Statistics in Astronomy

Jasper Wall

Bertinoro May 2009

University of British Columbia

Vancouver, B.C., Canada

Benvenuto!

1. Statistics and probability
2. Probability, Bayes, and Monte Carlo
3. Correlation and PCA
4. Hypothesis testing
5. Surveys and luminosity functions

JVW Bertinoro I

Statistics and Probability

Decision Time

Science is decision.

A list of what it is not science is infinite (and arguable):

- Building instruments
- Observing
- Reducing data
- Making pretty graphs
- Writing code
- Writing papers
- Reading the literature
- Learning tools: physics/astronomy/maths

Only **decision** counts.

We decide by comparing

Example: Is the faint smudge on an image a star or a galaxy?
- Measure FWHM of the point-spread function.
- Measure full-width-half-maximum, the FWHM of the image.
- The data set, the image of the object, is now represented by a *statistic*

► Decision!

Statistics are there for decision against a background.

Every measurement, parameter or value we derive requires an error estimate, a measure of range (expressed in terms of probability) that encompasses our belief of the true value of the parameter.

No measured quantity or property is of the slightest use in decision unless it has a `range quantity' attached.

What is or are statistics? Why?

A **statistic** is a quantity that summarizes data; it is the ultimate data-reduction.

It is a **property of the data** and nothing else. It may be a number, a mean for example, but it doesn't have to be.

It is a basis for using the data or experimental result to make a decision.

We need to know how to treat data with a view to decision,
to obtain the right statistics to use in drawing **statistical inference**.

It is the latter which is the branch of science; at times the term is loosely used to describe both the **descriptive values** and the science.

How to decide

The essence of **classical statistical analysis** is

(iii)the formulation of hypothesis,
(iv)the gathering of hypothesis-test data via experiment,
construction of a test-statistic.
(iii) comparison with the sampling distribution.

But we can't `rerun our experiments'.
Thus we don't know the underlying distributions of the variables:

- Small samples

- Poor experimental control

  - ► We have to be smarter than this

Probability:Distributions:Statistics:Inference

Statistics are combinations of Data – and Nothing Else

Example: **average**
- we expect it bears some relation to the true mean
- we calculate the *sampling distribution* ≡ the probability of various values it may assume if we (hypothetically) repeat the experiment many times.
- we then know the probability that some range around our single measurement will contain the true mean.

This is precisely the utility of statistics: they are laboriously-discovered combinations of observations which converge, for large sample sizes, to some underlying parameter we want to know.

Probability:Distributions:Statistics:Inference 3

the Bayesian way

A radically different way of making inferences focuses on the probabilities immediately, and to hell with statistics

Invert the reasoning just described: The **data** are unique and known!

Example: in the previous example it is the **mean that is unknown**, that should have probability attached to it. We instead calculate **the probability of various values of the mean**, given the data we have.

The approach comes far closer to answering the questions that we actually ask. Of course it allows us to make decisions.

So we should always use it, except for the buts:

but#1 - the brain works the other way
but#2 - other people work the other way, and we've gotta check them out
but#3 - the data may not be given to us in a form we can Bayesiate it
but#4 - there may not be a model

Probability is essential for us

(1) Astronomical measurements are subject to random measurement error and we need to have a common language of expression. If we quote an error, what is the unspoken assumption about it?

(2) The inability to do experiments on our subject matter leads us to draw conclusions by contrasting properties of controlled samples. They are usually 'too small', leading to 'statistical error'.

Example: `the distributions of luminosity in X-ray-selected Type I and Type II objects differ at the 95 per cent level of significance.'

Very often the strength of this conclusion is:
- dominated by the number of objects in the sample
- unaffected by observational error.

So: probability + conditionality + independence + Bayes' Theorem + prior + posterior probabilities.
=> **probability distributions**

We (still) cannot avoid statistics….

…and there are several reasons for this unfortunate situation:

5. Error (range) assignment - ours, and theirs – what do they mean?

2. How can data be used best? Or at all?

3. Correlation, testing the hypothesis, model fitting; how do we proceed?

4. Incomplete samples, samples from an experiment which cannot be rerun, upper limits; how can we use these to best advantage?

5. Others describe their data and conclusions in statistical terms. We need some self-defense.

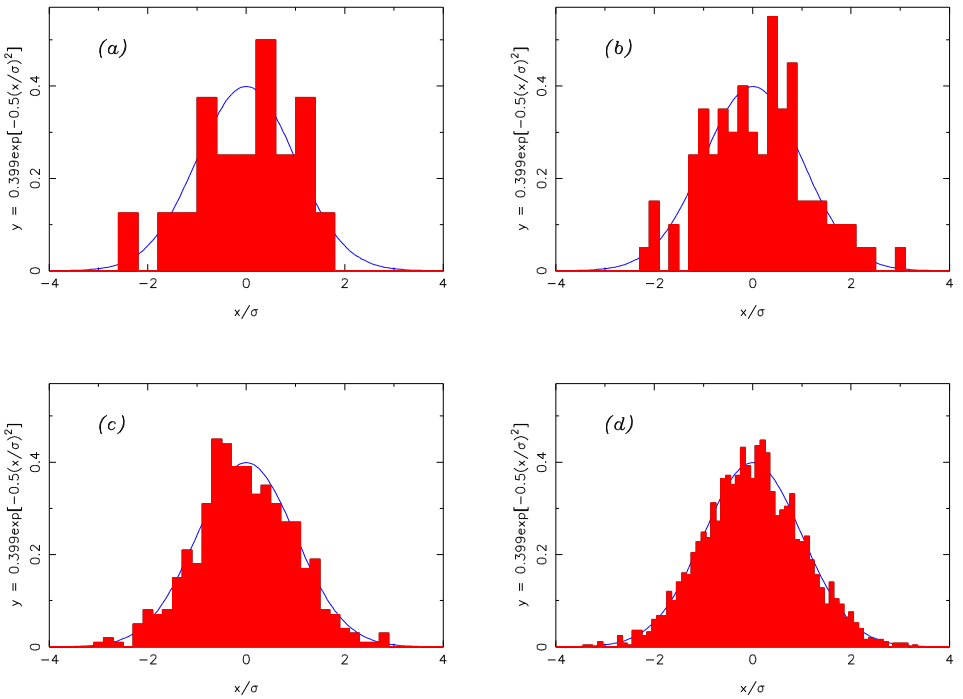
6. Above all, we must decide. The decision process cannot be done without some methodology, no matter how good the experiment.

Probability:Distributions:Statistics:Inference 2

'Probability' is crucial in the decision process

We have a built-in sense of probability

- from distributions or frequencies, which we 'know'
- from experience
- from data



Consider the eye-brain system observing an approaching person .....

.....It carries out a complete scientific experiment and makes a decision

Probability Distributions

Example: toss four `fair' coins. The probability of no heads is $(1/2)^4$; of one head $4 \times (1/2)^4$; of two heads $6 \times (1/2)^4$, etc. The sum of the possibilities for getting 0 heads to 4 heads is readily seen to be 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 the 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.

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

Probability Distributions 2

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 3

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

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

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

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

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

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 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**:

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

This coefficient can be derived in the following way. 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 $(1-\rho)^{(N-n)}$ 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 – an 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

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

For example, 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.

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

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

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

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 =** $\sqrt{}$ **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** ي **√t**, the **sky-limited** case.

Poisson Example, continued

There are the following further cases:

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

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

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

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

   for CCD of readout noise σ_{ccd}.

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

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

   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

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

from which it is easy to show that the mean is **μ** and the variance is **σ²**.

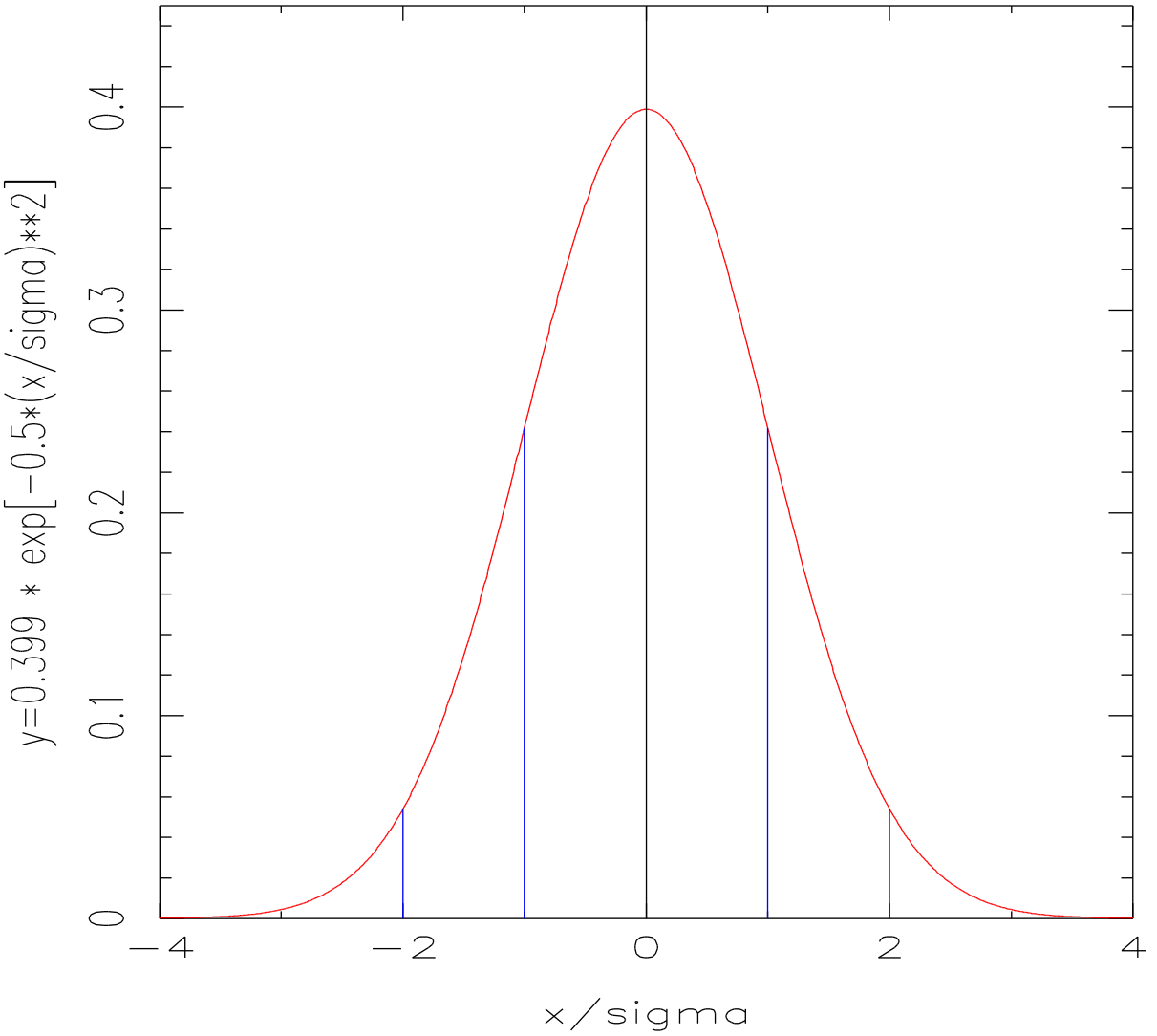
For the binomial when the sample size is very large, the discrete distribution tends to a continuous probability density

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

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 2



The Central Limit Theorem

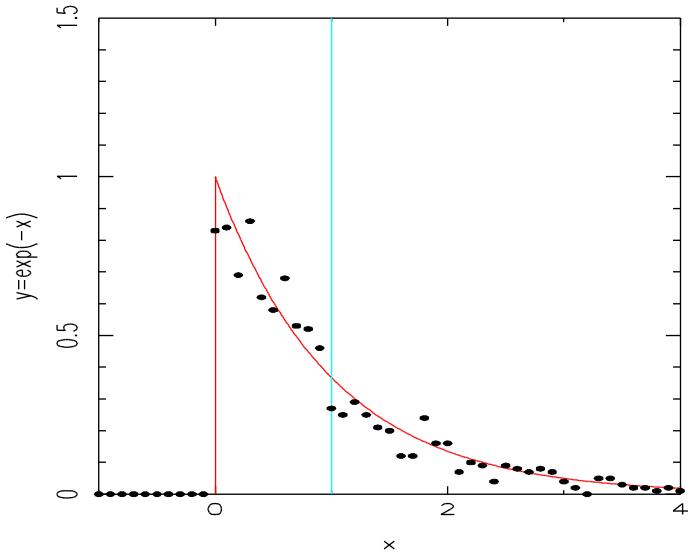
The true importance of the Gaussian distribution and its dominant position in experimental science, stems from the **Central Limit Theorem**. A non-rigorous statement of this is as follows.

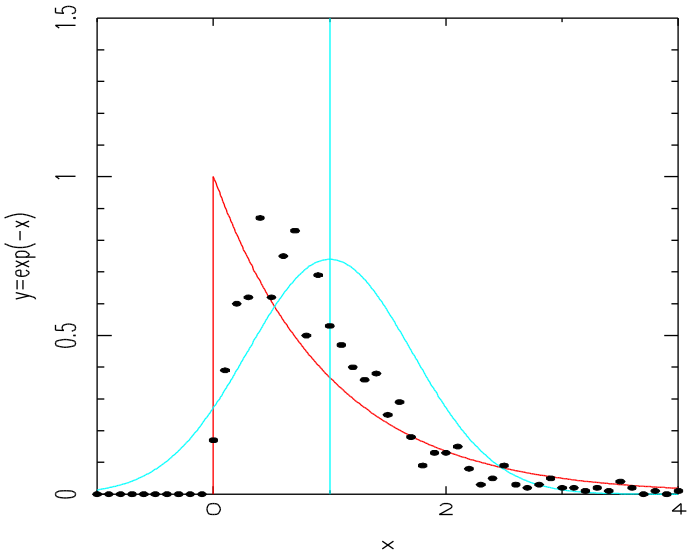
> Form averages M_n from repeatedly drawing n samples from a population x_i with finite mean μ, variance σ^2. Then the distribution of
>
> $$[\frac{(M_n - \mu)}{\sigma/\sqrt{n}}] \rightarrow \text{Gaussian distribution}$$
>
> with mean 0, variance 1, as n $\rightarrow \infty$.

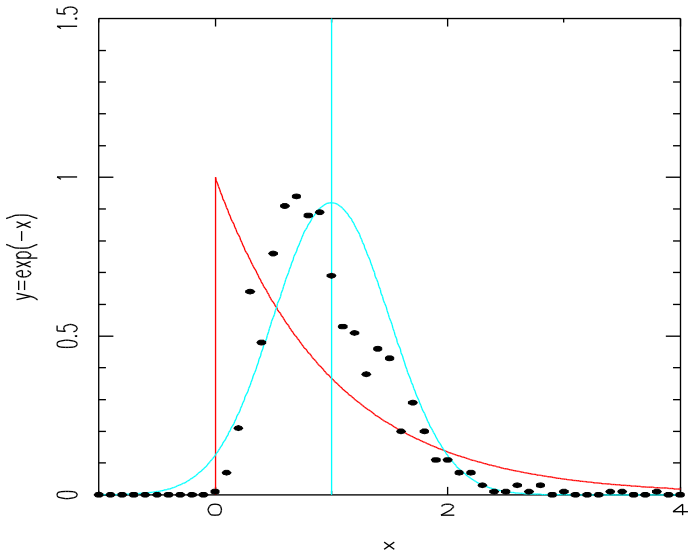
THIS MAY BE THE MOST REMARKABLE THEOREM EVER

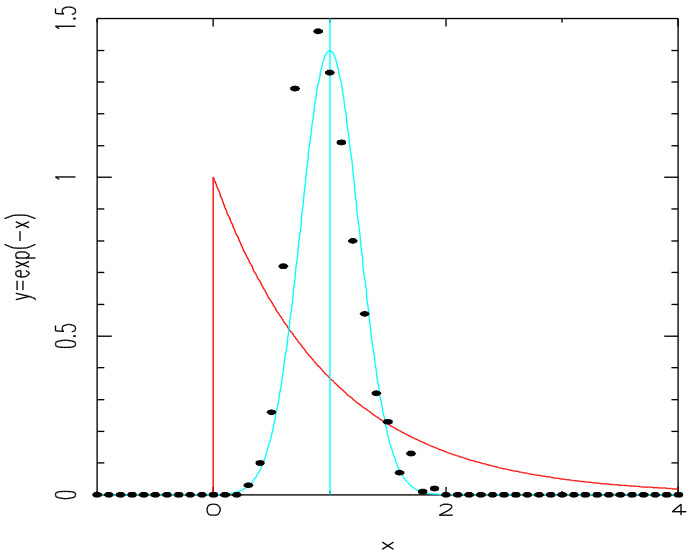
- It says that averaging will produce a Gaussian distribution of results - **no matter the shape of distribution from which the sample is drawn**.
- Eyeball integration counts!
- Errors on averaged samples will always look `Gaussian'.
- The Central Limit Theorem shapes our entire view of experimentation.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

The Central Limit Theorem – Example 1

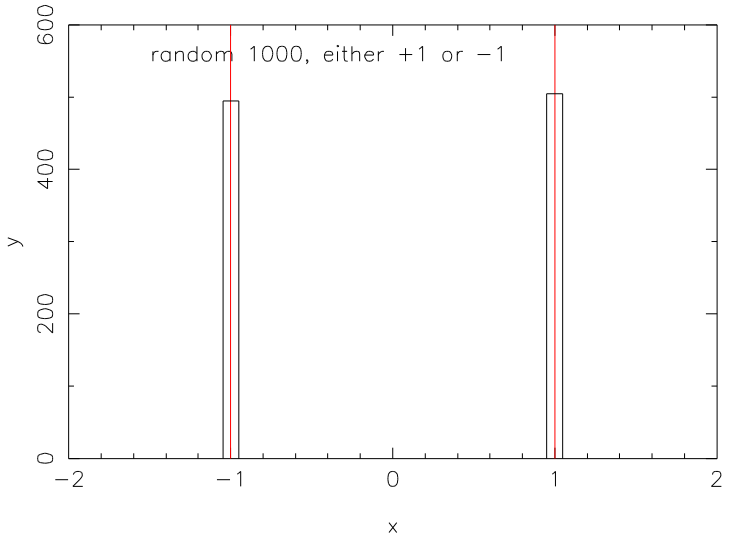


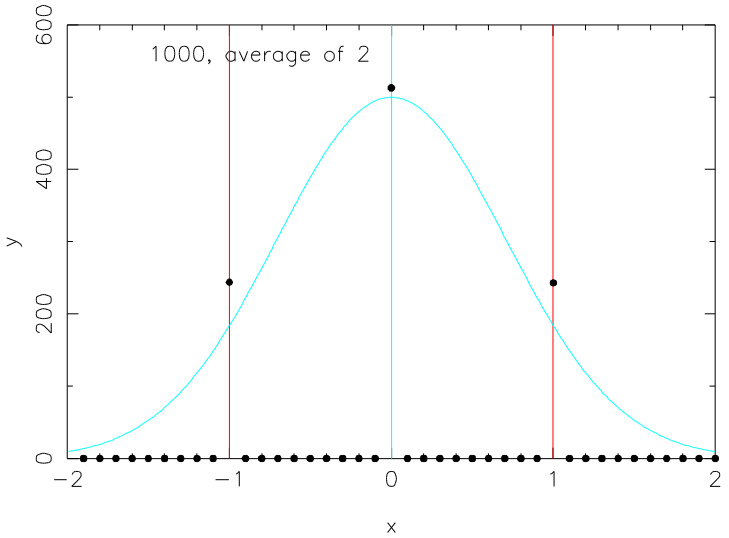


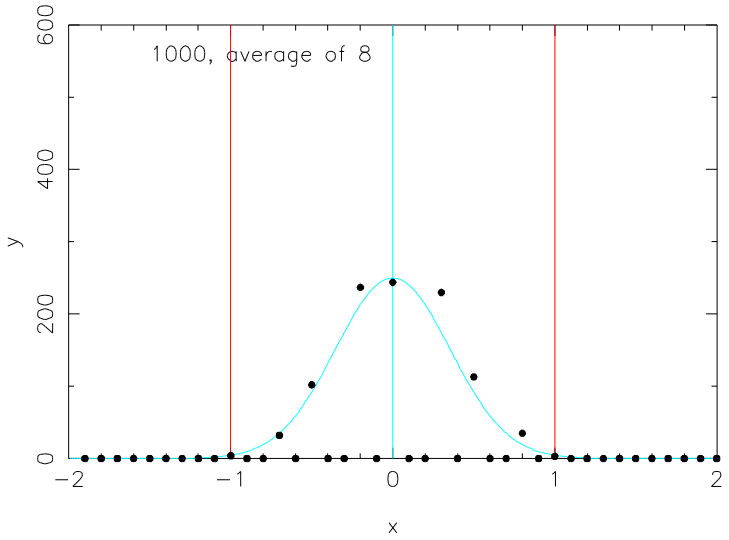


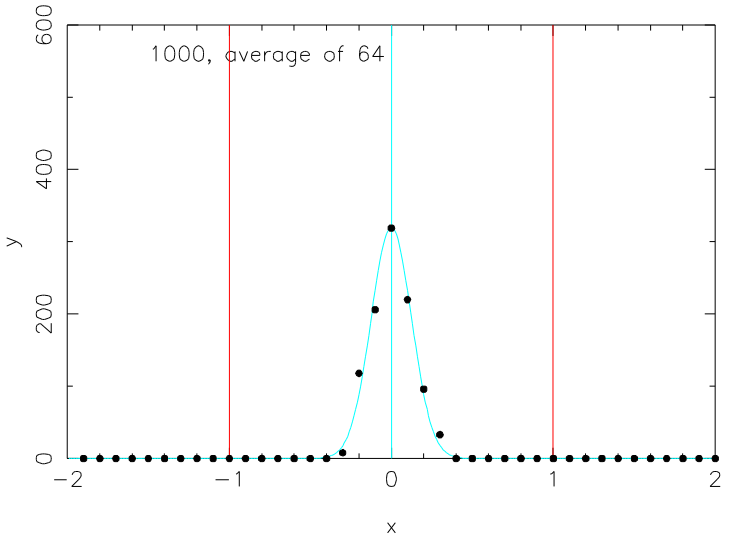


The Central Limit Theorem – Example 2





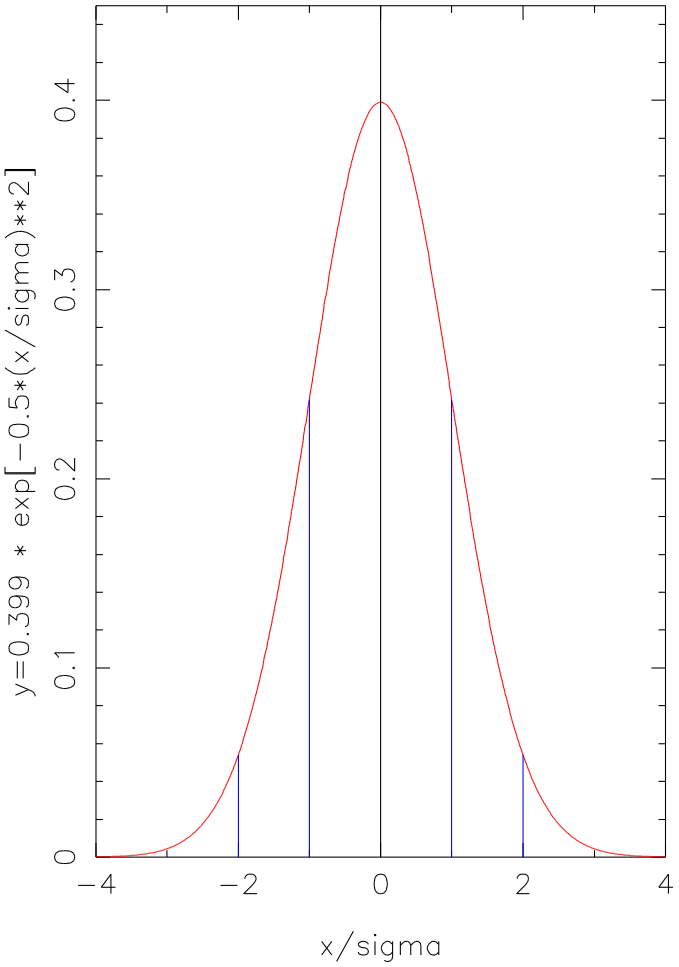




Gaussian tails

Table 1: The tails of the Gaussian distribution

m	Percentage area under the Gaussian curve in region: $> m\sigma$ (one tail)	$< -m\sigma, > m\sigma$ (both tails)	$-m\sigma < m\sigma$ (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



End Bertinoro 1