Lecturenotes Statistics II – Contents

The Central Limit Theorem and Binning

How is the sum of two independent random variables

$$y^r \;=\; x_1^r + x_2^r\ . \tag{1}$$

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

$$g(y) \;=\; \frac{dG(y)}{dy} \;=\; \int_{-\infty}^{+\infty} f_1(x)\ f_2(y-x)\ dx\ . \tag{2}$$

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

$$g_2(y) \;=\; \begin{cases} y & \text{for} \quad 0 \le y \le 1, \\ 2-y & \text{for} \quad 1 \le y \le 2, \\ 0 & \text{elsewhere.} \end{cases} \tag{3}$$

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

$$g_3(y) \;=\; \begin{cases} y^2/2 & \text{for} \;\; 0 \le y \le 1, \\ (-2y^2+6y-3)/2 & \text{for} \;\; 1 \le y \le 2, \\ (y-3)^2/2 & \text{for} \;\; 2 \le y \le 3, \\ 0 & \text{elsewhere.} \end{cases} \tag{4}$$

The convolution (2) 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}:

$$\phi(t) \;=\; \langle e^{itx^r} \rangle \;=\; \int_{-\infty}^{+\infty} e^{itx}\, f(x)\, dx \;. \tag{5}$$

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

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

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

$$y^r \;=\; x_1^r + \ldots + x_N^r \;. \tag{6}$$

The characteristic function of y^r is

$$\phi_y(t) \;=\; \prod_{i=1}^{N} \phi_{x_i}(t) \tag{7}$$

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

$$g(y) \;=\; \frac{1}{2\pi} \int_{-\infty}^{+\infty} dt\, e^{-ity}\, \phi_y(t)\ . \tag{8}$$

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

$$\widehat{g}_N(\overline{x}) \;=\; N\, g_N(N\overline{x})\ . \tag{9}$$

Proof:

Substitute $y = N\overline{x}$ into $g_N(y)\,dy$:

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

Example:

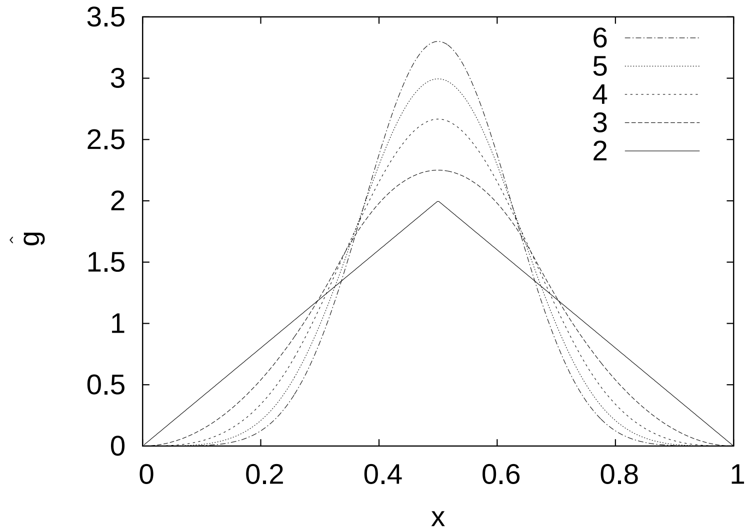


Figure 1: Probability densities for the arithmetic means of two to six uniformly distributed random variables, $\widehat{g}_2(\overline{x})$ and $\widehat{g}_3(\overline{x})$, respectively.

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

$$\phi_y(t) \;=\; \left[\phi_x(t)\right]^N \;. \tag{10}$$

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,

$$\phi_{\overline{x}}(t) \;=\; \phi_y\left(\frac{t}{N}\right) \;=\; \left[\phi_x\left(\frac{t}{N}\right)\right]^N \;. \tag{11}$$

Example: The normal distribution.

The characteristic function is obtained by Gaussian integration

$$\phi(t) \;=\; \exp\left(-\frac{1}{2}\sigma^2 t^2\right) \;. \tag{12}$$

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

$$\phi_y(t) \;=\; [\phi(t)]^2 \;=\; \exp\left(-\frac{1}{2} 2\sigma^2 t^2\right) \;. \tag{13}$$

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 $t \to t/2$:

$$\phi_{\overline{x}}(t) \;=\; \exp\left(-\frac{1}{2}\frac{\sigma^2}{2} t^2\right) \;. \tag{14}$$

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

$$\phi_x(t) \;=\; 1 \;-\; \frac{\sigma_x^2}{2}\, t^2 \;+\; \mathcal{O}(t^3)\;. \tag{15}$$

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

$$\sigma^2(\overline{x}^{\,r}) \;=\; \frac{\sigma^2(x^r)}{N}\;. \tag{16}$$

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

$$\phi(t) = \int_{-\infty}^{+\infty} dx\,\frac{e^{itx}}{\pi\,(1+x^2)} \;=\; \exp(-|t|)\;. \tag{17}$$

The integration involves the residue theorem. Using equation (11) for the characteristic function of the mean of N random variables, we find

$$\phi_{\overline{x}}(t) = \left[\exp\left(-\frac{|t|}{N}\right)\right]^n = \exp(-|t|)\ . \tag{18}$$

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.

Generalized limit theorems exist for broad distributions, like those given by a probability density

$$f(x) = \frac{a}{2(1+|x|)^{1+a}},\ 0 < a \leq 2\,.$$

This goes sometimes under the name Levy statistics, but is in the Russian literature more closely associated with work by Khintchine.

Binning

The notion of introduced here should not be confused with histograming! 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 be neglected, once NBIN is sufficiently large compared to the autocorrelation time.

Example:

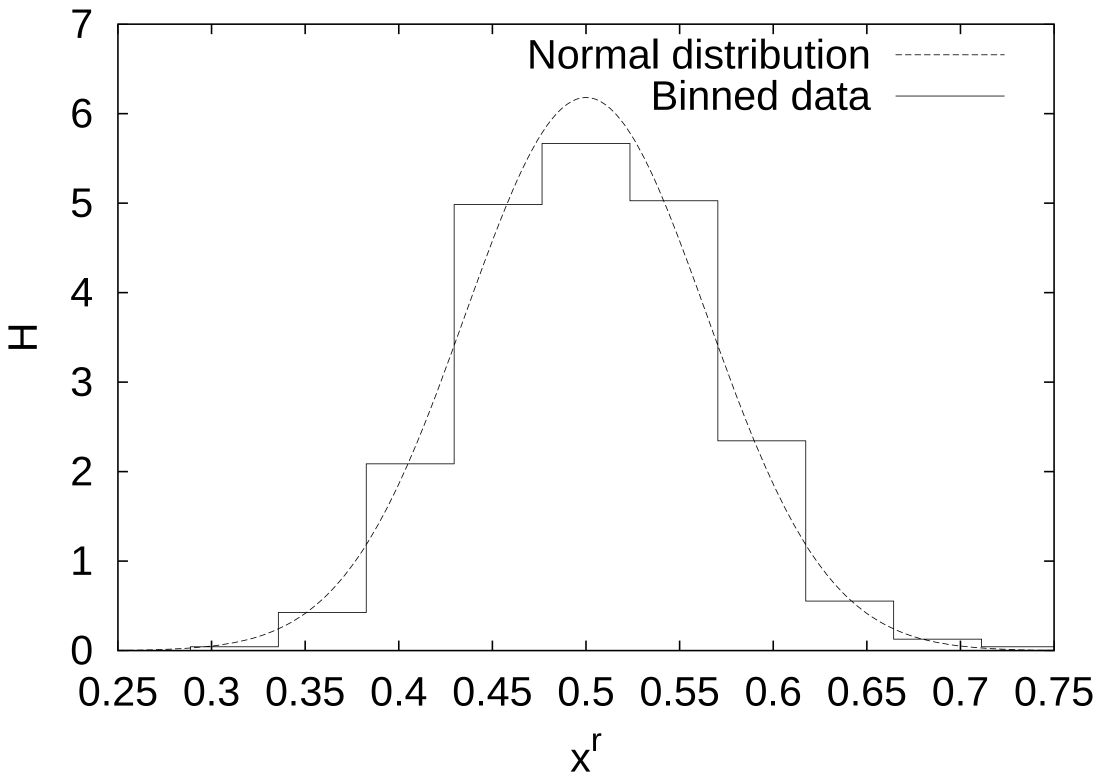


Figure 2: Comparison of a histogram of 500 binned data with the normal distribution $\sqrt{(120/\pi)}\,\exp[-120\,(x-1/2)^2]$. Each binned data point is the average of 20 uniformly distributed random numbers. Assignment a0108_02.

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

$$\sigma = \sigma(\overline{x}^{\,r}) \quad \text{with} \quad \overline{x}^{\,r} = \frac{1}{N}\sum_{i=1}^{N} x_i^r \tag{19}$$

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 the probability content p of the confidence intervals (19) to be

$$p = p(n) = G(n\sigma) - G(-n\sigma) = \frac{1}{\sqrt{2\pi}}\int_{-n}^{+n} dx\, e^{-\frac{1}{2}x^2} = \text{erf}\left(\frac{n}{\sqrt{2}}\right) \, . \tag{20}$$

Table 1: 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

$$\overline{x} \in [\widehat{x} - n\sigma, \widehat{x} + n\sigma] \iff \widehat{x} \in [\overline{x} - n\sigma, \overline{x} + n\sigma] \tag{21}$$

solves the problem. Conventionally, these estimates are quoted as

$$\widehat{x} = \overline{x} \pm \triangle\overline{x} \tag{22}$$

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

$$(s_x'^{\,r})^2 \;=\; \frac{1}{N}\sum_{i=1}^{N}(x_i^r - \overline{x}^{\,r})^2 \tag{23}$$

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

$$\langle (s_x'^{\,r})^2\rangle \;\neq\; \sigma_x^2\ . \tag{24}$$

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 (23) 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

$$(s_x^r)^2 \;=\; \frac{N}{N-1}\,(s_x'^{\,r})^2 \;=\; \frac{1}{N-1}\sum_{i=1}^{N}(x_i^r - \overline{x}^{\,r})^2 \; . \tag{25}$$

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

$$(s_{\overline{x}}^r)^2 \;=\; \frac{1}{N(N-1)}\sum_{i=1}^{N}(x_i^r - \overline{x}^{\,r})^2 \; . \tag{26}$$

Our workhorse for error bar calculations from data is the routines `steb0.f` of `ForLib` used via

```
CALL STEB0(NDAT, DAT, DATM, DATV, DATME) .
```

See today's classwork.

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

$$d^r \;=\; \frac{D^r}{\sigma_D} \tag{27}$$

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

$$P \;=\; P(|d^r| \le d) \;=\; G_0(d) - G_0(-d) \;=\; 1 - 2\,G_0(-d) \;=\; \mathrm{erf}\left(\frac{d}{\sqrt{2}}\right)\;. \tag{28}$$

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

$$Q \;=\; 1 - P \;=\; 2\,G_0(-d) \;=\; 1 - \mathrm{erf}\left(\frac{d}{\sqrt{2}}\right)\;. \tag{29}$$

If the data set are sampled from identical Gaussian distributions, Q is a uniformly distributed random variable in the range $[0, 1)$. Example: See today's classwork.

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

$$t^r = \frac{\overline{x}^{\,r} - \widehat{x}}{s^r_{\overline{x}}} \tag{30}$$

is given by the probability density

$$f(t) = \frac{1}{(N-1)\,B(1/2,(N-1)/2)} \left(1 + \frac{t^2}{N-1}\right)^{-\frac{N}{2}} . \tag{31}$$

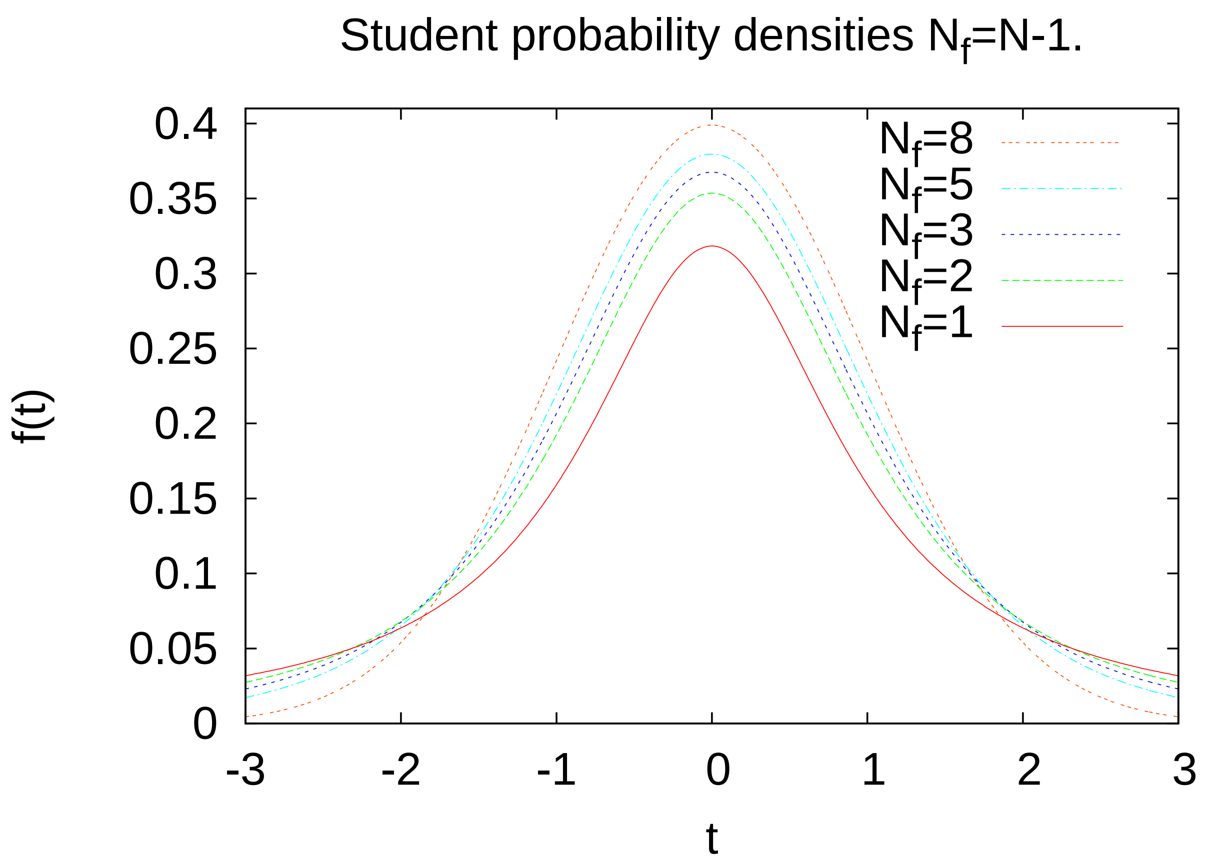


Figure 3: Probability densities of some student distributions.

Here $B(x,y)$ is the beta function. The fall-off is a power law $|t|^{-f}$ for $|t| \to \infty$. 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
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.

Student Difference Test

This test takes into account that only a finite number of events are sampled. Let the following (normal distributed) data be given

$$\overline{x} \;\text{ calculated from }\; M \;\text{ events, }\; i.e., \;\; \sigma^2_{\overline{x}} \;=\; \sigma^2_x/M \tag{32}$$

$$\overline{y} \;\text{ calculated from }\; N \;\text{ events, }\; i.e., \;\; \sigma^2_{\overline{y}} \;=\; \sigma^2_y/N \tag{33}$$

and an unbiased estimators of the variances

$$s^2_{\overline{x}} = s^2_x/M = \frac{\sum_{i=1}^{M}(x_i - \overline{x})^2}{M\,(M-1)} \;\text{ and }\; s^2_{\overline{y}} = s^2_y/N = \frac{\sum_{j=1}^{N}(y_j - \overline{y})^2}{N\,(N-1)}\;. \tag{34}$$

Under the additional assumption $\sigma^2_x = \sigma^2_y$ the probability

$$P(|\overline{x} - \overline{y}| > d) \tag{35}$$

is is determined by the Student distribution function. Example: Todays's classwork.

More examples: 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 question: Assume data are sampled from the same normal distribution, when are two measured error bars consistent and when not? Answered later by: χ^2 Distribution, Error of the Error Bar, and Variance Ratio Test.

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

For normally distributed data with mean zero the random variable

$$(\chi^r)^2 = \sum_{i=1}^{f} (y_i^r)^2 \ , \tag{36}$$

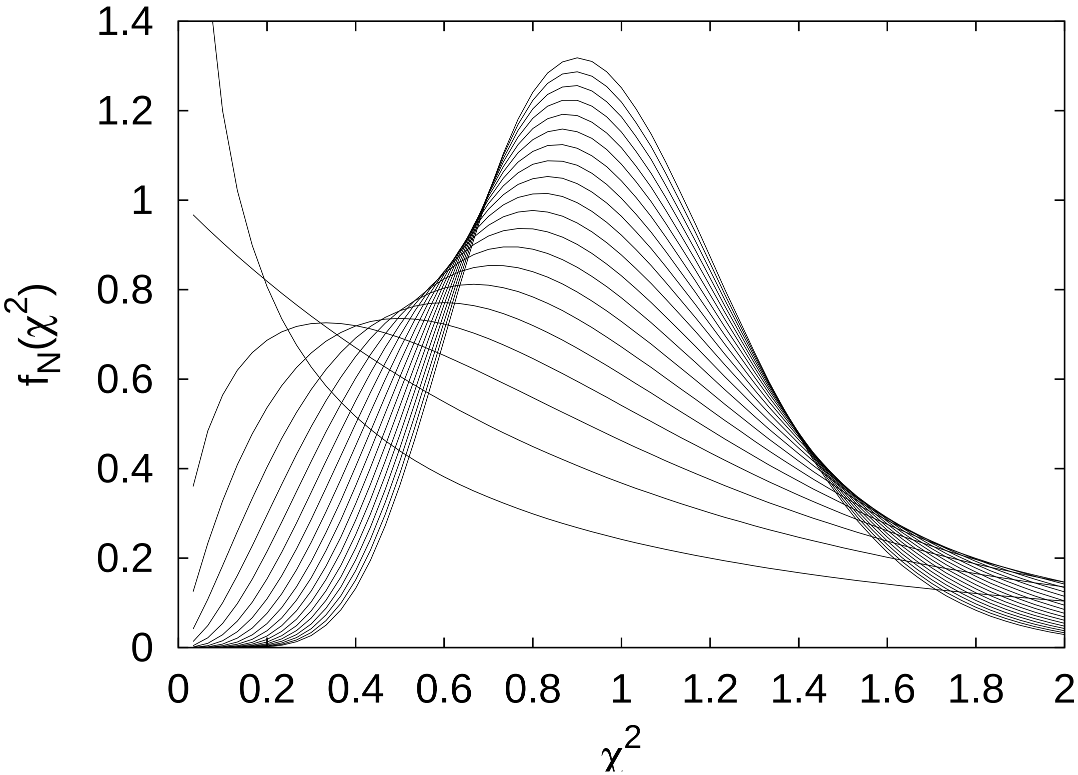
defines the χ^2 **distribution** of f degrees of freedom. The study of the variance $(s_x^r)^2$ of a Gaussian sample can be reduced to the χ^2-distribution with $f = N - 1$

$$(\chi_f^r)^2 = \frac{(N-1)\,(s_x^r)^2}{\sigma_x^2} = \sum_{i=1}^{N} \frac{(x_i^r - \overline{x}^{\,r})^2}{\sigma_x^2} \ . \tag{37}$$

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

$$f_N(\chi^2) \ = \ N f(N\chi^2) \ = \frac{a\ e^{-a\chi^2}\ \left(a\chi^2\right)^{a-1}}{\Gamma(a)} \ \text{ where } \ a = \frac{N}{2} \ . \tag{38}$$

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 $F = 1$. With $f_i = N_i - 1,\ i = 1, 2$

$$\text{The probability } \frac{f_1}{f_2} F < w \text{ is } H(w) = 1 - B_I\left(\frac{1}{w+1}, \frac{1}{2} f_2, \frac{1}{2} f_1\right) \, .$$

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

Examples: Today's classwork.

Statistical Bootstrap for Mean Values

This resampling approach is best done in three steps:

1. Block (bin) the data into `NDAT` blocked data.
2. Sample without replacement from the blocked data a sample of `NBTR` mean estimates, the bootstrap sample. See subroutine `bootstrp.f`.
3. Sort the bootstrp sample and estimate confidence limits from the Cumulative Distribution Function (CDF). See subroutine `bootstrp.f`.

Example: See classwork (pi estimate again).

Typically, the method is useful when one has relatively few independent data and/or the distribution of the estimated mean values is not Gaussian.