ASTRO509-3

Probability Distributions

"I have had my results for a long time: but I do not yet know how I am to arrive at them."

Johann Carl Friedrich Gauss 1777 -1855

ASTR509 © Jasper Wall Fall term 2013

In the previous episode......

We considered the (shaky) basis for defining probabilities (= `belief'); at the rules which define probability theory, and at conditional probability.

This led us to Bayes' theorem:

$$\mathrm{prob(B \mid A)} = \frac{\mathrm{prob(A \mid B) prob(B)}}{\mathrm{prob(A)}}$$

We considered how the terms broke down into 'posterior distribution' (what we want to know), the 'likelihood' (relative probability of the data given the model), the 'prior distribution', and the 'normalization' term.

We demonstrated Bayes in action with simple examples, urns and supernovae.

We discussed how it is that 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. i.e. we showed that 'inverse' problems are the real problems.

We started to work on priors, used them in simple examples, noting that the 'prior' really means 'what we know apart from the data'.

We discussed characterization of the posterior distribution.

Probability distributions

Example: toss four `fair' coins. The probabilities:

- no heads = $(1/2)^4$;
- one head = 4 x $(1/2)^4$;
- two heads = 6 x $(1/2)^4$, etc.
- sum of the possibilities for getting 0 heads to 4 heads = 1.0.

If x is the number of heads (0,1,2,3,4), we have a set of probabilities

prob(x) = (1/16,1/4,3/8,1/4,1/16);

we have a probability distribution describing expectation of occurrence of event **x**.

This probability distribution is discrete; there is a discrete set of outcomes and so a discrete set of probabilities for those outcomes.

Probability distributions 2

In other words we have a mapping between the outcomes of the experiment and a set of integers.

- sometimes the set of outcomes maps onto real numbers instead; here we discretize the range of real numbers into little ranges within which we assume the probability does not change.

- If **x** is the real number that indexes outcomes, we associate with it a probability density **f(x);** the probability that we will get a number `near’ **x**, say within a tiny range **δx,** is **prob(x) δx**.

- we loosely refer to ‘probability distributions' with discrete outcomes or not.

Probability distributions 3

Formally: if ***x*** is a continuous random variable, then ***f(x)*** is its **probability density function**, commonly termed **probability distribution**, when

1. Probability $[a < x < b] = \int_a^b f(x)dx$

2. $\int_{-\infty}^{\infty} f(x)dx = 1$, and

3. ***f(x)*** is a single-valued non-negative number for all real ***x***.

The corresponding **cumulative distribution function** is $F(x) = \int_{-\infty}^{x} f(y)dy$

Probability distributions and distribution functions may be similarly defined for sets of discrete values of ***x.***

Distributions may be **multivariate**, functions of more than one variable.

Probability distributions 4

Quantifiers - **location** (where is the `centre'?)
- **dispersion** (what is the `spread'?)

These quantifiers can be given by the first two **moments of the distributions:**

$$\mu_1(\text{mean}) = \mu = \int_{-\infty}^{\infty} x f(x)\, dx \tag{1}$$

$$\mu_2(\text{variance}) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu_1)^2 f(x)\, dx \tag{2}$$

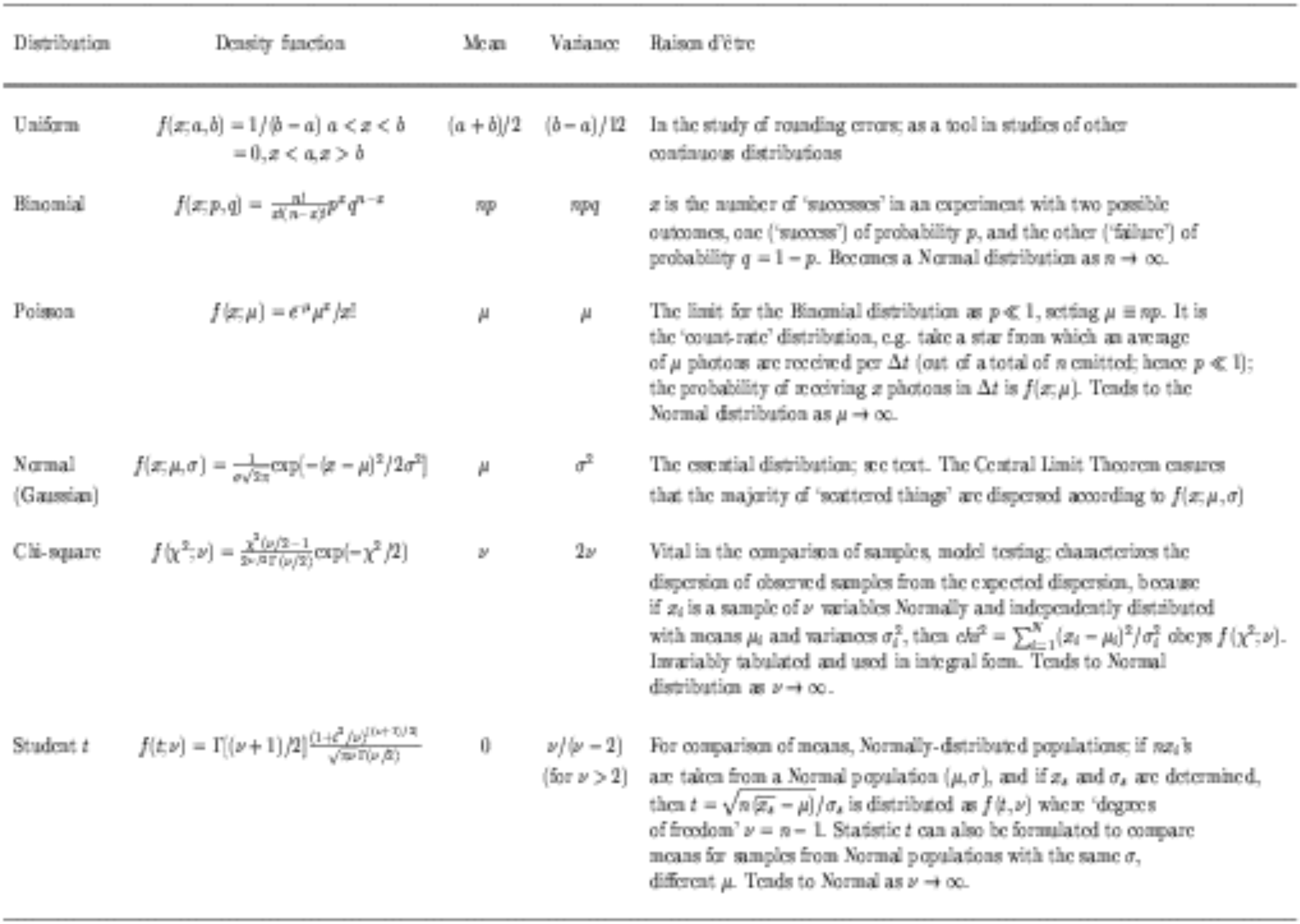
Other moments, particularly the third moment (‘skewness’) can play a prominent role; but these two are far the most important.

Probability distributions 5

There are probability distributions we can calculate resulting from ideal experiments, outcomes or combinations of these.

The best-known are the UNIFORM, BINOMIAL, POISSON and GAUSSIAN (or NORMAL) distributions, and these have a bunch of hangers-on...........

The common probability density functions



Distribution	Density function	Mean	Variance	Raison d'être
Uniform	$f(x;a,b) = 1/(b-a)\ a < x < b$ $= 0, x < a, x > b$	$(a+b)/2$	$(b-a)/12$	In the study of rounding errors; as a tool in studies of other continuous distributions
Binomial	$f(x;p,q) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$	np	npq	x is the number of 'successes' in an experiment with two possible outcomes, one ('success') of probability p, and the other ('failure') of probability $q = 1-p$. Becomes a Normal distribution as $n \to \infty$.
Poisson	$f(x;\mu) = e^{-\mu}\mu^x/x!$	μ	μ	The limit for the Binomial distribution as $p \ll 1$, setting $\mu \equiv np$. It is the 'count-rate' distribution, e.g. take a star from which an average of μ photons are received per Δt (out of a total of n emitted; hence $p \ll 1$); the probability of receiving x photons in Δt is $f(x;\mu)$. Tends to the Normal distribution as $\mu \to \infty$.
Normal (Gaussian)	$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-(x-\mu)^2/2\sigma^2]$	μ	σ^2	The essential distribution; see text. The Central Limit Theorem ensures that the majority of 'scattered things' are dispersed according to $f(x;\mu,\sigma)$
Chi-square	$f(\chi^2;\nu) = \frac{\chi^{2(\nu/2-1)}}{2^{\nu/2}\Gamma(\nu/2)}\exp(-\chi^2/2)$	ν	2ν	Vital in the comparison of samples, model testing; characterizes the dispersion of observed samples from the expected dispersion, because if x_i is a sample of ν variables Normally and independently distributed with means μ_i and variances σ_i^2, then $chi^2 = \sum_{i=1}^{N}(x_i-\mu_i)^2/\sigma_i^2$ obeys $f(\chi^2;\nu)$. Invariably tabulated and used in integral form. Tends to Normal distribution as $\nu \to \infty$.
Student t	$f(t;\nu) = \Gamma[(\nu+1)/2]\frac{(1+t^2/\nu)^{-(\nu+1)/2}}{\sqrt{\pi\nu}\Gamma(\nu/2)}$	0	$\nu/(\nu-2)$ (for $\nu > 2$)	For comparison of means, Normally-distributed populations; if nx_i's are taken from a Normal population (μ,σ), and if $\bar{x}_s$ and σ_s are determined, then $t = \sqrt{n(\bar{x}_s-\mu)}/\sigma_s$ is distributed as $f(t,\nu)$ where 'degrees of freedom' $\nu = n-1$. Statistic t can also be formulated to compare means for samples from Normal populations with the same σ, different μ. Tends to Normal as $\nu \to \infty$.

The Binomial Distribution

There are two outcomes - `success' or `failure'. This common distribution gives the chance of **n** successes in **N** trials, with the probability of a success at each trial **ρ**, and successive trials are independent. This probability is

$$\mathrm{prob}(n) = \left(\begin{array}{c} N \\ n \end{array} \right) \rho^n (1-\rho)^{N-n}.$$

Bernoulli, Johann, 1667-1748

The leading term, the combinatorial coefficient, gives the number of distinct ways of choosing **n** items out of **N**:

$$\left(\begin{array}{c} N \\ n \end{array} \right) = \frac{N!}{n!(N-n)!}.$$

The Binomial Distribution 2

This coefficient can be derived as follows.

There are **N!** equivalent ways of arranging the **N** trials. However there are **n!** permutations of the successes, and **(N-n)!** permutations of the failures, which correspond to the same result –

namely, exactly **n** successes, arrangement unspecified. Since we require not just **n** successes (probability **ρ^{n}**) but exactly **n** successes, we need exactly **N-n** failures, probability **(1- ρ)$^{(N-n)}$** as well. The binomial distribution follows from this argument.

The binomial distribution has a mean value given by

$$\sum_{n=0}^{N} n\,\mathrm{prob}(n) = Np$$

and a variance or mean square value of

$$\sum_{n=0}^{N} (n - Np)^2 \mathrm{prob}(n) = Np(1-p).$$

The Binomial Distribution - Example

In a sample of 100 galaxy clusters selected by automatic techniques,10 contain a dominant central galaxy. We plan to check a different sample of 30 clusters, now selected by X-ray emission. **How many of these clusters do we expect to have a dominant central galaxy?**

If we assume that the 10 per cent probability holds for the X-ray sample, then the chance of getting **n** dominant central galaxies is

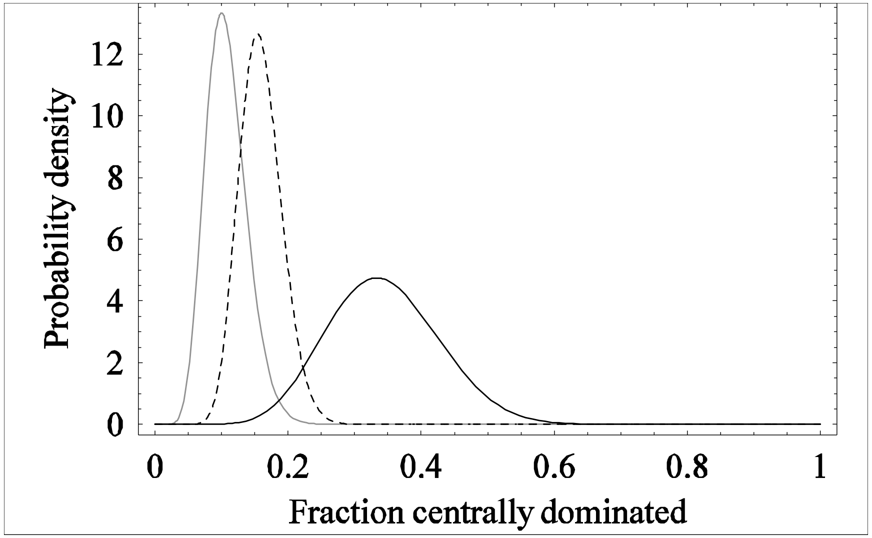
$$\mathrm{prob}(n) = \left(\begin{array}{c} 30 \\ n \end{array} \right) 0.1^n 0.9^{30-n}.$$

E.g.the chance of getting 10 is about 1%; if we found this many we would be suspicious that the X-ray cluster population differed from the general population.

Suppose we made these observations and did find 10 centrally-dominated clusters. **What can we do with this information?**

A Bayesian calculation that parallels the supernova example! Assuming the X-ray galaxies are a homogeneous set, we can deduce the probability distribution for the fraction of these galaxies that have a dominant central galaxy. A relevant prior would be the results for the original larger survey……….

Binomial example, continued ....



The posterior probability distribution for the observation that 10/30 X-ray-selected clusters are centrally-dominated. The dark black line uses a uniform prior distribution for this fraction; the dashed line uses the prior derived from an assumed previous sample in which 10 out of 100 clusters had dominant central members. The light curve shows the distribution for this earlier sample.

The figure makes clear that the data are not really sufficient to alter our prior very much. For example, there is only a 10 per cent chance that the centrally-dominant fraction exceeds even 0.2; the possibility of it being as high as 33% is completely negligible. **Our X-ray clusters differ markedly from the general population.**

The Poisson Distribution

The Poisson distribution derives from the binomial in the limiting case of very rare events and a large number of trials, so that although **ρ** → 0, **N ρ** → (a finite value). Calling this finite mean value **μ**, the Poisson distribution is

$$\mathrm{prob}(n) = \frac{\mu^n}{n!} e^{-\mu}.$$

The variance of the Poisson distribution is also **μ**.

Example : Village blacksmiths are/were occasionally kicked by the horse they were shoeing, say on average, 3 times per year. How often would they have good years with no kicks? How often would they have bad years, say 10 kicks?

Poisson, Siméon-Denis, 1781-1840

The Poisson Distribution - Example 2

A familiar example of a process obeying Poisson statistics is the number of photons arriving during an integration. The probability of a photon arriving in a fixed interval of time is (often) small. The arrivals of successive photons are independent. Thus the conditions necessary for the Poisson distribution are met.

Hence, if the integration over time **t** of photons arriving at a rate **λ** has a mean of **μ = λt** photons, then the fluctuation on this number will be **σ = √ μ**, because we know that the variance is **μ.**

(In practice we usually only know the number of photons in a single exposure, rather than the mean number; obviously we can then only estimate the **μ**.)

For **photon-limited** observations, such as CCD images or spectra,

$$\mu = \lambda t \text{ while } \sigma = \sqrt{\lambda t}.$$

If we ``integrate" more,

$$\sigma \propto \sqrt{t}, \text{ while signal} \propto t.$$

Thus **Signal/Noise** ∝ √**t**, the **sky-limited** case.

Poisson example 2, continued

There are the following further cases:

1. *Photon-limited*, *e.g.* CCD observations of faint objects:

$$S/N \propto \frac{\mu}{\sqrt{\mu}}, \ \textit{or} \ \propto \sqrt{t}$$

2. *Readout-limited*, *e.g.* CCD observations of bright objects:

$$S/N \propto \frac{\mu}{\sigma_{ccd}}, \ \textit{or} \ \propto t$$

   for CCD of readout noise σ_{ccd}.

3. *Receiver-limited*, *e.g.* radio astronomy:

$$S/N \propto \frac{S}{\sigma_{rec}/\sqrt{t}}, \ \textit{or} \ \propto \sqrt{t}$$

   for receiver of thermal noise σ_{rec}.

The Gaussian (Normal) Distribution

Both the Binomial and the Poisson distributions tend to the Gaussian distribution, large **N** in the case of the Binomial, large **μ** in the case of the Poisson.
The (univariate) Gaussian (Normal) distribution is

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

from which it is easy to show that the mean is **μ** and the variance is $\boldsymbol{\sigma^2}$.

The Gaussian (Normal) Distribution 2

For the binomial when the sample size is very large, the discrete distribution tends to a continuous probability density

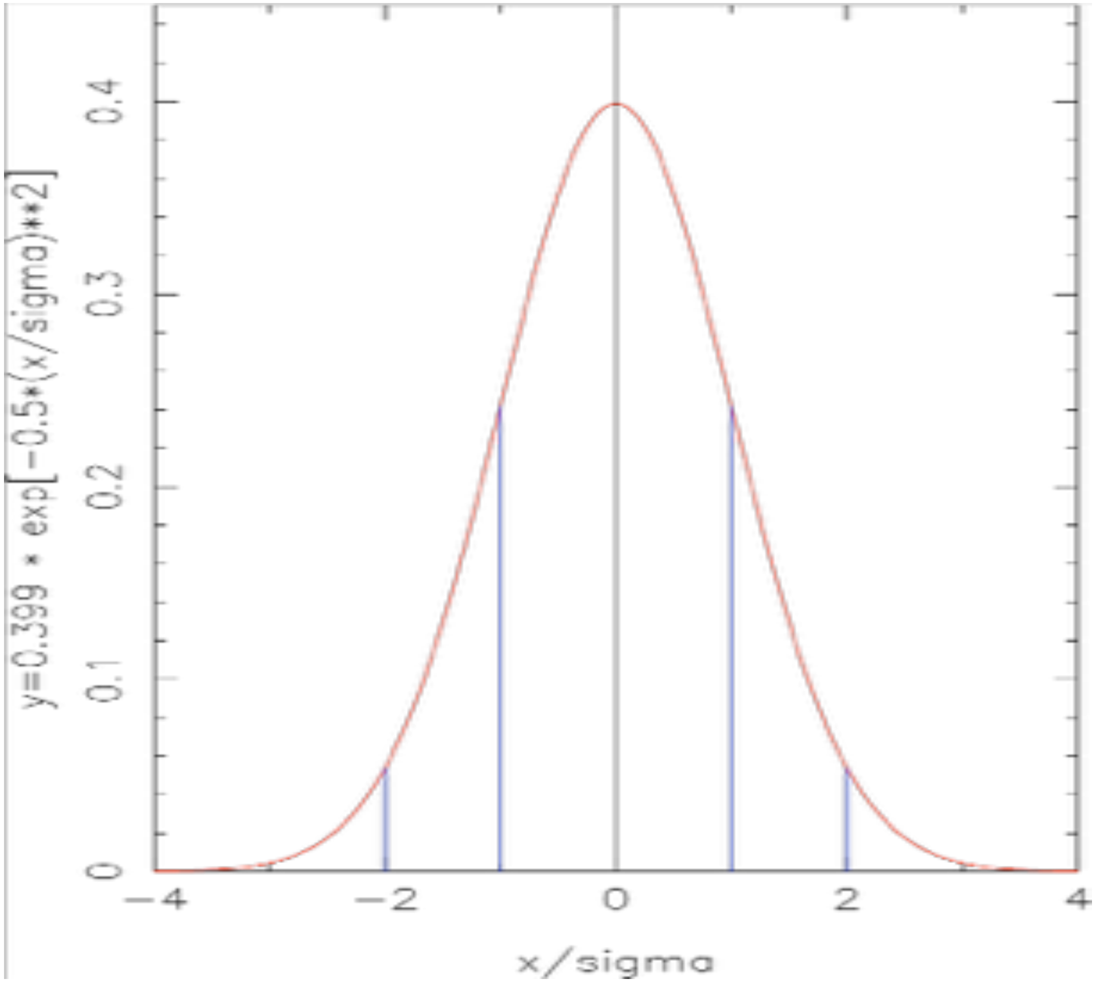
$$\mathrm{prob}(n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(n-\mu)^2}{2\sigma^2}\right]$$

in which the mean **μ** = **N p** and variance **σ²** = **N p (1-p)** are still given by the parent formulae for the binomial distribution.

Here is an instance of the discrete changing to the continuous distribution:

in this approximation we can treat **n** as a continuous variable (because **n** changes by one unit at a time, being an integer **=>** the fractional change **1/n** is small).

The Gaussian Distribution 3



The Central Limit Theorem

The true importance of the Gaussian distribution and its dominant position in experimental science, stems from the **Central Limit Theorem**. A non-rigorous statement of this is as follows.

Form averages M_n from repeatedly drawing n samples from a population x_i with finite mean μ, variance σ^2. Then the distribution of

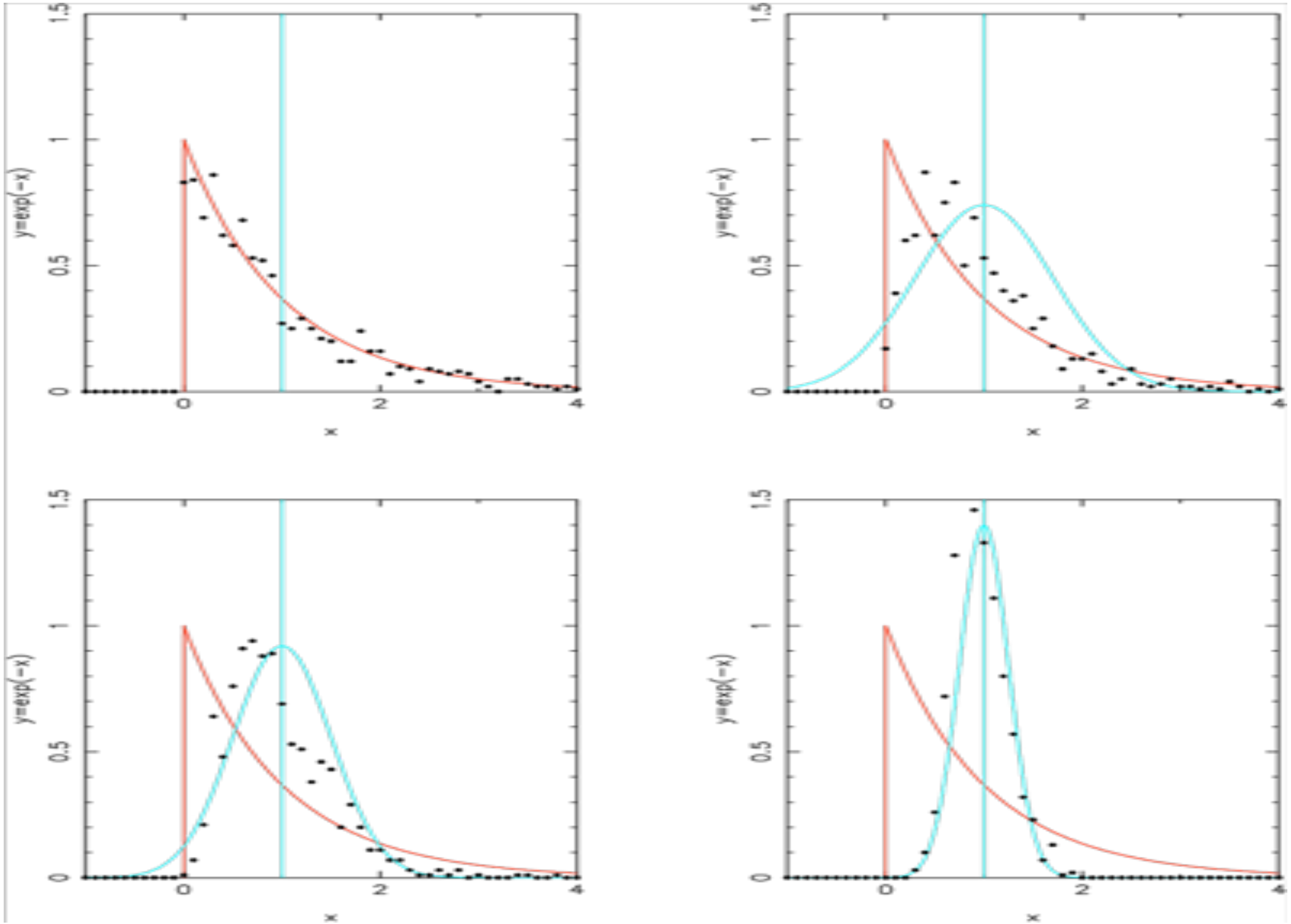
$$[\frac{(M_n - \mu)}{\sigma/\sqrt{n}}] \rightarrow \text{Gaussian distribution}$$

with mean 0, variance 1, as n $\rightarrow \infty$.

THIS MAY BE THE MOST REMARKABLE THEOREM EVER

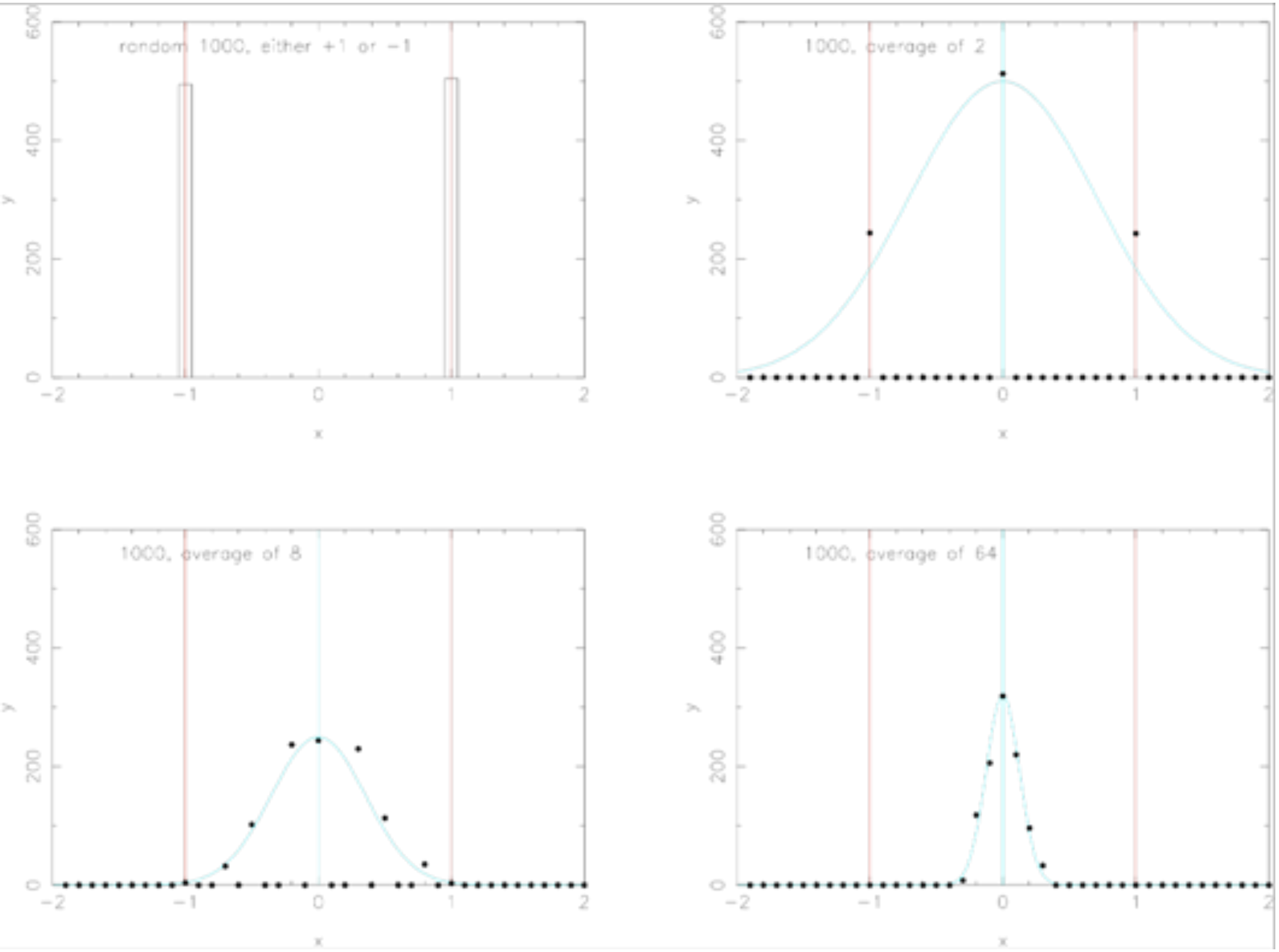
- It says that averaging will produce a Gaussian distribution of results - **no matter the shape of distribution from which the sample is drawn**.
- Eyeball integration counts!
- Errors on averaged samples will always look `Gaussian'.
- **The Central Limit Theorem shapes our entire view of experimentation. => error language of sigmas, describing tails of Gaussian distributions.**

The Central Limit Theorem - Example 1



200 values drawn from exponential distribution with cutoff; averages of 1, 2, 4, 16

The Central Limit Theorem - Example 2



Gaussian tails

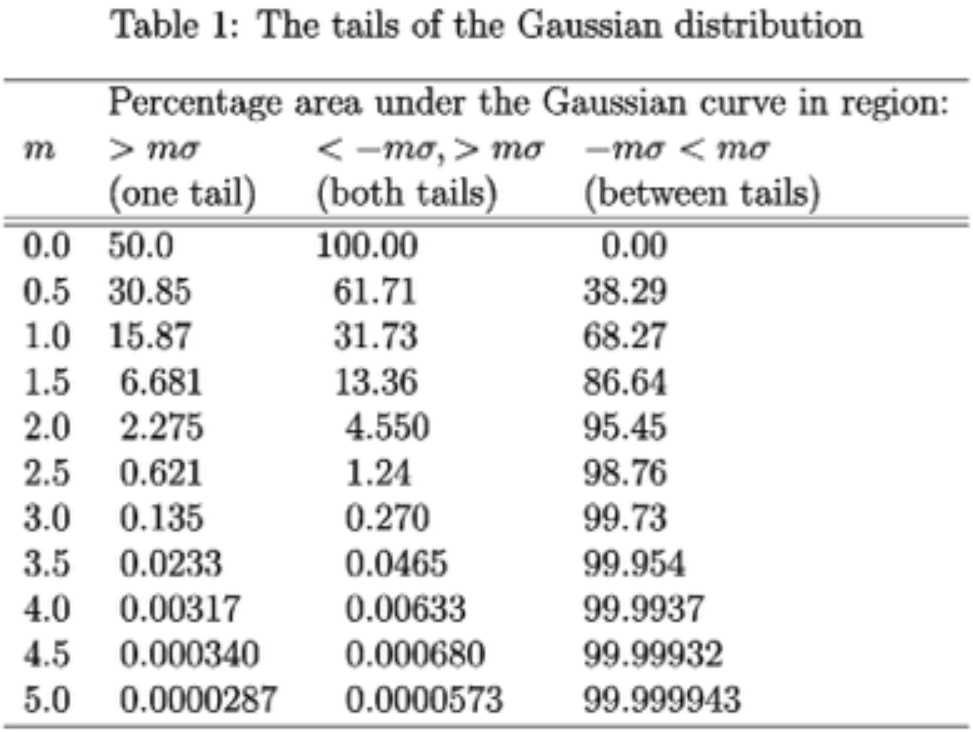
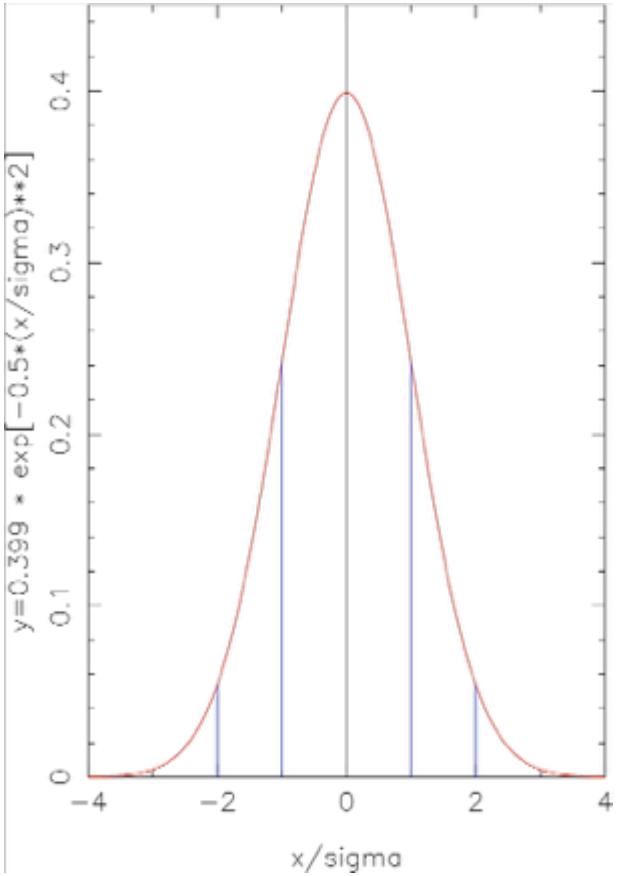


Table 1: The tails of the Gaussian distribution

	Percentage area under the Gaussian curve in region:		
m	$> m\sigma$	$< -m\sigma, > m\sigma$	$-m\sigma < m\sigma$
	(one tail)	(both tails)	(between tails)
0.0	50.0	100.00	0.00
0.5	30.85	61.71	38.29
1.0	15.87	31.73	68.27
1.5	6.681	13.36	86.64
2.0	2.275	4.550	95.45
2.5	0.621	1.24	98.76
3.0	0.135	0.270	99.73
3.5	0.0233	0.0465	99.954
4.0	0.00317	0.00633	99.9937
4.5	0.000340	0.000680	99.99932
5.0	0.0000287	0.0000573	99.999943



The Power Law Distribution

- $N(>L) = K\,L^{\gamma+1,}$ integral form, $N > L$, or
- $dN = (\gamma+1)\,K\,L^{\gamma}\,dL$, differential form, dN in dL
- not formally a probability distribution because $\int = \infty$;
- normally there are physical bounds so that it works
- scale-free:
  $f(cL) = (cL)^{\gamma} = Const\ \mathbf{x}\ (L)^{\gamma} = Const\ \mathbf{x}\ f(L)$
- Steep power laws $-4 < \gamma < 0$ pop up in astronomy frequently
- Criticality: earthquakes, stock-market fluctuations, forest fires, sub-networks on the internet, sand-piles…..Salpeter Mass Function, source counts (number-magnitude counts), primordial fluctuation spectrum….

The Power Law Distribution 2

Why do disasters occur?

There are many pitfalls and combinations of pitfalls

1. It is totally different in character from binomial - Poisson - Gaussian

Characterizing by means or variances completely misleading.

The Central Limit Theorem fails us badly.

2. The index!

- differential or integral?
- binning: uniform or on a $\Delta\mathrm{logL}$ scale? n.b. $d(logL) = Const\ \boldsymbol{x}\ L^{-1}\ \ dL$

The Power Law Distribution - Example

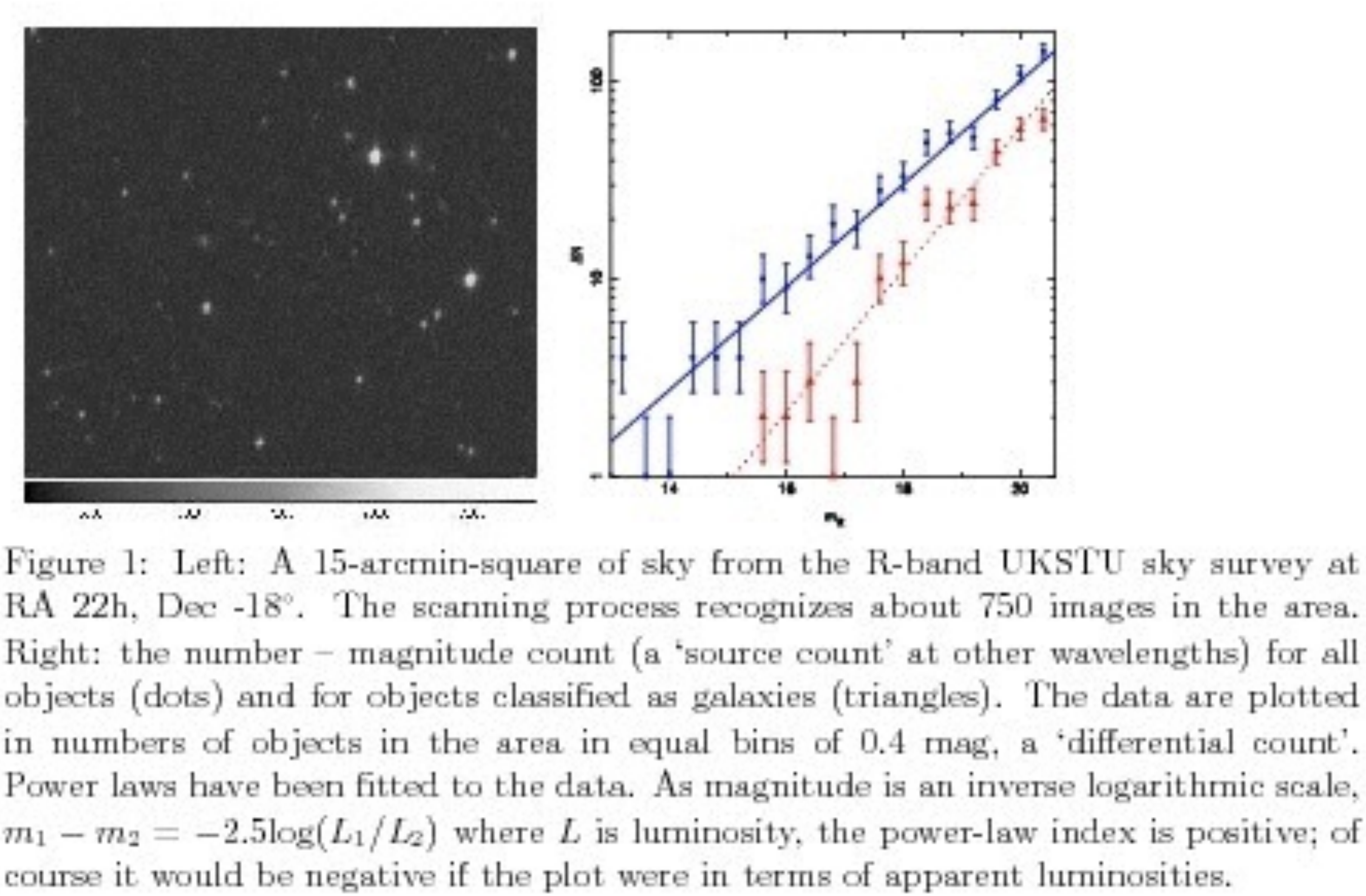


Figure 1: Left: A 15-arcmin-square of sky from the R-band UKSTU sky survey at RA 22h, Dec -18°. The scanning process recognizes about 750 images in the area. Right: the number – magnitude count (a 'source count' at other wavelengths) for all objects (dots) and for objects classified as galaxies (triangles). The data are plotted in numbers of objects in the area in equal bins of 0.4 mag, a 'differential count'. Power laws have been fitted to the data. As magnitude is an inverse logarithmic scale, $m_1 - m_2 = -2.5\log(L_1/L_2)$ where L is luminosity, the power-law index is positive; of course it would be negative if the plot were in terms of apparent luminosities.