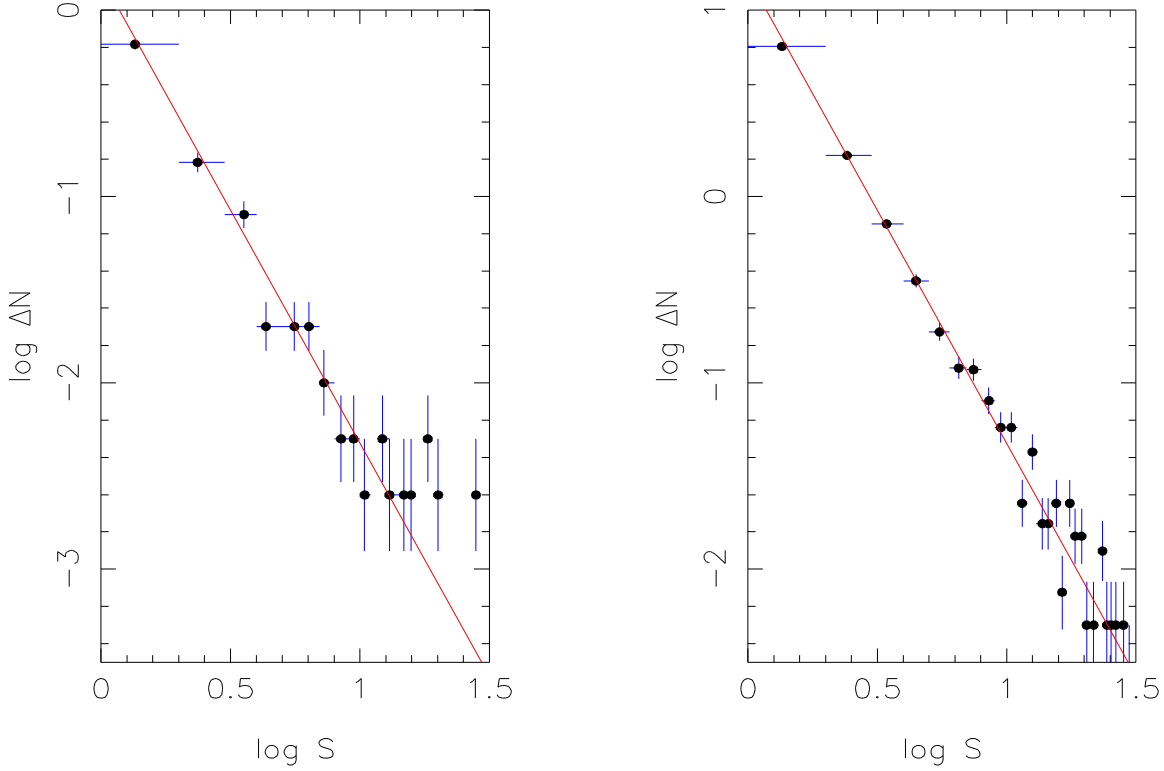
$$\alpha(x) = \int^x f(x)dx,$$

so that the required transformation is $x = x(\alpha)$.

We therefore need $x = f^{-1}(a)$, the inverse function of the integral of $f(x)$

Randoms from frequency distribution – 2

EXAMPLE: A source-count distribution is given by $\mathbf{f(x)dx = -1.5x^{-2.5}dx}$, a `Euclidean' differential source count. Here $\mathbf{d\alpha = -1.5x^{-2.5}dx}$, $\mathbf{\alpha = x^{-1.5}}$, and the transformation is $\mathbf{x = f^{-1}(\alpha) = \alpha^{-1/1.5}}$.

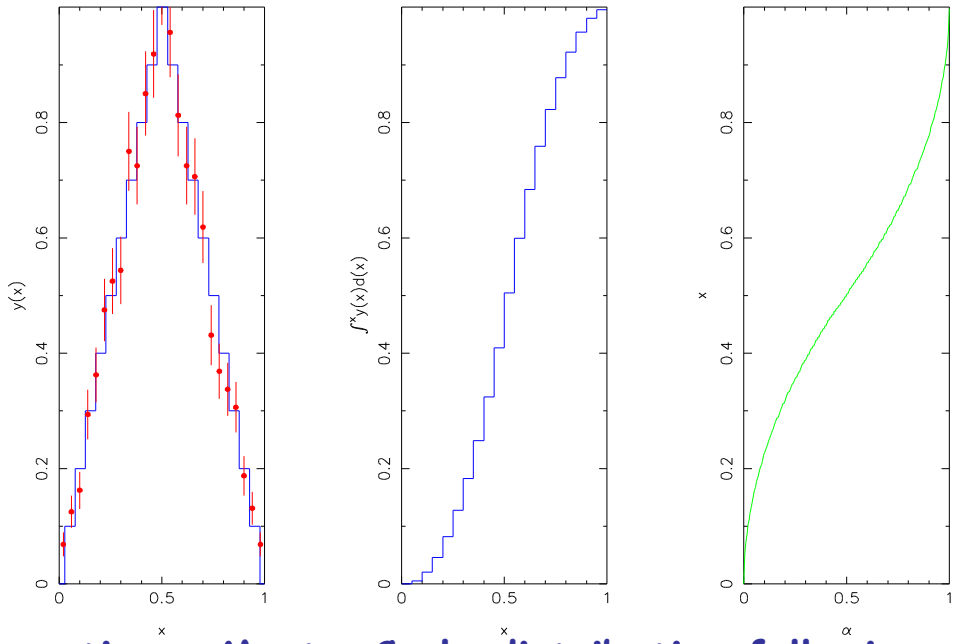


Differential source counts generated via Monte Carlo sampling with an initial uniform deviate, obeying the source-count law $N(>S) = kS^{-1.5}$. The straight line in each shows the anticipated count with slope -2.5. left - k = 1.0, 400 trials, right - k = 100, 4000 trials.

Randoms from frequency distribution – 3

The very same procedure works if we don't have a functional form for **f(x)dx**.
If this is a histogram, we need simply to calculate the integral version, and perform the reverse function operation as before.

EXAMPLE:



An example of generating a Monte-Carlo distribution following a known histogram. Left: the step-ladder histogram, with points from 2000 trials, produced by a) integrating the function (middle) and b) transforming the axes to produce f^{-1} of the integrated distribution (right). The points with √N error bars in the left diagram are from drawing 2000 uniformly-distributed random numbers and transforming them according to the right diagram.

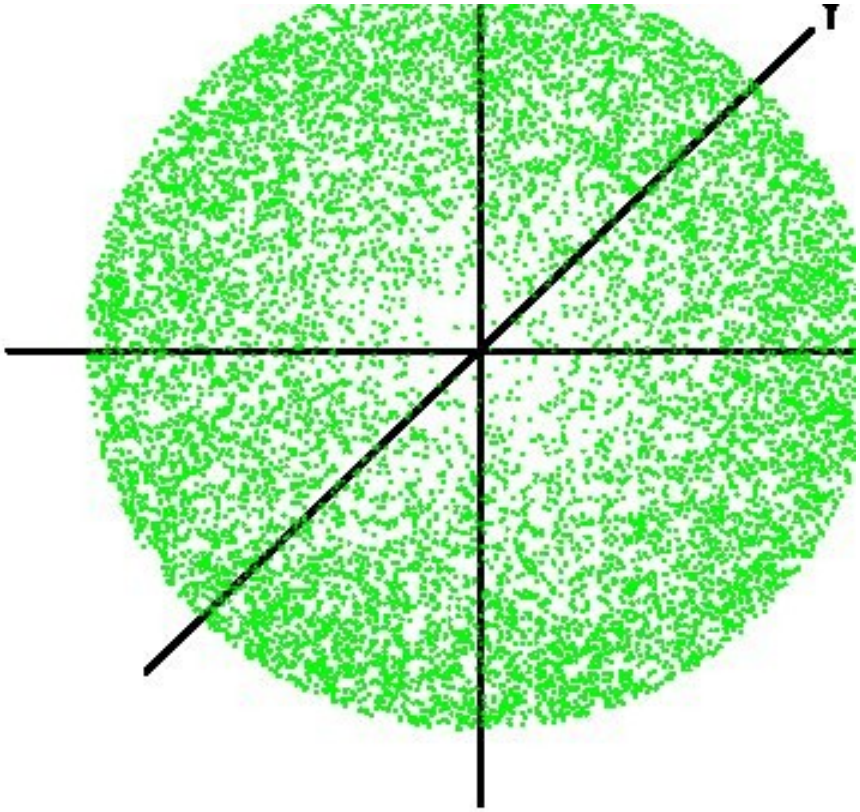
CHALLENGE I: Random number generation

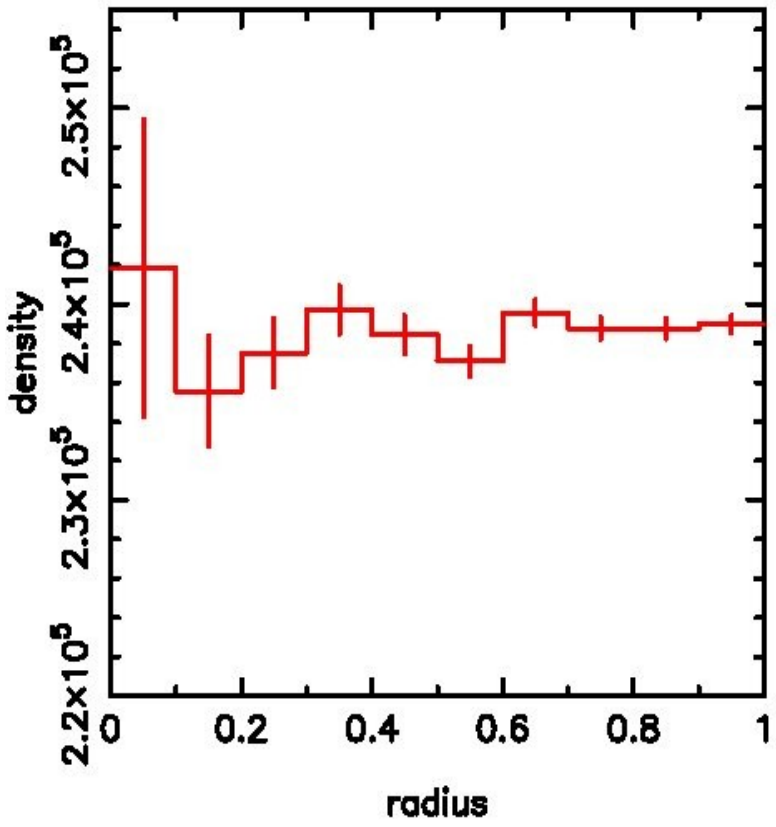
1. Get a random number generator working. Test that it works by checking its uniformity over the generator range (usually 0 to 1). Make sure you can scale this range, e.g. over -10 to +10.

2. Using your random-number subroutine, make a Poisson generator, by having a small number of photons per time bin, starting with, say $\mu=3$ photons on average. Compile the distribution in the time bins over a large number of bins, say 100. Now make the 'integration time' longer so that the numbers rise from an average of 3 to 10 to 50. At what point does your distribution become indistinguishable from a Gaussian with mean = μ and $\sigma = \sqrt{\mu}$?

3. Make your own universe....

3. Enter your very own Monte Carlo Universe





Make a universe: uniformly fill it with 1000000 objects of random (x,y,z) at radii from 0 to 1.0. Make sure that you have managed to make it uniform by checking the densities of objects in radius shells

End Bertinoro 2