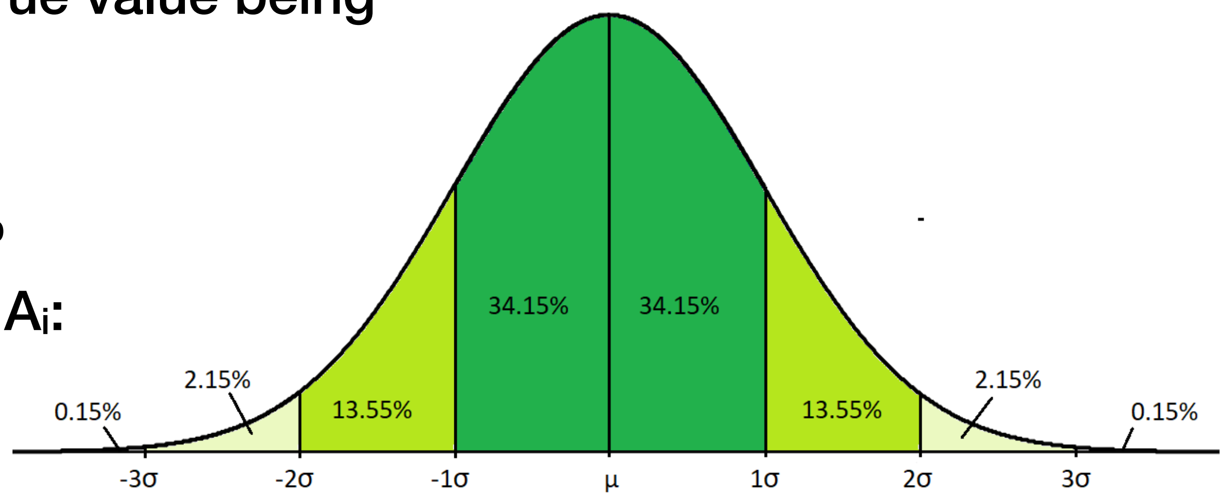
Statistical errors

Expressing a statistical estimate as A ± σ, the meaning normally is

- σ represents one standard deviation of the computed mean value A
- under the assumption of normal-distributed fluctuations

Then, the probability of the true value being

- within [A-σ,A+σ] is 68%
- within [A-2σ,A+2σ] is 95%
- within [A-3σ,A+3σ] is 99.7%



For M independent samples A_i:

$$\bar{A} = \frac{1}{M}\sum_{i=1}^{M} A_i$$

$$\sigma_A = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(A_i - \bar{A})^2} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(A_i^2 - \bar{A}^2)}$$

$$= \sqrt{\overline{A^2} - (\bar{A})^2}$$

The estimated standard deviation of the distribution of values {A_i}

But the "error bar" is the standard deviation of the mean of {A_i}

The mean value fluctuates less than the width σ_A of the distribution
- imagine taking the number of samples M to infinity:

$$\sigma_A = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(A_i - \bar{A})^2}$$

will approach a constant value
- the standard deviation of the distribution

$$\bar{A} = \frac{1}{M}\sum_{i=1}^{M} A_i$$

will approach a constant value
- the actual value <A> of A

$\longrightarrow$ σ_A cannot be the proper statistical error of A

Variances add: variance of the sum $\sum_{i=1}^{M} A_i$ is $M\sigma_A^2$

- standard deviation of the sum is $\sqrt{M}\sigma_A$

- divide by M; standard deviation of the mean is $\sigma_A/\sqrt{M}$

- here M should be replaced by M-1 (reflecting infinite uncertainty if M=1)

$$\sigma = \sqrt{\frac{1}{M(M-1)}\sum_{i=1}^{M}(A_i - \bar{A})^2} = \sqrt{\frac{1}{M(M-1)}\sum_{i=1}^{M}(A_i^2 - \bar{A}^2)} = \sqrt{\frac{\overline{A^2} - (\bar{A})^2}{M-1}}$$

Data binning

The statistical error ("error bar") has its conventional meaning only if the values $\{A_i\}$ are normal distributed

- typically they obey some completely different distribution

Apply **central limit theorem** to obtain normal distributed "bin averages"

A bin average is based on M samples as before, but now B of them

- B different mean values (estimates of A): $\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_B$

$$\bar{A}_b = \frac{1}{M}\sum_{i=1}^{M} A_{b,i}$$

$A_{b,i}$ is value #i belonging to bin b

Regardless of the distribution of individual values

- if M is large enough, the bin averages are normal-distributed

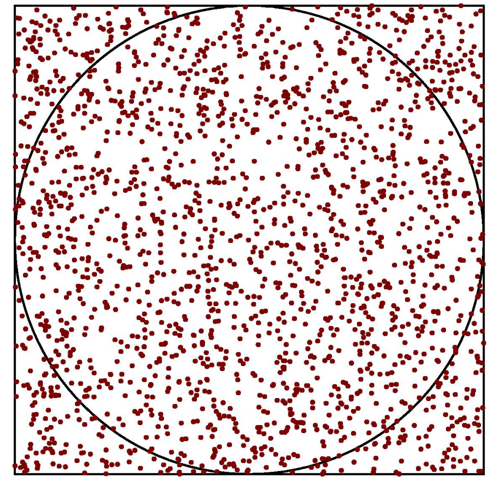
Use standard formulas with the bin data:

$$\bar{A} = \frac{1}{B}\sum_{b=1}^{B} \bar{A}_b \qquad \sigma = \sqrt{\frac{1}{B(B-1)}\sum_{b=1}^{B}(\bar{A}_b - \bar{A})^2} = \sqrt{\frac{1}{B(B-1)}\sum_{b=1}^{B}(\bar{A}_b^2 - \bar{A}^2)} = \sqrt{\frac{\overline{A^2} - (\bar{A})^2}{B-1}}$$

Emergence of normal distribution

- example: sampling f=1circle in square
- lets just consider the estimate of the mean <f>

$$f(x,y) = \begin{cases} 1, & \text{if } x^2+y^2 \le 1 \\ 0, & \text{if } x^2+y^2 < 1 \end{cases}$$



For each sample, the probabilities of f=0,1 are:

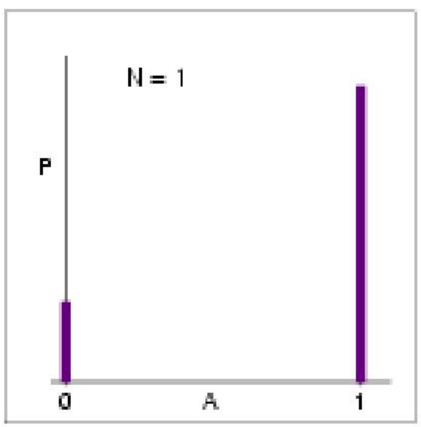
$$P(f=1) = \pi/4, \quad P(f=0) = 1-\pi/4$$

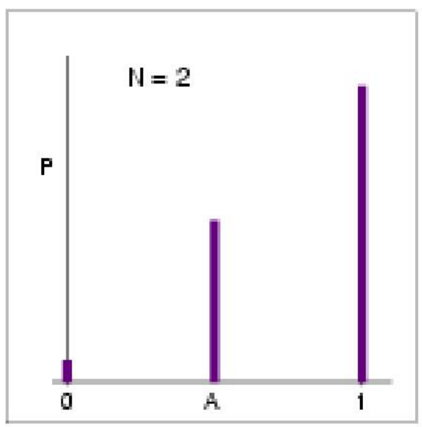
For N samples, the possible average values A are

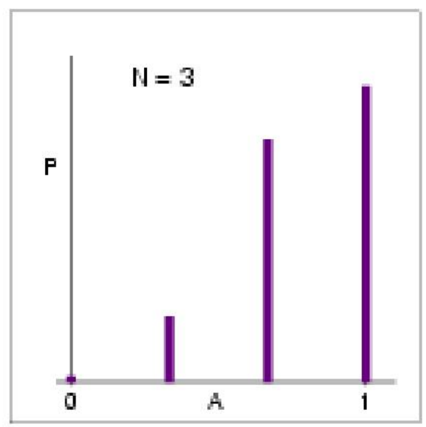
$$A \in \left\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\right\}$$

the probabilities of these averages are

$$P\left(A = \frac{m}{N}\right) = \frac{N!}{m!(N-m)!}\left(\frac{\pi}{4}\right)^m \left(1-\frac{\pi}{4}\right)^{N-m}$$





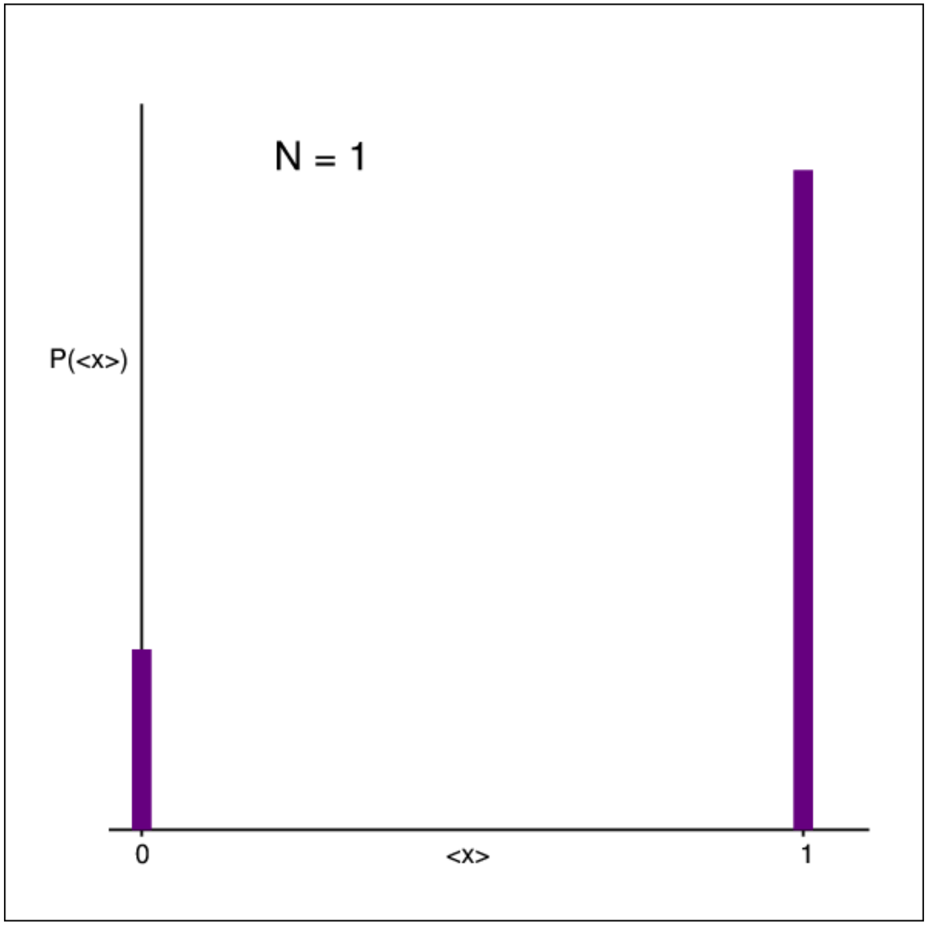


Evolution of P(A)
from N=1 to 100

Note: We can think of the probability distribution of a continuum of A values

P(A) is a sum of delta-functions; reflects discrete set of possible values

For large N, a small broadening of the deltas (e.g., bars or Gaussians) give a continuous distribution



$$P(A) = \sum_{m=0}^{N} \frac{N!}{m!(N-m)!} \left(\frac{\pi}{4}\right)^m \left(1 - \frac{\pi}{4}\right)^{N-m} \delta(A - m/N)$$

Modified circle integration

Function with singularity. Inside circle of radius 1:

$$f(r) = r^{-\alpha}, \quad r = \sqrt{x^2 + y^2}$$

integrable if $\alpha < 2$

$$I = \int_{-1}^{1} dy \int_{-1}^{1} dx f(x,y), \qquad f(x,y) = r^{-\alpha}, \text{ if } r \leq 1, \qquad f(x,y) = 0, \text{ if } r > 1$$

Distribution of radius r inside circle: ***P(r)=2r*** ***(0 ≤ r ≤ 1)***

$$\int_0^1 P(r)dr = 2\int_0^1 r dr = 1$$

Distribution of function values inside the circle:

$$P(f)df = P(r)\left|\frac{dr}{df}\right| df = \frac{2}{\alpha} f^{-1-2/\alpha} df$$

outside:

$$P(f) = (1 - \pi/4)\delta(f)$$

$$P(f) = \frac{\pi}{4}\frac{2}{\alpha} f^{-1-2/\alpha}\Theta(f-1) + \left(\frac{\pi}{4} - 1\right)\delta(f)$$

Distribution of average A of f based on N samples:

$$P(A) = \int_0^\infty df_N \cdots \int_0^\infty df_1 P(f_N)\cdots P(f_1)\delta[A - (f_1 + \cdots + f_N)/N]$$

$$P(A) = \int_0^\infty df_N \cdots \int_0^\infty df_1 P(f_N) \cdots P(f_1)\delta[A - (f_1 + \cdots + f_N)/N]$$

Should become normal distribution for large N

What is large enough (e.g., to use for data binning)?

$\alpha = 3/2$

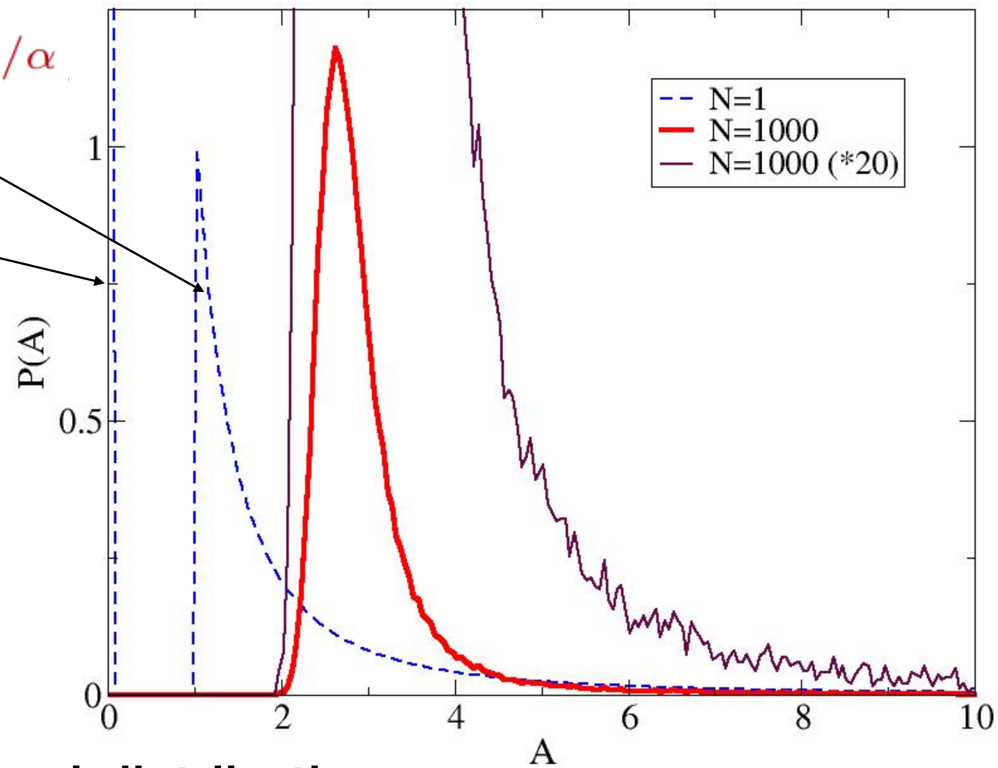
How can we compute the probability distribution?

$\frac{2}{\alpha} f^{-1-2/\alpha}$

Monte Carlo sampling

- we can get P(A)
- not just <A>

$\delta(f)$



There is a delta fktn at A=0 in the N=1000 result, with a very small amplitude $(1-\pi/4)^{1000}$

For N=1000, there is still a “fat tail”

- larger N needed to approximte normal distribution

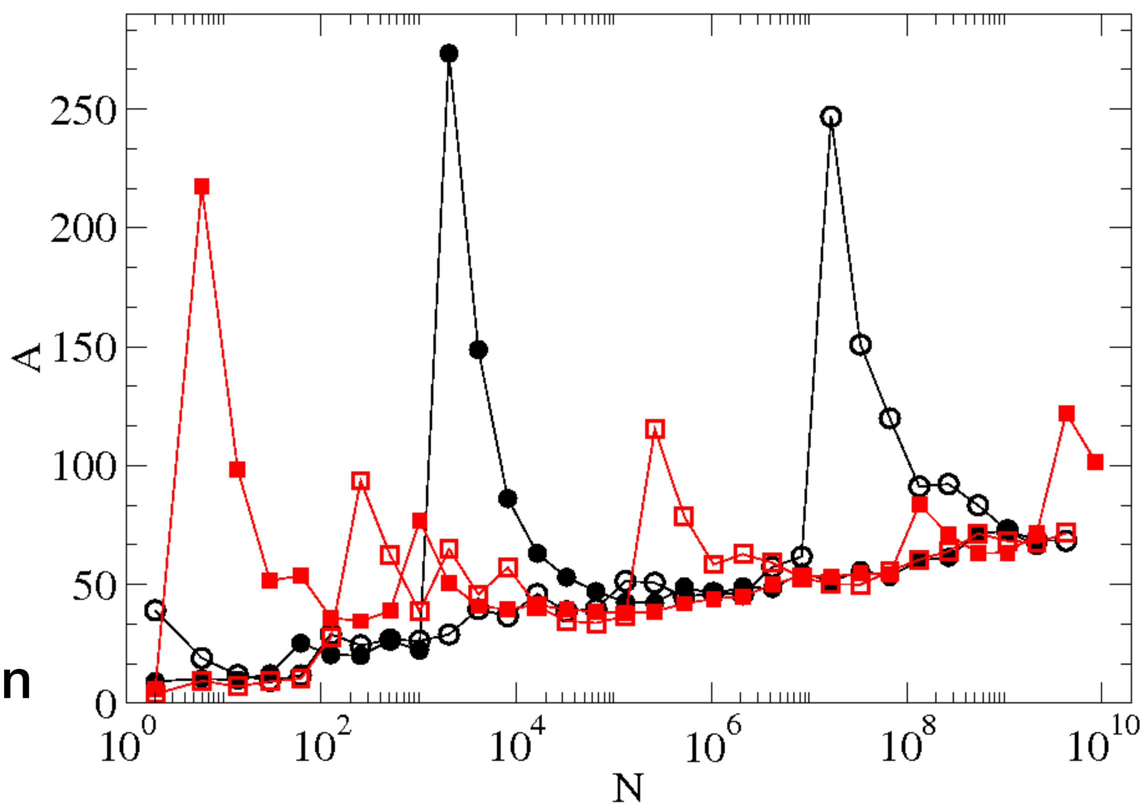
What happens if the function is not integrable?

Example: boarderline case: $f(r) = r^{-\alpha}, \alpha = 2$

- singularity at r=0, log divergence vs lower cut-off r_0

$$\int_0^{2\pi} d\phi \int_{r_0}^{1} \frac{r\,dr}{r^2} = -2\pi \ln(r_0) = 2\pi \ln(1/r_0)$$

Four independent simulations
- partial averages based on N samples



How is the divergence manifested in MC sampling?

Rare-event behavior
- due to fat tails in P(A)

Occasional very large f values give huge contributions to A, cause spikes in A(N)

The overall behavior of <A(N)>, i.e., the peak of the of distribution of P(A(N)), shows a log behavior

Numerical integration on a mesh vs MC sampling

Scaling of the computational effort:

- may depend on the dimensionality and the required precision ε

Mesh-based method: time ~ $M(\varepsilon)^D \times g(\varepsilon)$

- where $g(\varepsilon)$ depends on integrand and method

Monte Carlo sampling method: $\varepsilon \sim N^{-1/2}$, time ~ $\varepsilon^{-2} \times h(f)$

- where $h(f)$ depends on the function f
- time scaling not explicitly dependent on the dimensionality D

Which type of method is better?

- for given desired precision ε

The above scaling forms show that MC sampling should be better above some dimensionality D

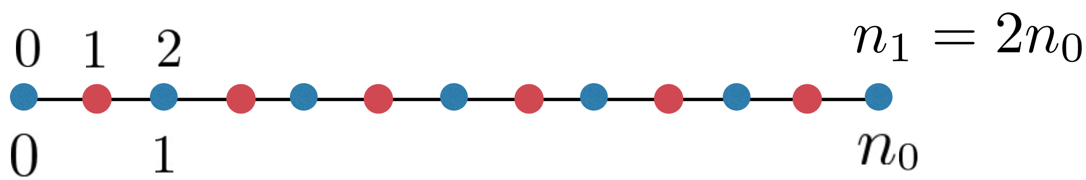
- in practice, mesh-based methods are difficult even for D=3
- MC sampling can work well even in very high dimensions
  - unless the integrand is strongly varying (low probability of hitting contributing parts of the volume

Romberg integration

Idea: Use two or more trapezoidal integral estimates, extrapolate

- step sizes (decreasing order) $h_0, h_1, \ldots, h_m$, integral estimates $I_0, I_1, \ldots, I_m$
- use polynomial of order n to fit and extrapolate to h=0
- error for given h scales as h^2 (+ higher even powers only)
- use polynomial P(x) with $x=h^2$

Simplest case: 2 points (m=1), using $h_0=(b-a)/n_0$ and $h_1=h_0/2$ ($x_1=x_0/4$)



Function evaluation once only for each point needed

$$I_0 = I_\infty + \epsilon x_0, \quad I_1 = I_\infty + \epsilon x_0/4$$

$$\rightarrow \quad I_\infty = \frac{4}{3}I_1 - \frac{1}{3}I_0 \quad + O(h_0^4) \quad [O(x_0^2)]$$

reducing h by 50%

- error should be 1/4 of previous
- ε is unknown factor, eliminated

Computation cost doubled, error reduced by two powers of h_0!

Generalizes easily to the case of m estimates