Tutorial on Markov Chain Monte Carlo Simulations and Their Statistical Analysis (in Fortran)

Bernd Berg

Singapore MCMC Meeting, March 2004

Overview

1. Lecture I/II: Statistics as Needed.

2. Lecture II: Markov Chain Monte Carlo (MC).

3. Lecture III: Statistical Analysis of MC Data.

4. Lecture IV: Multicanonical Simulations.

Reference

Markov Chain Monte Carlo Simulations and Their Statistical Analysis I (with web-based Fortran code)

by

Bernd A. Berg

To be published by World Scientific.

Contents of Lecture I

1. Probability Distributions and Sampling
2. Random Numbers and the Fortran Code
3. Gaussian Distribution, Confidence Intervals and Heapsort
4. The Central Limit Theorem and Binning
5. Gaussian Error Analysis for Large and Small Samples
6. χ^2 Distribution, the Error of the Error Bar, and the Variance Ratio Test
7. The Jackknife Approach

Probability Distributions and Sampling

Probability or chance is one of those basic concepts in science which elude a derivation from more fundamental principles.

A **sample space** is a set of **events**, also called **measurements** or **observations**, whose occurrence depends on chance.

Carrying out independent repetitions of the same experiment is called **sampling**. The outcome of each experiment provides an event called **data point**. In N such experiments we may find the event A to occur with **frequency** n, $0 \leq n \leq N$. The **probability** assigned to the event A is a number $P(A)$, $0 \leq P(A) \leq 1$, so that

$$P(A) \;=\; \lim_{N\to\infty} \frac{n}{N}. \tag{1}$$

This equation is sometimes called the **frequency definition of probability**.

Let us denote by $P(a,b)$ the probability that $x^r \in [a,b]$ where x^r is a **random variable** drawn in the interval $(-\infty,+\infty)$ with the **probability density** $f(x)$.

Then,

$$P(a,b) = \int_a^b f(x)\ dx. \tag{2}$$

Knowledge of all probabilities $P(a,b)$ implies

$$f(x) = \lim_{y \to x^-} \frac{P(y,x)}{x-y} \ \geq\ 0\ . \tag{3}$$

The **(cumulative) distribution function** of the random variable x^r is defined as

$$F(x) = P(x^r \leq x) = \int_{-\infty}^{x} f(x)\, dx\ . \tag{4}$$

A particularly important case is the **uniform probability distribution** for random numbers between $[0,1)$,

$$u(x) = \begin{cases} 1 & \text{for} \quad 0 \leq x < 1; \\ 0 & \text{elsewhere.} \end{cases} \tag{5}$$

The corresponding distribution function is

$$U(x) = \int_{-\infty}^{x} u(x)\, dx = \begin{cases} 0 & \text{for} \quad x < 0; \\ x & \text{for} \quad 0 \leq x \leq 1; \\ 1 & \text{for} \quad x > 1. \end{cases} \tag{6}$$

Remarkably, the uniform distribution allows for the construction of general probability distributions. Let

$$y = F(x) = \int_{-\infty}^{x} f(x')\, dx'$$

and assume that inverse $x = F^{-1}(y)$ exist. For y^r being a uniformly distributed random variable in the range $[0, 1)$ it follows that

$$x^r = F^{-1}(y^r) \tag{7}$$

is distributed according to the probability density $f(x)$.

An example is the **Cauchy distribution**

$$f_c(x) = \frac{\alpha}{\pi\,(\alpha^2 + x^2)} \quad \text{and} \quad F_c(x) = \int_{-\infty}^{x} f_c(x')\,dx' = \frac{1}{2} + \frac{1}{\pi}\,\tan^{-1}\left(\frac{x}{\alpha}\right),\ \alpha > 0\,. \tag{8}$$

The Cauchy distributed random variable x^r is generated from the uniform $y^r \in [0,1)$ through

$$x^r = \alpha\,\tan(2\pi y^r)\,. \tag{9}$$

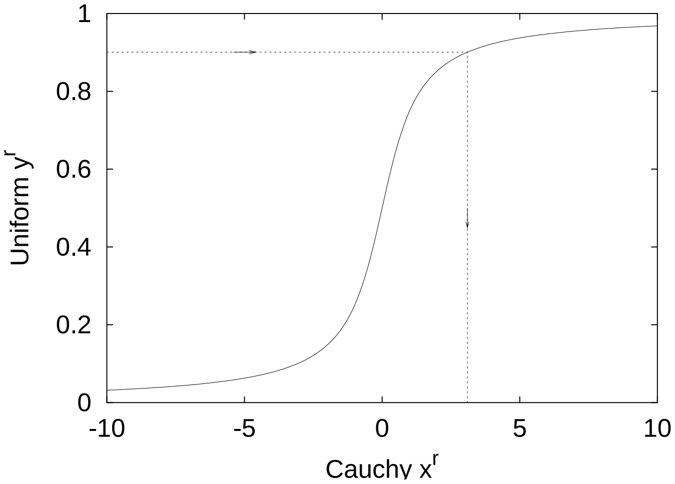


Figure 1: Mapping of the uniform to the Cauchy distribution.

Random Numbers and Fortran Code

According to Marsaglia and collaborators a list of desirable properties for random number generators is:

(i) *Randomness*. The generator should pass stringent tests for randomness.

(ii) *Long period*.

(iii) *Computational efficiency*.

(iv) *Repeatability*. Initial conditions (seed values) completely determine the resulting sequence of random variables.

(v) *Portability*. Identical sequences of random variables may be produced on a wide variety of computers (for given seed values).

(vi) *Homogeneity*. All subsets of bits of the numbers are random.

Physicists have added a number of their applications as new tests. In particular the exact solution of the $2d$ Ising model is used. Here the random number generator by Marsaglia and collaborators is provided. It relies on a combination of two generators:

x_n from a lagged Fibonacci series $I_n = I_{n-r} - I_{n-s} \bmod 2^{24}$, $r = 97$, $s = 33$.

y_n from the arithmetic series $I - k,\ I - 2k,\ I - 3k,\ \dots,\ \bmod\,[2^{24} - 3]$.

For most applications this generator is a good compromise. Our Fortran code which implements Marsaglia random numbers consists of three subroutines:

`rmaset.f` to set the initial state of the random number generator.

`ranmar.f` which provides one random number per call.

`rmasave.f` to save the final state of the generator.

In addition, `rmafun.f` is a function version of `ranmar.f` and calls to these two routines are freely interchangeable. Related is also the subroutine `rmacau.f`, which generates Cauchy random numbers with $\alpha = 1$.

The subroutine `rmaset.f` initializes the generator to mutually independent sequences of random numbers for distinct pairs of

$$-1801 \le \texttt{iseed1} \le 29527 \ \ \text{and} \ \ -9373 \le \texttt{iseed2} \le 20708 \ . \tag{10}$$

This property makes the generator quite useful for parallel processing.

Table 1: Illustration of a start and a continuations run of the Marsaglia random number generator using the program `mar.f` with the default seeds (a0102_02).

```
RANMAR INITIALIZED.                  MARSAGLIA CONTINUATION.
idat, xr = 1  0.116391063            idat, xr = 1  0.495856345
idat, xr = 2  0.96484679             idat, xr = 2  0.577386141
idat, xr = 3  0.882970393            idat, xr = 3  0.942340136
idat, xr = 4  0.420486867            idat, xr = 4  0.243162394
extra xr =    0.495856345            extra xr =    0.550126791
```

How to run and get the FORTRAN code?

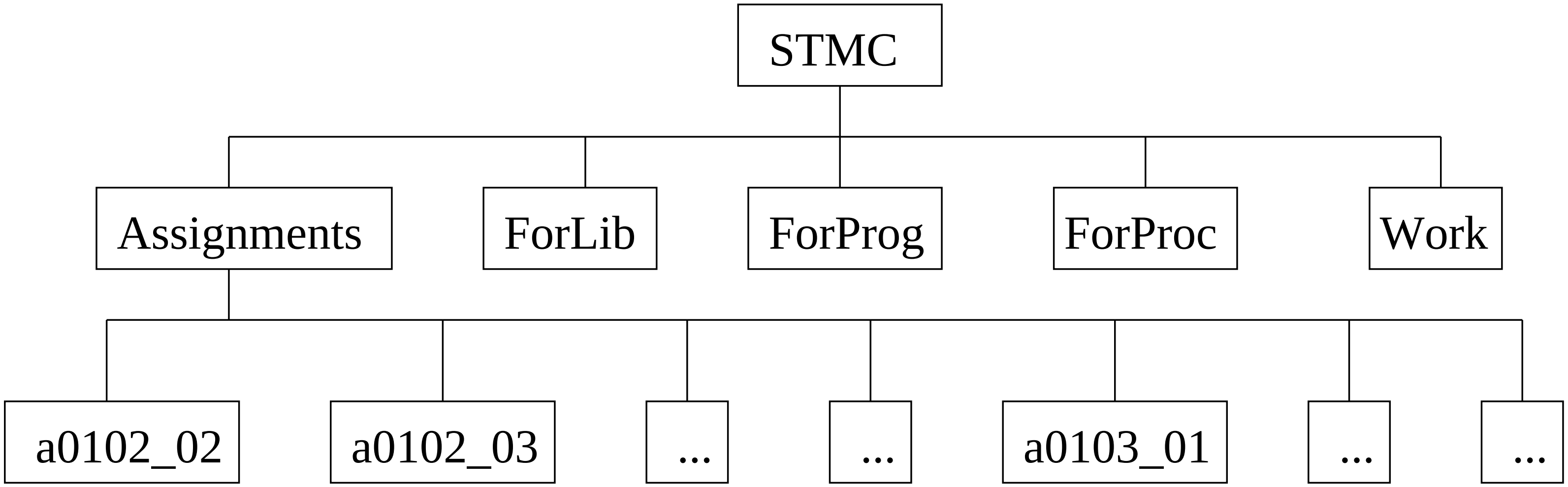


Figure 2: The Fortran routines are provided and prepared to run in the a tree structure of folders depicted in this figure. This tree of directories unfolds from the downloaded file.

To **download** the Fortran code book visit the website

$$\mathbf{http://b\_berg.home.comcast.net/}$$

and follow the instructions given there.

The code is provided in the directories `ForLib`, `ForProg` and `ForProc`. `ForLib` contains a library of functions and subroutines which is closed in the sense that no reference to non-standard functions or subroutines outside the library is ever made. Fortran programs are contained in the folder `ForProg` and procedures for interactive use in `ForProc`.

Assignment: Marsaglia random numbers. Run the program `mar.f` to reproduce the results of table 1. Understand how to re-start the random number generator as well as how to perform different starts when the continuation data file `ranmar.d` does not exist. Note: You find `mar.f` in `ForProg/Marsaglia` and it includes subroutines from `ForLib`. To compile properly, `mar.f` has to be located two levels down from a root directory `STMC`. The solution is found in the subdirectory `Assignments/a0102_02`.

The hyperstructure of program dependencies introduced between the levels of the STMC directory tree should be kept!

Gaussian Distribution

The **Gaussian** or **normal distribution** is of major importance. Its probability density is

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2/(2\sigma^2)} \tag{11}$$

where σ^2 is the **variance** and $\sigma > 0$ the **standard deviation**. The Gaussian distribution function $G(x)$ is related to that of variance $\sigma^2 = 1$ by

$$G(x) = \int_{-\infty}^{x} g(x')\, dx' = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x/\sigma} e^{-(x'')^2/2}\, dx'' = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{x}{\sigma\sqrt{2}}\right) \ . \tag{12}$$

In principle we could now generate **Gaussian random numbers**. However, the numerical calculation of the inverse error function is slow and makes this an impractical procedure. Much faster is to express the product probability density of two independent Gaussian distributions in polar coordinates

$$\frac{1}{2\pi\,\sigma^2}\,e^{-x^2/(2\sigma^2)}\,e^{-y^2/(2\sigma^2)}\,dx\,dy = \frac{1}{2\pi\,\sigma^2}\,e^{-r^2/(2\sigma^2)}\,d\phi\,r dr\,,$$

and to use the relations

$$x^r = r^r\,\cos\phi^r \quad \text{and} \quad y^r = r^r\,\sin\phi^r\ . \tag{13}$$

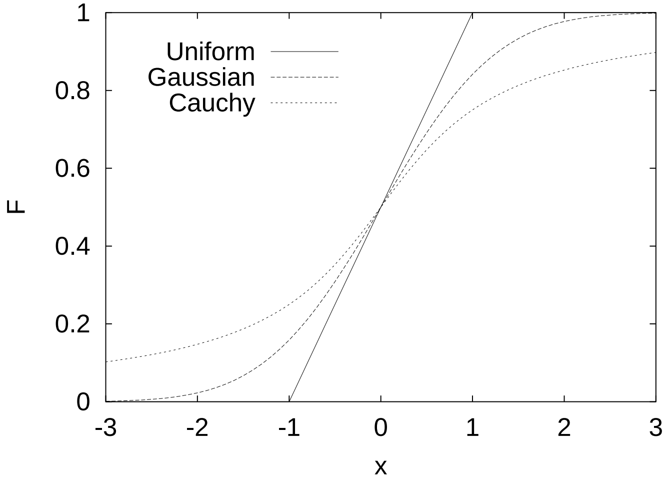


Figure 3: Cumulative distribution functions for a uniform, a Gaussian and a Cauchy distribution (Assignment a0101_02).

Confidence Intervals and Heapsort

Let a distribution function $F(x)$ and $0 \le q \le 1$ be given. One defines **q-tiles** (also called **quantiles** or **fractiles**) x_q by means of

$$F(x_q) \ = \ q \ . \tag{14}$$

The **median** $x_{\frac{1}{2}}$ is often the **typical** value of the random variable x^r.

The precise probability content of the confidence intervals

$$[x_q, x_{1-q}] = [-n\sigma, n\sigma] \ \text{ for } \ n = 1, 2$$

of the normal distribution is $p = 68.27\%$ for one σ and $p = 95.45\%$ for two σ.

The **peaked distribution function**

$$F_q(x) = \begin{cases} F(x) \text{ for } F(x) \le \frac{1}{2}, \\ 1 - F(x) \text{ for } F(x) > \frac{1}{2}. \end{cases} \tag{15}$$

provides a useful way to visualize probability intervals of a distribution.

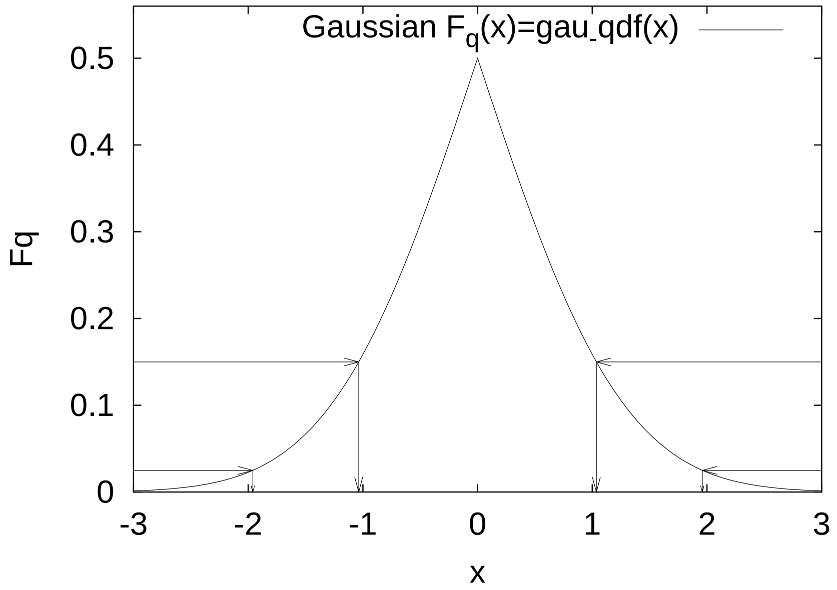


Figure 4: Gaussian peaked distribution function and estimates of x_q for the 70% (approximately $1\,\sigma$) and 95% (approximately $2\,\sigma$) confidence intervals.

Sampling provides us with an empirical distribution function and in practice the problem is to estimate confidence intervals from the empirical data.

Assume we generate n random number $x_1, ..., x_n$ according to a probability distribution $F(x)$. The n random numbers constitute then a **sample**.

We may re-arrange the x_i in increasing order. Denoting the smallest value by x_{π_1}, the next smallest by x_{π_2}, etc., we arrive at

$$x_{\pi_1} \leq x_{\pi_2} \leq \ldots \leq x_{\pi_n} \tag{16}$$

where $\pi_1, \ldots, \pi_n$ is a permutation of $1, \ldots, n$. Each of the x_{π_i} is then called an **order statistic**. An estimator for the distribution function $F(x)$ is then the **empirical distribution function**

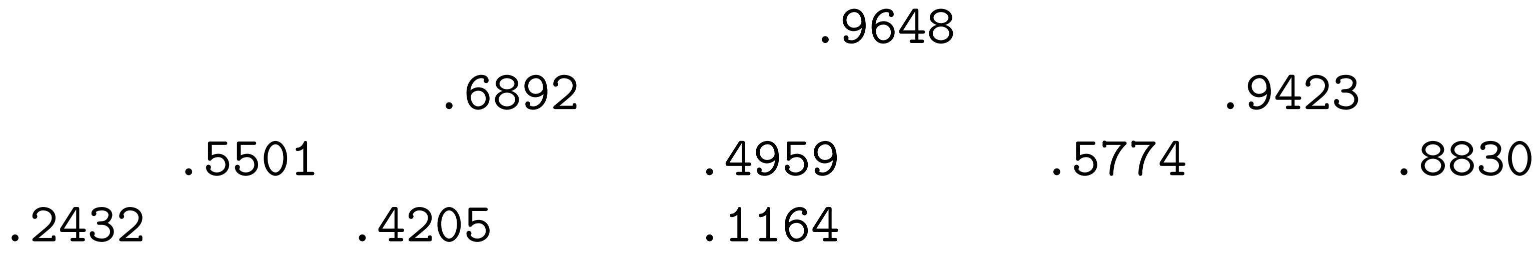
$$\overline{F}(x) = \frac{i}{n} \quad \text{for} \quad x_{\pi_i} \leq x < x_{\pi_{i+1}}, \; i = 0, 1, \ldots, n-1, n \tag{17}$$

with the definitions $x_{\pi_0} = -\infty$ and $x_{\pi_{n+1}} = +\infty$.

To calculate $\overline{F}(x)$ and the corresponding peaked distribution function, one needs an efficient way to **sort** n data values in ascending (or descending) order. This is provided by the **heapsort**, which relies on two steps: First the data are arranged in a heap, then the heap is sorted.

A **heap** is a partial ordering so that the number at the top is larger or equal than the two numbers in the second row, provided at least three numbers x_i exist. Each number of the second row has under it two smaller or equal numbers in the third row and so on, until all numbers $x_i, \ldots, x_n$ are exhausted. Example:

Table 2: The heap, generated from Marsaglia's first ten random numbers.



Finally, the heap is sorted. The computer time needed to succeed with the sorting process grows only like $n \log_2 n$ because there are $\log_2 n$ levels in the heap, see Knuth for a detailed discussion of sorting algorithms.

Example:

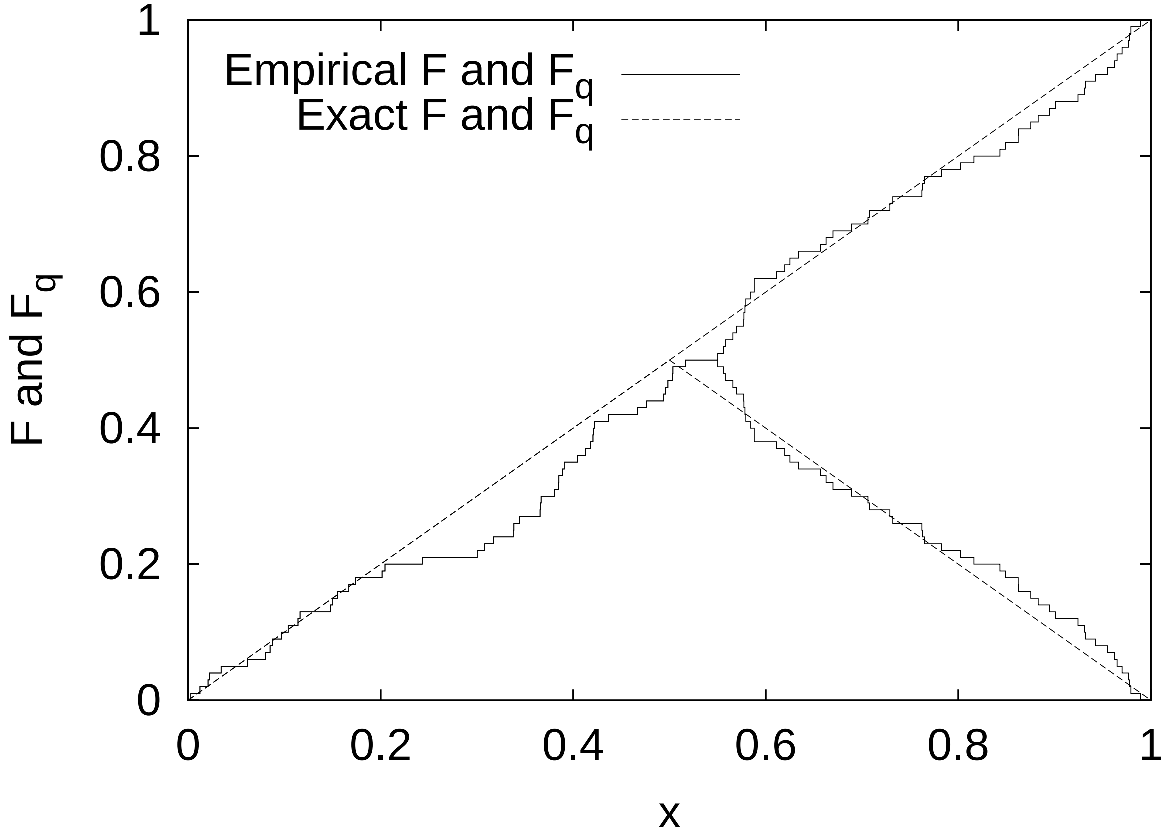


Figure 5: Cumulative and peaked distribution functions for Marsaglia's first 100 uniform random numbers (assignment a0106_02).

The Central Limit Theorem and Binning

How is the sum of two independent random variables

$$y^r \;=\; x_1^r + x_2^r \; . \tag{18}$$

distributed? We denote the probability density of y^r by $g(y)$. The corresponding cumulative distribution function is given by

$$G(y) = \int_{x_1+x_2 \le y} f_1(x_1)\; f_2(x_2)\; dx_1\; dx_2 == \int_{-\infty}^{+\infty} f_1(x)\; F_2(y-x)\; dx$$

where $F_2(x)$ is the distribution function of the random variable x_2^r. We take the derivative and obtain the probability density of y^r

$$g(y) \;=\; \frac{dG(y)}{dy} \;=\; \int_{-\infty}^{+\infty} f_1(x)\; f_2(y-x)\; dx \; . \tag{19}$$

The probability density of a sum of two independent random variables is the **convolution of the probability densities** of these random variables.

Example: Sums of uniform random numbers, corresponding to the sums of an uniformly distributed random variable $x^r \in (0,1]$:

(a) Let $y^r = x^r + x^r$, then

$$g_2(y) \;=\; \begin{cases} y & \text{for} \quad 0 \le y \le 1, \\ 2-y & \text{for} \quad 1 \le y \le 2, \\ 0 & \text{elsewhere.} \end{cases} \tag{20}$$

(b) Let $y^{\,r} = x^r + x^r + x^r$, then

$$g_3(y) \;=\; \begin{cases} y^2/2 & \text{for} \;\; 0 \le y \le 1, \\ (-2y^2+6y-3)/2 & \text{for} \;\; 1 \le y \le 2, \\ (y-3)^2/2 & \text{for} \;\; 2 \le y \le 3, \\ 0 & \text{elsewhere.} \end{cases} \tag{21}$$

The convolution (19) takes on a simple form in **Fourier space**. In statistics the **Fourier transformation** of the probability density is known as **characteristic function**, defined as the expectation value of e^{itx^r}:

$$\phi(t) \ = \ \langle e^{itx^r} \rangle \ = \ \int_{-\infty}^{+\infty} e^{itx} \, f(x) \, dx \ . \tag{22}$$

The characteristic function is particularly useful for investigating sums of random variables, $y^r = x_1^r + x_2^r$:

$$\phi_y(t) = \langle e^{(itx_1^r + itx_2^r)} \rangle = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{itx} \, e^{ity} \, f_1(x_1) \, f_2(x_2) \, dx_1 \, dx_2 = \phi_{x_1}(t) \, \phi_{x_2}(t) \ . \tag{23}$$

The characteristic function of a sum of random variables is the product of their characteristic functions. The result generalizes immediately to N random variables

$$y^r \ = \ x_1^r + \ldots + x_N^r \ . \tag{24}$$

The characteristic function of y^r is

$$\phi_y(t) \;=\; \prod_{i=1}^{N} \phi_{x_i}(t) \tag{25}$$

and the probability density of y^r is the Fourier back-transformation of this characteristic function

$$g(y) \;=\; \frac{1}{2\pi} \int_{-\infty}^{+\infty} dt\, e^{-ity}\, \phi_y(t) \;. \tag{26}$$

The **probability density of the sample mean** is obtained as follows: The arithmetic mean of y^r is $\overline{x}^{\,r} = y^r/N$. We denote the probability density of y^r by $g_N(y)$ and the probability density of the arithmetic mean by $\widehat{g}_N(\overline{x})$. They are related by

$$\widehat{g}_N(\overline{x}) \;=\; N\, g_N(N\overline{x}) \;. \tag{27}$$

This follows by substituting $y = N\overline{x}$ into $g_N(y)\,dy$:

$$1 = \int_{-\infty}^{+\infty} g_N(y)\,dy = \int_{-\infty}^{+\infty} g_N(N\overline{x})\,2d\overline{x} = \int_{-\infty}^{+\infty} \widehat{g}_N(\overline{x})\,d\overline{x}\ .$$

Example:

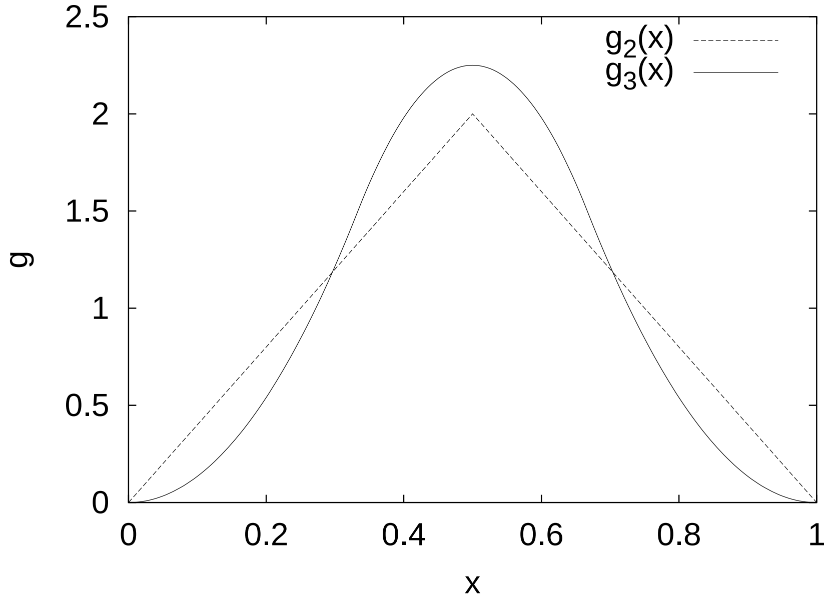


Figure 6: Probability densities for the arithmetic means of two and three uniformly distributed random variables, $\widehat{g}_2(\overline{x})$ and $\widehat{g}_3(\overline{x})$, respectively.

This suggests that sampling leads to convergence of the mean by reducing its variance. We use the characteristic function to understand the general behavior. The characteristic function of a sum of independent random variables is the product of their individual characteristic functions

$$\phi_y(t) \;=\; [\phi_x(t)]^N \;. \tag{28}$$

The characteristic function for the corresponding arithmetic average is

$$\phi_{\overline{x}}(t) \;=\; \int_{-\infty}^{+\infty} d\overline{x}\, e^{it\overline{x}}\, \widehat{g}_N(\overline{x}) \;=\; \int_{-\infty}^{+\infty} N d\overline{x}\, e^{it\overline{x}}\, g_N(N\overline{x})$$

$$= \int_{-\infty}^{+\infty} dy\, \exp\left(i\,\frac{t}{N}\,y\right)\, g_N(y)\;.$$

Hence,

$$\phi_{\overline{x}}(t) \;=\; \phi_y\left(\frac{t}{N}\right) \;=\; \left[\phi_x\left(\frac{t}{N}\right)\right]^N \;. \tag{29}$$

Example: The normal distribution.

The characteristic function is obtained by Gaussian integration

$$\phi(t) \;=\; \exp\left(-\frac{1}{2}\sigma^2 t^2\right) \; . \tag{30}$$

Defining $y^r = x^r + x^r$ we have

$$\phi_y(t) \;=\; [\phi(t)]^2 \;=\; \exp\left(-\frac{1}{2}2\sigma^2 t^2\right) \; . \tag{31}$$

This is the characteristic function of a Gaussian with variance $2\sigma^2$. We obtain the characteristic function of the arithmetic average $\overline{x}^{\,r} = y^r/2$ by the substitution $t \to t/2$:

$$\phi_{\overline{x}}(t) \;=\; \exp\left(-\frac{1}{2}\frac{\sigma^2}{2}t^2\right) \; . \tag{32}$$

The variance is reduced by a factor of two.

The Central Limit Theorem

To simplify the equations we restrict ourselves to $\widehat{x} = 0$. Let us consider a probability density $f(x)$ and assume that its moment exists, implying that the characteristic function is a least two times differentiable, so that

$$\phi_x(t) \;=\; 1 \;-\; \frac{\sigma_x^2}{2}\, t^2 \;+\; \mathcal{O}(t^3)\;. \tag{33}$$

The leading term reflects the the normalization of the probability density and the first moment is $\phi'(0) = \widehat{x} = 0$. The characteristic function of the mean becomes

$$\phi_{\overline{x}}(t) \;=\; \left[1 \;-\; \frac{\sigma_x^2}{2N^2} t^2 \;+\; \mathcal{O}\left(\frac{t^3}{N^3}\right)\right]^N \;=\; \exp\left[-\frac{1}{2}\,\frac{\sigma_x^2}{N}\, t^2\right] + \;\mathcal{O}\left(\frac{t^3}{N^2}\right)\;.$$

The probability density of the arithmetic mean $\overline{x}^{\,r}$ converges towards the Gaussian probability density with variance

$$\sigma^2(\overline{x}^{\,r}) \;=\; \frac{\sigma^2(x^r)}{N}\;. \tag{34}$$

A Counter example: The Cauchy distribution provides an instructive, case for which the central limit theorem does not work. This is expected as its second moment does not exist. Nevertheless, the characteristic function of the Cauchy distribution exists. For simplicity we take $\alpha = 1$ and get

$$\phi(t) = \int_{-\infty}^{+\infty} dx \, \frac{e^{itx}}{\pi\,(1+x^2)} \;=\; \exp(-|t|) \;. \tag{35}$$

The integration involves the residue theorem. Using equation (29) for the characteristic function of the mean of N random variables, we find

$$\phi_{\overline{x}}(t) = \left[\exp\left(-\frac{|t|}{N}\right)\right]^n \;=\; \exp(-|t|) \;. \tag{36}$$

The surprisingly simple result is that the probability distribution for the mean values of N independent Cauchy random variables agrees with the probability distribution of a single Cauchy random variable. Estimates of the Cauchy mean cannot be obtained by sampling. Indeed, the mean does not exist.

Binning

The notion of introduced here should not be confused with histogramming! Binning means here that we group NDAT data into NBINS bins, where each binned data point is the arithmetic average of

$$\texttt{NBIN} = [\texttt{NDAT}/\texttt{NBINS}] \quad \text{(Fortran integer division.)}$$

original data points. Preferably NDAT is a multiple of NBINS. The purpose of the binning procedure is twofold:

1. When the the central limit theorem applies, the binned data will become practically Gaussian, as soon as NBIN becomes large enough. This allows to apply Gaussian error analysis methods even when the original are not Gaussian.

2. When data are generated by a Markov process subsequent events are correlated. For binned data these correlations are reduced and can in practical applications be neglected, once NBIN is sufficiently large when compared to the autocorrelation time,

Example:

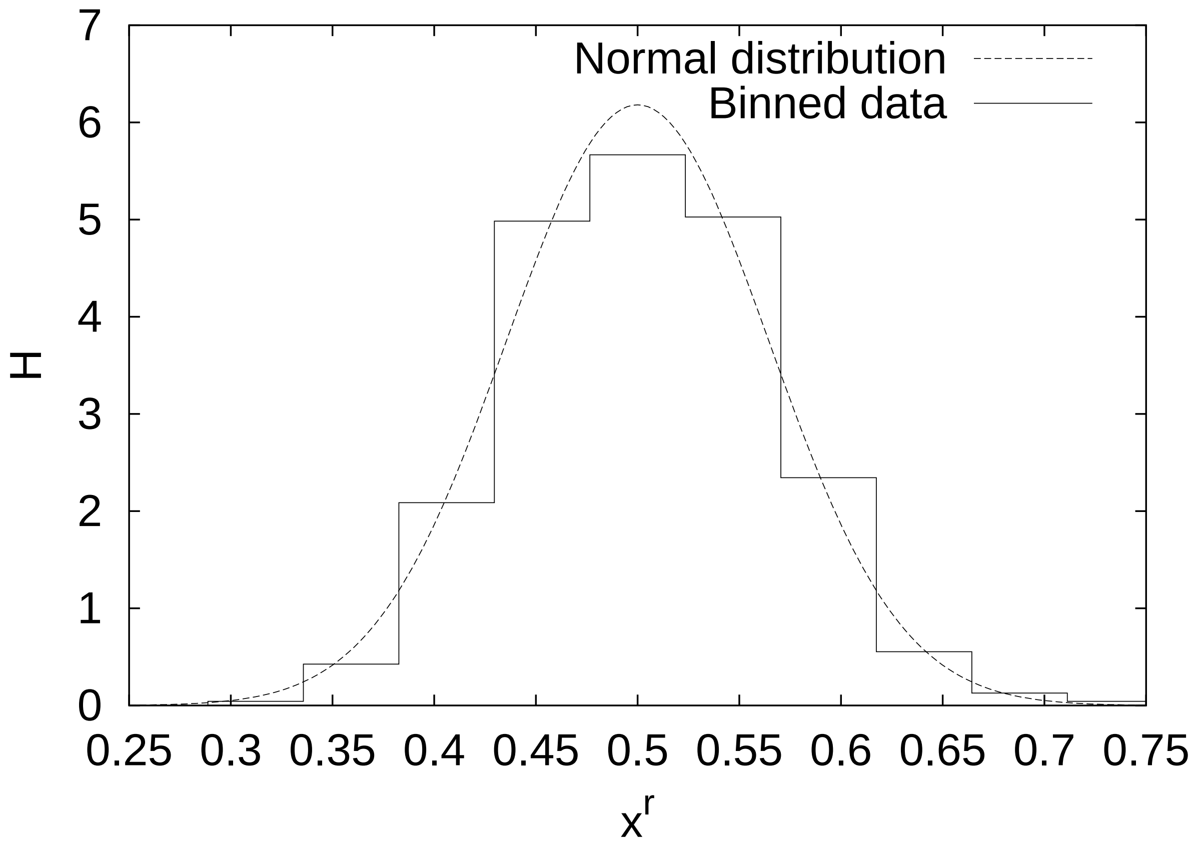


Figure 7: Comparison of a histogram of 500 binned data with the normal distribution $\sqrt{(120/\pi)}\,\exp[-120\,(x-1/2)^2]$. Each binned data point is the average of 20 uniformly distributed random numbers. Assignment a0108_02.

Gaussian Error Analysis for Large and Small Samples

The central limit theorem underlines the importance of the normal distribution. Assuming we have a large enough sample, the arithmetic mean of a suitable expectation value becomes normally distributed and the calculation of the confidence intervals is reduced to studying the normal distribution. It has become the convention to use the **standard deviation** of the sample mean

$$\sigma = \sigma(\overline{x}^{\,r}) \quad \text{with} \quad \overline{x}^{\,r} = \sum_{i=1}^{N} x_i^r \tag{37}$$

to indicate its confidence intervals $[\widehat{x} - n\sigma, \widehat{x} + n\sigma]$ (the dependence of σ on N is suppressed). For a Gaussian distribution equation (12) yields the probability content p of the confidence intervals (37) to be

$$p = p(n) = G(n\sigma) - G(-n\sigma) = \frac{1}{\sqrt{2\pi}} \int_{-n}^{+n} dx\, e^{-\frac{1}{2}x^2} = \mathrm{erf}\left(\frac{n}{\sqrt{2}}\right) \ . \tag{38}$$

Table 3: Probability content p of Gaussian confidence intervals $[\widehat{x}-n\sigma,\widehat{x}+n\sigma]$, $n=1,\dots,6$, and $q=(1-p)/2$. Assignment a0201_01.

n	1	2	3	4	5
p	.68	.95	1.0	1.0	1.0
q	.16	.23E-01	.13E-02	.32E-04	.29E-06

In practice the roles of $\overline{x}$ and $\widehat{x}$ are interchanged: One would like to know the likelihood that the **unknown** exact expectation value $\widehat{x}$ will be in a certain confidence interval around the measured sample mean. The relationship

$$\overline{x}\in[\widehat{x}-n\sigma,\widehat{x}+n\sigma] \iff \widehat{x}\in[\overline{x}-n\sigma,\overline{x}+n\sigma] \tag{39}$$

solves the problem. Conventionally, these estimates are quoted as

$$\widehat{x} \;=\; \overline{x}\pm\triangle\overline{x} \tag{40}$$

where the **error bar** $\triangle\overline{x}$ is often an **estimator** of the exact standard deviation.

An obvious estimator for the variance σ_x^2 is

$$(s'^{\,r}_x)^2 \;=\; \frac{1}{N}\sum_{i=1}^{N}(x_i^r - \overline{x}^{\,r})^2 \tag{41}$$

where the prime indicates that we shall not be happy with it, because we encounter a **bias**. An estimator is said to be biased when its expectation value does not agree with the exact result. In our case

$$\langle (s'^{\,r}_x)^2 \rangle \;\neq\; \sigma_x^2 \;. \tag{42}$$

An estimator whose expectation value agrees with the true expectation value is called **unbiased**. For the variance it is rather straightforward to construct an unbiased estimator $(s_x^r)^x$. The bias of the definition (41) comes from replacing the exact mean $\widehat{x}$ by its estimator $\overline{x}^{\,r}$. The latter is a random variable, whereas the former is just a number.

Some algebra shows that the desired **unbiased estimator of the variance** is given by

$$(s_x^r)^2 = \frac{N}{N-1}(s_x'^{\,r})^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i^r - \overline{x}^{\,r})^2 \; . \tag{43}$$

Correspondingly, the unbiased estimator of the variance of the sample mean is

$$(s_{\overline{x}}^r)^2 = \frac{1}{N(N-1)}\sum_{i=1}^{N}(x_i^r - \overline{x}^{\,r})^2 \; . \tag{44}$$

Gaussian difference test:

In practice one is often faced with the problem to compare two different empirical estimates of some mean. How large must $D = \overline{x} - \overline{y}$ be in order to indicate a real difference? The quotient

$$d^r = \frac{D^r}{\sigma_D} \tag{45}$$

is normally distributed with expectation zero and variance one, so that

$$P \;=\; P(|d^r| \le d) \;=\; G_0(d) - G_0(-d) \;=\; 1 - 2\,G_0(-d) \;=\; \mathrm{erf}\left(\frac{d}{\sqrt{2}}\right) \;. \tag{46}$$

The **likelihood that the observed difference** $|\overline{x} - \overline{y}|$ **is due to chance** is defined to be

$$Q \;=\; 1 - P \;=\; 2\,G_0(-d) \;=\; 1 - \mathrm{erf}\left(\frac{d}{\sqrt{2}}\right) \;. \tag{47}$$

If the assumption is correct, then Q is a uniformly distributed random variable in the range $[0, 1)$. **Examples** are:

Table 4: Gaussian difference tests (assignment a0201_06).

$\overline{x}_1 \pm \sigma_{\overline{x}_1}$	1.0 ± 0.1	1.0 ± 0.1	1.0 ± 0.1	1.0 ± 0.05	1.000 ± 0.025
$\overline{x}_2 \pm \sigma_{\overline{x}_2}$	1.2 ± 0.2	1.2 ± 0.1	1.2 ± 0.0	1.2 ± 0.00	1.200 ± 0.025
Q	0.37	0.16	0.046	0.000063	0.15×10^{-7}

Gosset's Student Distribution

We ask the question: What happens with the Gaussian confidence limits when we replace the variance $\sigma^2_{\overline{x}}$ by its estimator $s^2_{\overline{x}}$ in statements like

$$\frac{|\overline{x}-\widehat{x}|}{\sigma_{\overline{x}}} < 1.96 \ \ \text{with} \ \ 95\% \ \ \text{probability}.$$

For sampling from a Gaussian distribution the answer was given by Gosset, who published his article 1908 under the pseudonym *Student* in Biometrika. He showed that the distribution of the random variable

$$t^r = \frac{\overline{x}^{\,r} - \widehat{x}}{s^r_{\overline{x}}} \tag{48}$$

is given by the probability density

$$f(t) = \frac{1}{(N-1)\,B(1/2,(N-1)/2)} \left(1 + \frac{t^2}{N-1}\right)^{-\frac{N}{2}} . \tag{49}$$

Here $B(x,y)$ is the beta function. The fall-off is a power law $|t|^{-f}$ for $|t| \to \infty$, instead of the exponential fall-off of the normal distribution. Confidence probabilities of the Student distribution are:

N \ S	1.0000	2.0000	3.0000	4.0000	5.0000
2	.50000	.70483	.79517	.84404	.87433
3	.57735	.81650	.90453	.94281	.96225
4	.60900	.86067	.94233	.97199	.98461
8	.64938	.91438	.98006	.99481	.99843
16	.66683	.93605	.99103	.99884	.99984
32	.67495	.94567	.99471	.99963	.99998
64	.67886	.95018	.99614	.99983	1.0000
INFINITY:	.68269	.95450	.99730	.99994	1.0000

For $N \leq 4$ we find substantial deviations from the Gaussian confidence levels. Up to two standard deviations reasonable approximations of Gaussian confidence limits are obtained for $N \geq 16$ data. If desired, the Student distribution function can always be used to calculate the exact confidence limits. When the central limit theorem applies, we can **bin a large set of non-Gaussian data into 16 almost Gaussian data** and reduce the error analysis to Gaussian methods.

Student difference test

This test is a generalization of the Gaussian difference test. It takes into account that only a finite number of events are sampled. As before it is assumed that the events are drawn from a normal distribution. Let the following data be given

$$\begin{aligned} \overline{x} \ \text{ calculated from } \ M \ \text{ events, } \ i.e., \ \sigma^2_{\overline{x}} &= \sigma^2_x/M && (50) \\ \overline{y} \ \text{ calculated from } \ N \ \text{ events, } \ i.e., \ \sigma^2_{\overline{y}} &= \sigma^2_y/N && (51) \end{aligned}$$

and an unbiased estimators of the variances

$$s^2_{\overline{x}} = s^2_x/M = \frac{\sum_{i=1}^M (x_i - \overline{x})^2}{M\,(M-1)} \ \text{ and } \ s^2_{\overline{y}} = s^2_y/N = \frac{\sum_{j=1}^N (y_j - \overline{y})^2}{N\,(N-1)} \ . \tag{52}$$

Under the **additional assumption** $\sigma^2_x = \sigma^2_y$ the discussion which leads to the Student distribution applies and the probability

$$P(|\overline{x} - \overline{y}| > d) \tag{53}$$

is determined by the Student distribution function in the same way as the probability of the Gaussian difference test is determined by the normal distribution.

Examples for the Student difference test for $\overline{x}_1 = 1.00 \pm 0.05$ from M data and $\overline{x}_2 = 1.20 \pm 0.05$ from N data (assignment a0203_03):

M	512	32	16	16	4	3	2
N	512	32	16	4	4	3	2
Q	0.0048	0.0063	0.0083	0.072	0.030	0.047	0.11

The Gaussian difference test gives $Q = 0.0047$. For $M = N = 512$ the Student Q value is practically identical with the Gaussian result, for $M = N = 16$ it has almost doubled. Likelihoods above a 5% cut-off, are only obtained for $M = N = 2$ (11%) and $M = 16$, $N = 4$ (7%). The latter result looks a bit surprising, because its Q value is smaller than for $M = N = 4$. The explanation is that for $M = 16$, $N = 4$ data one would expect the $N = 4$ error bar to be two times larger than the $M = 16$ error bar, whereas the estimated error bars are identical.

This leads to the problem: Assume data are sampled from the same normal distribution, when are two measured error bars consistent and when not?

χ^2 Distribution, Error of the Error Bar, Variance Ratio Test

The distribution of the random variable

$$(\chi^r)^2 = \sum_{i=1}^{N} (y_i^r)^2 \ , \tag{54}$$

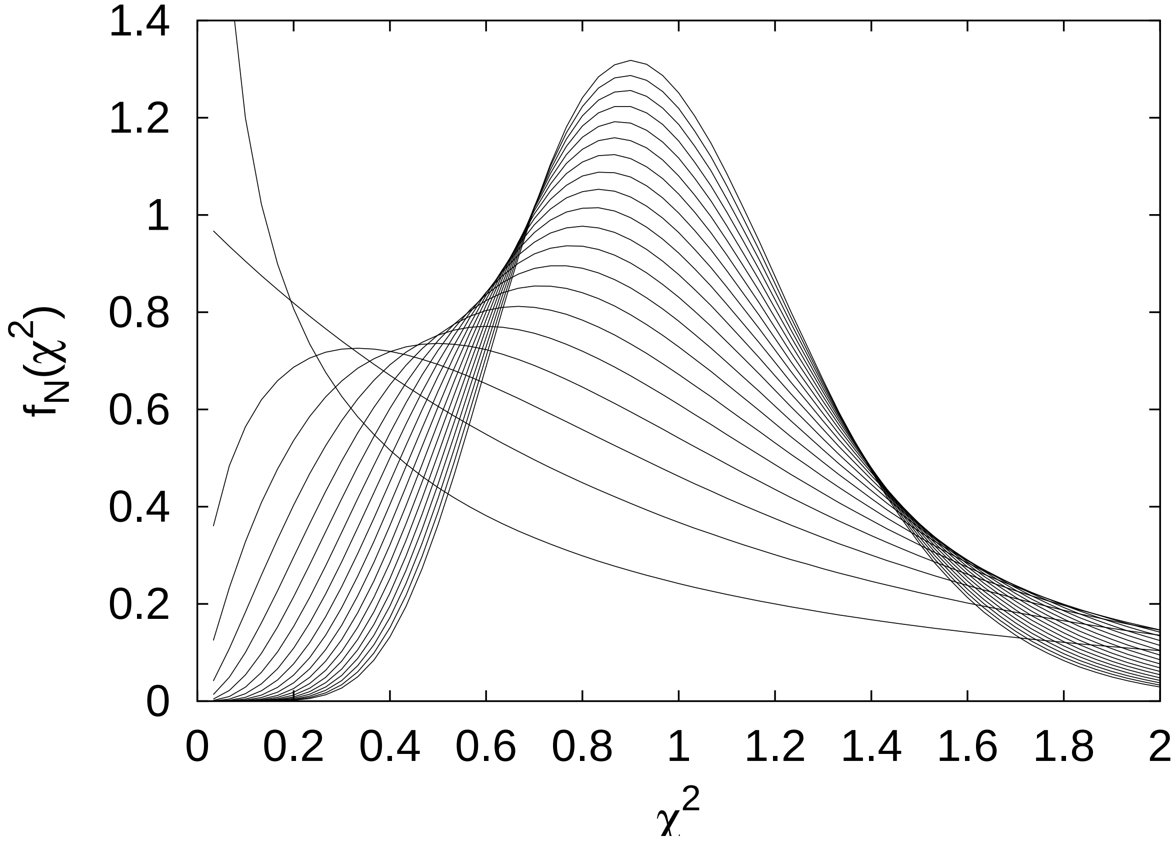
where each y_i^r is normally distributed, defines the $\boldsymbol{\chi^2}$ **distribution** with N degrees of freedom. The study of the variance $(s_x^r)^2$ of a Gaussian sample can be reduced to the χ^2-distribution with $f = N - 1$ degrees of freedom

$$(\chi_f^r)^2 = \frac{(N-1)\,(s_x^r)^2}{\sigma_x^2} = \sum_{i=1}^{N} \frac{(x_i^r - \overline{x}^{\,r})^2}{\sigma_x^2} \ . \tag{55}$$

The probability density of χ^2 **per degree of freedom (pdf)** is

$$f_N(\chi^2) \ = \ N f(N\chi^2) \ = \frac{a\ e^{-a\chi^2}\ \left(a\chi^2\right)^{a-1}}{\Gamma(a)} \ \ \text{where} \ \ a = \frac{N}{2} \ . \tag{56}$$

For $N = 1, 2, \ldots, 20$ this probability density is plotted in the figure and we can see the central limit theorem at work. Picking the curves at $\chi^2/N = 1$, increasing $f_N(\chi^2)$ values correspond to increasing $N = 1, 2, \ldots, 20$. Assignment a0202_03.



The Error of the Error Bar

For normally distributed data the number of data alone determines the errors of error bars, because the χ^2 distribution is exactly known. One does not have to rely on estimators! Confidence intervals for variance estimates $s_x^2 = 1$ from `NDAT` data (assignment a0204_01):

```
                        q        q        q      1-q      1-q

      NDAT=2**K      .025     .150     .500     .850     .975

          2    1     .199     .483    2.198   27.960 1018.255
          4    2     .321     .564    1.268    3.760   13.902
          8    3     .437     .651    1.103    2.084    4.142
         16    4     .546     .728    1.046    1.579    2.395
         32    5     .643     .792    1.022    1.349    1.768
       1024   10     .919     .956    1.001    1.048    1.093
      16384   14     .979     .989    1.000    1.012    1.022
```

Variance ratio test (F-test)

We assume that two sets of normal data are given together with estimates of their variances $\sigma^2_{x_1}$ and $\sigma^2_{x_2}$: $\left(s^2_{x_1}, N_1\right)$ and $\left(s^2_{x_2}, N_2\right)$. We would like to test whether the ratio $F = \frac{s^2_{x_1}}{s^2_{x_2}}$ differs from $F = 1$ in a statistically significant way. The probability $\frac{f_1}{f_2} F < w$, where $f_i = N_i - 1,\ i = 1, 2$, is know to be

$$H(w) = 1 - B_I \left(\frac{1}{w+1}, \frac{1}{2} f_2, \frac{1}{2} f_1 \right) . \tag{57}$$

Examples (assignment a0505_02):

$\triangle \overline{x}_1$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
N_1	16	16	64	1024	2048	32	1024	16
$\triangle \overline{x}_2$	1.0	1.0	1.0	1.05	1.05	2.0	2.0	2.0
N_2	16	8	16	1024	2048	8	256	16
Q	1.0	0.36	0.005	0.12	0.027	0.90	0.98	0.01

This test allows us later to compare the efficiency of MC algorithms.

The Jackknife Approach

Jackknife estimators allow to correct for the bias and the error of the bias. The method was introduced in the 1950s in papers by Quenouille and Tukey. The jackknife method is **recommended as the standard** for error bar calculations. In unbiased situations the jackknife and the usual error bars agree. Otherwise the jackknife estimates are improvements, so that one cannot loose.

The unbiased estimator of the expectation value $\widehat{x}$ is

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Normally bias problems occur when one estimates a non-linear function of $\widehat{x}$:

$$\widehat{f} = f(\widehat{x}) \ . \tag{58}$$

Typically, the bias is of order $1/N$:

$$\text{bias}\ (\overline{f}) \ = \ \widehat{f} - \langle \overline{f} \rangle \ = \ \frac{a_1}{N} + \frac{a_2}{N^2} + O(\frac{1}{N^3}) \ . \tag{59}$$

Unfortunately, we lost the ability to estimate the variance $\sigma^2(\overline{f}) = \sigma^2(f)/N$ via the standard equation

$$s^2(\overline{f}) \;=\; \frac{1}{N} s^2(f) \;=\; \frac{1}{N\,(N-1)} \sum_{i=1}^{N} (f_i - \overline{f})^2 \;, \tag{60}$$

because $f_i = f(x_i)$ is not a valid estimator of $\widehat{f}$. Also it is in non-trivial applications almost always a bad idea to to use standard error propagation formulas with the aim to deduce $\triangle\overline{f}$ from $\triangle\overline{x}$. Jackknife methods are not only easier to implement, but also more precise and far more **robust**.

The error bar problem for the estimator $\overline{f}$ is conveniently overcome by using **jackknife estimators** $\overline{f}^J$, f_i^J, defined by

$$\overline{f}^J \;=\; \frac{1}{N} \sum_{i=1}^{N} f_i^J \;\; \text{with} \;\; f_i^J \;=\; f(x_i^J) \;\; \text{and} \;\; x_i^J \;=\; \frac{1}{N-1} \sum_{k \neq i} x_k \;. \tag{61}$$

The estimator for the variance $\sigma^2(\overline{f}^J)$ is

$$s_J^2(\overline{f}^J) \;=\; \frac{N-1}{N}\sum_{i=1}^{N}(f_i^J - \overline{f}^J)^2\ . \tag{62}$$

Straightforward algebra shows that in the unbiased case the estimator of the jackknife variance (62) reduces to the normal variance (60).

Notably only of order N (not N^2) operations are needed to construct the jackknife averages $x_i^J,\ i=1,\dots,N$ from the original data.