Statistical Methods & Tools

Wouter Verkerke
(NIKHEF)

Roadmap for this course

- In this course I aim to follow the 'HEP workflow' to organize the discussion of statistics issues
  - My experience is that this most intuitive for HEP Phd students
- Basics (15 slides)
  - Distributions, the Central Limit Theorem
- Event classification (54 slides)
  - Hypothesis testing
  - Machine learning
- Parameter estimation (64 slides)
  - Estimators: Maximum Likelihood and Chi-squared
  - Mathematical tools for model building
  - Practical issues arising with minimization
- Confidence intervals, limits, significance (54 slides)
  - Hypothesis testing (again), Bayes Theorem
  - Frequentist statistics, Likelihood-based intervals
- Likelihood principle, Systematics and nuisance parameters (53 slides)
  - Likelihood principle and conditioning
  - Systematic uncertainties as nuisance parameters
  - Treatment of nuisance parameters in statistical inference

Basics

— Basic distributions – Binomial, Poisson, Gaussian
— Central Limit Theorem
— Covariance, correlations

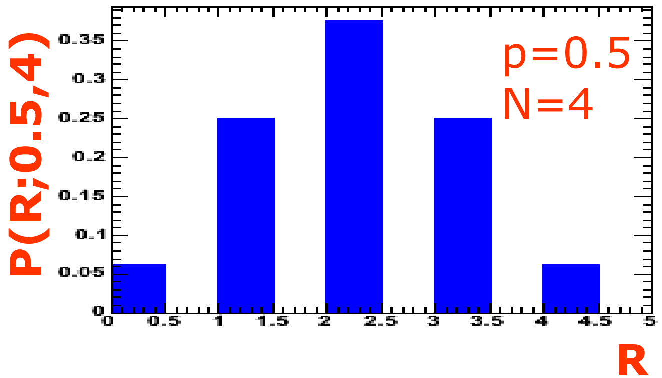
Basic Distributions – The binomial distribution

- Simple experiment – Drawing marbles from a bowl
  - Bowl with marbles, fraction **p** are black, others are white
  - Draw **N** marbles from bowl, *put marble back after each drawing*
  - Distribution of **R** black marbles in drawn sample:

Probability of a specific outcome e.g. 'BBBWBWW'

Number of equivalent permutations for that outcome

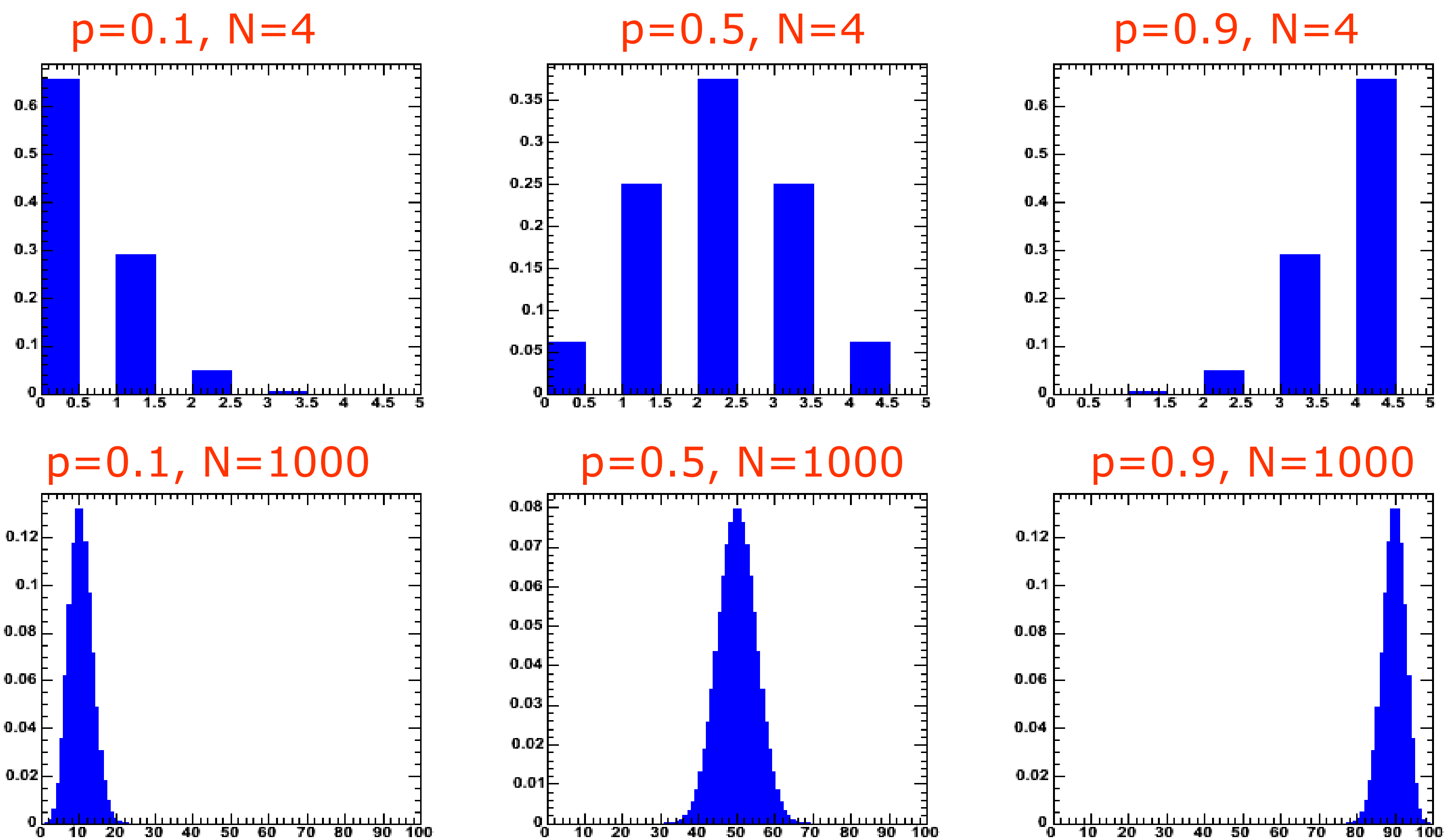
$$P(R;p,N) = p^R (1-p)^{N-R} \frac{N!}{R!(N-R)!}$$



Binomial distribution

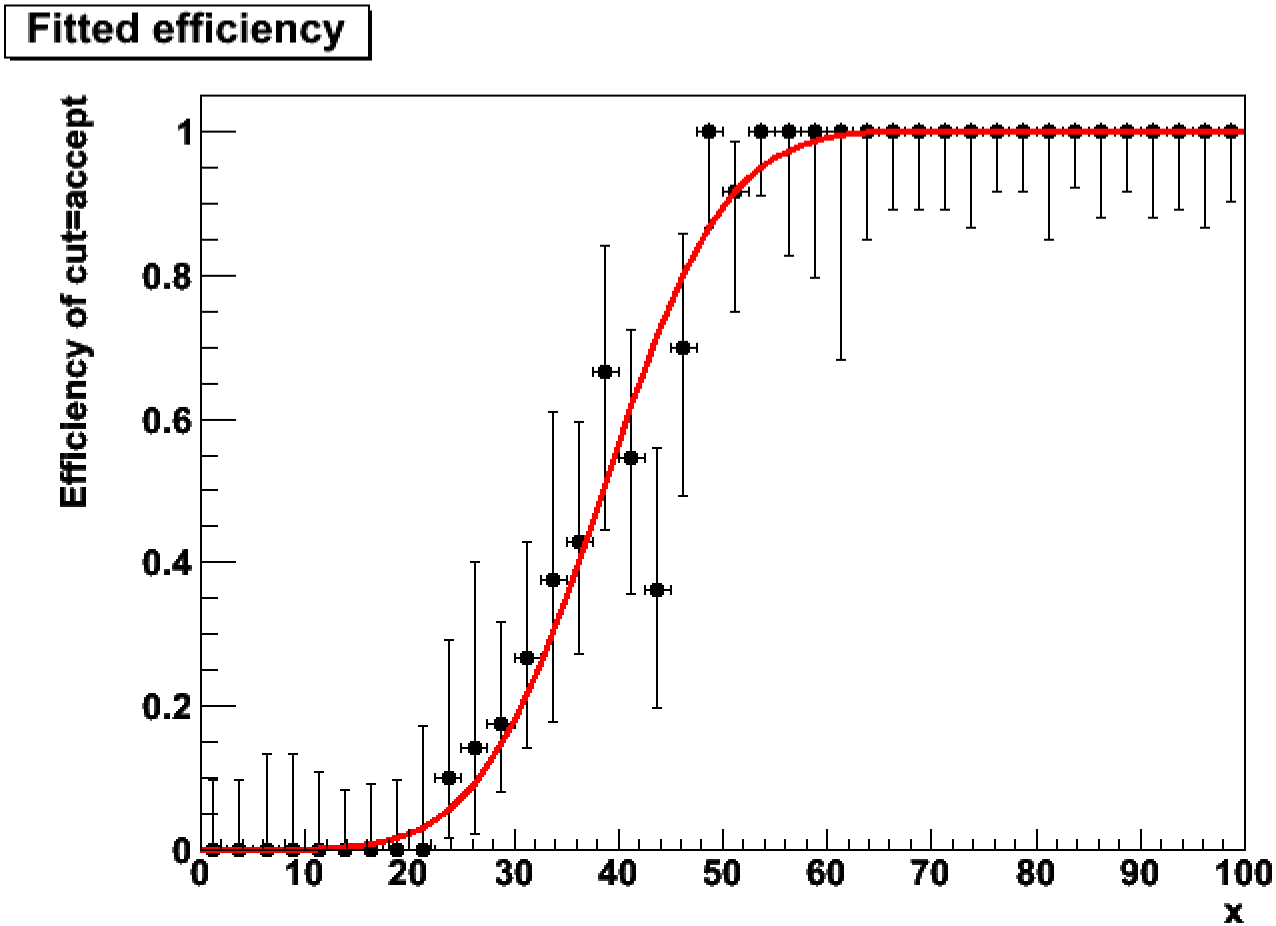
Properties of the binomial distribution

- Mean: $$\langle r \rangle = n \cdot p$$
- Variance: $$V(r) = np(1-p) \quad \Rightarrow \quad \sigma = \sqrt{np(1-p)}$$



HEP example – Efficiency measurement

- Example: trigger efficiency turn-on curve



Basic Distributions – the Poisson distribution

- Sometimes we don't know the equivalent of the number of drawings
  - **Example: Geiger counter**
  - Sharp events occurring in a (time) continuum
- What distribution to we expect in measurement over fixed amount of time?
  - Divide time interval λ in n finite chunks,
  - Take binomial formula with $p=\lambda/n$ and let $n \to \infty$

$$P(r;\lambda/n,n) = \frac{\lambda^r}{n^r}\left(1-\frac{\lambda}{n}\right)^{n-r}\frac{n!}{r!(n-r)!}$$

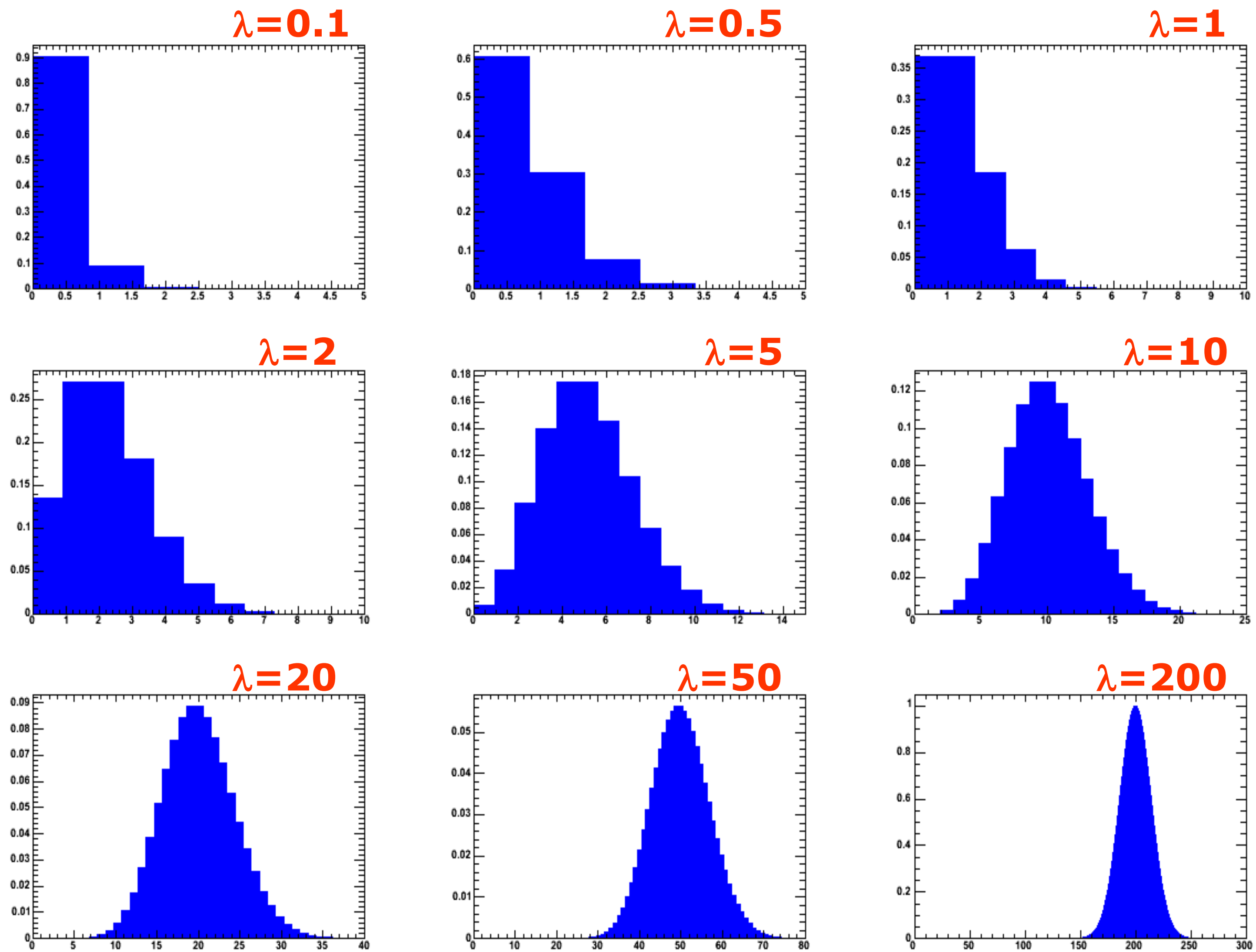
$$\lim_{n\to\infty}\frac{n!}{r!(n-r)!} = n^r,$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-r} = e^{-\lambda}$$

$$P(r;\lambda) = \frac{e^{-\lambda}\lambda^r}{r!}$$

←Poisson distribution

Properties of the Poisson distribution



More properties of the Poisson distribution $P(r;\lambda)=\frac{e^{-\lambda}\lambda^r}{r!}$

- Mean, variance:

$$\langle r \rangle = \lambda$$

$$V(r)=\lambda \quad \Rightarrow \quad \sigma=\sqrt{\lambda}$$

- Convolution of 2 Poisson distributions is also a Poisson distribution with $\lambda_{ab}=\lambda_a+\lambda_b$

$$P(r)=\sum_{r_A=0}^{r} P(r_A;\lambda_A)P(r-r_A;\lambda_B)$$

$$= e^{-\lambda_A}e^{-\lambda_B}\sum \frac{\lambda_A^{r_A}\lambda_B^{r-r_A}}{r_A!(r-r_A)!}$$

$$= e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}\sum_{r_{A=0}}^{r}\frac{r!}{(r-r_A)!}\left(\frac{\lambda_A}{\lambda_A+\lambda_B}\right)^{r_A}\left(\frac{\lambda_B}{\lambda_A+\lambda_B}\right)^{r-r_A}$$

$$= e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}\left(\frac{\lambda_A}{\lambda_A+\lambda_B}+\frac{\lambda_B}{\lambda_A+\lambda_B}\right)^r$$

$$= e^{-(\lambda_A+\lambda_B)}\frac{(\lambda_A+\lambda_B)^r}{r!}$$

Basic Distributions – The Gaussian distribution

- Look at *Poisson distribution* in limit of *large N*

$$P(r;\lambda) = e^{-\lambda}\frac{\lambda^r}{r!}$$

Take log, substitute, r = λ + x,
and use $\ln(r!) \approx r\ln r - r + \ln\sqrt{2\pi r}$

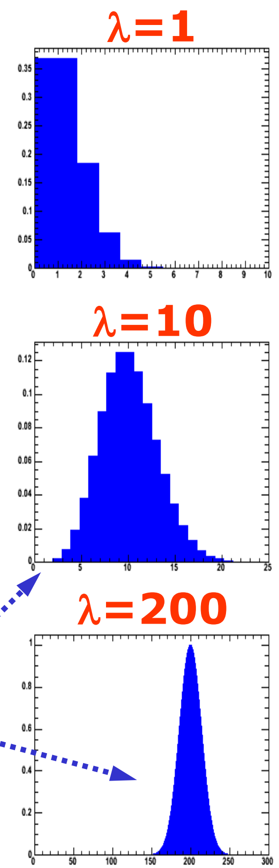
$$\begin{aligned}\ln(P(r;\lambda)) &= -\lambda + r\ln\lambda - (r\ln r - r) - \ln\sqrt{2\pi r}\\ &= -\lambda + r\left[\ln\lambda - \ln(\lambda(1+\frac{x}{\lambda}))\right] + (\lambda + x) - \ln\sqrt{2\pi\lambda}\\ &\approx x - (\lambda - x)\left(\frac{x}{\lambda} + \frac{x^2}{2\lambda^2}\right) - \ln(2\pi\lambda)\\ &\approx \frac{-x^2}{2\lambda} - \ln(2\pi\lambda)\end{aligned}$$

$\ln(1+z) \approx z - z^2/2$

Take exp

$$P(x) = \frac{e^{-x^2/2\lambda}}{\sqrt{2\pi\lambda}}$$

Familiar Gaussian distribution,
(approximation reasonable for N>10)



Properties of the Gaussian distribution

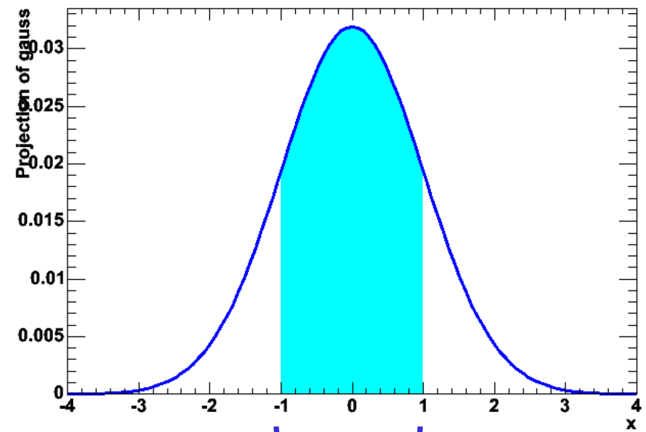
$$P(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}e^{-(x-\mu)^2/2\sigma^2}$$

- *Mean* and *Variance*

$$\langle x\rangle=\int_{-\infty}^{+\infty}xP(x;\mu,\sigma)dx=\mu$$

$$V(x)=\int_{-\infty}^{+\infty}(x-\mu)^2P(x;\mu,\sigma)dx=\sigma^2$$

$$\sigma=\sigma$$



- Integrals of Gaussian

68.27% within 1σ	90% → 1.645σ
95.43% within 2σ	95% → 1.96σ
99.73% within 3σ	99% → 2.58σ
	99.9% → 3.29σ

The Gaussian as 'Normal distribution'

- Why are errors usually Gaussian?

- The ***Central Limit Theorem*** says
  - If you take the sum X of N independent measurements x_i, each taken from a distribution of mean m_i, a variance $V_i=\sigma_i^2$, the distribution for x

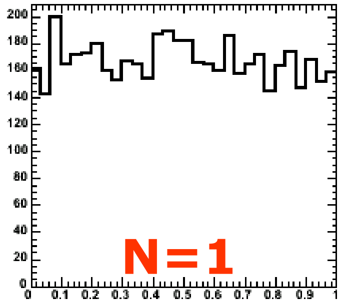
    (a) has expectation value $$\langle X \rangle = \sum_i \mu_i$$

    (b) has variance $$V(X) = \sum_i V_i = \sum_i \sigma_i^2$$

    (c) becomes Gaussian as N → ∞

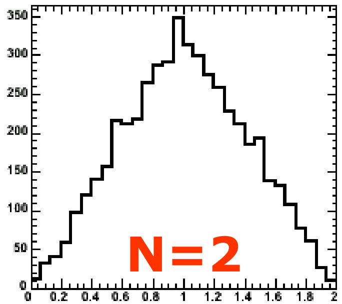
  - *Small print: tails converge very slowly in CLT, be careful in assuming Gaussian shape beyond 2σ*

Demonstration of Central Limit Theorem



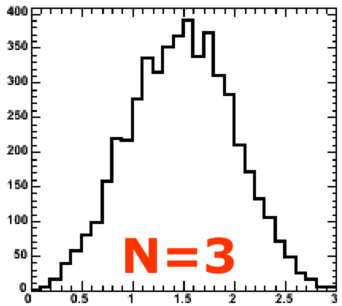
← 5000 numbers taken at random from a uniform distribution between [0,1].

- Mean = $^{1}/_{2}$, Variance = $^{1}/_{12}$

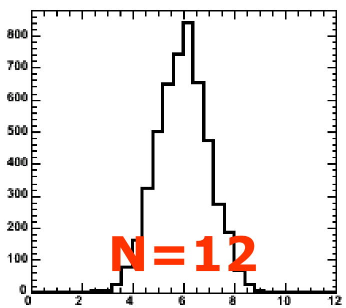


← 5000 numbers, each the sum of 2 random numbers, i.e. $X = x_1+x_2$.

- Triangular shape



← Same for 3 numbers, $X = x_1 + x_2 + x_3$



← Same for 12 numbers, overlaid curve is exact Gaussian distribution

Central Limit Theorem – repeated measurements

- Common case 1 : Repeated identical measurements
  i.e. $\mu_i = \mu$, $\sigma_i = \sigma$ for all i

C.L.T

$$\langle X \rangle = \sum_i \mu_i = N\mu \quad \Rightarrow \quad \langle \bar{x} \rangle = \frac{X}{N} = \mu$$

$$V(\bar{x}) = \sum_i V_i(\bar{x}) = \frac{1}{N^2} \sum_i V_i(X) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{N}}$$

← Famous sqrt(N) law

Central Limit Theorem – repeated measurements

- Common case 2 : Repeated measurements with identical means but different errors (i.e weighted measurements, $\mu_i = \mu$)

$$\bar{x} = \frac{\sum x_i / \sigma_i^2}{\sum 1/\sigma_i^2}$$

Weighted average

$$V(\bar{x}) = \frac{1}{\sum 1/\sigma_i^2} \Rightarrow \sigma(\bar{x}) = \frac{1}{\sqrt{\sum 1/\sigma_i^2}}$$

'Sum-of-weights' formula for error on weighted measurements