Basics of Data Analysis

Measurement uncertainties / Error analysis

Slides developed and adapted by

Kirill Bolotin, Martin Weinelt, and Niclas Müller

Literature

- P. R. Bevington, and D. K. Robinson, “*Data Reduction and Error Analysis for the Physical Sciences*”
- P. Möhrke, and B.-U. Runge, “*Arbeiten mit Messdaten*”
- “*Guide to uncertainty propagation and Error analysis*” Stony Brook U.
- “*A beginner guide to uncertainty of measurement*” by Stephanie Bell
- Slides and python code examples will be uploaded to website

Basics of Data Analysis

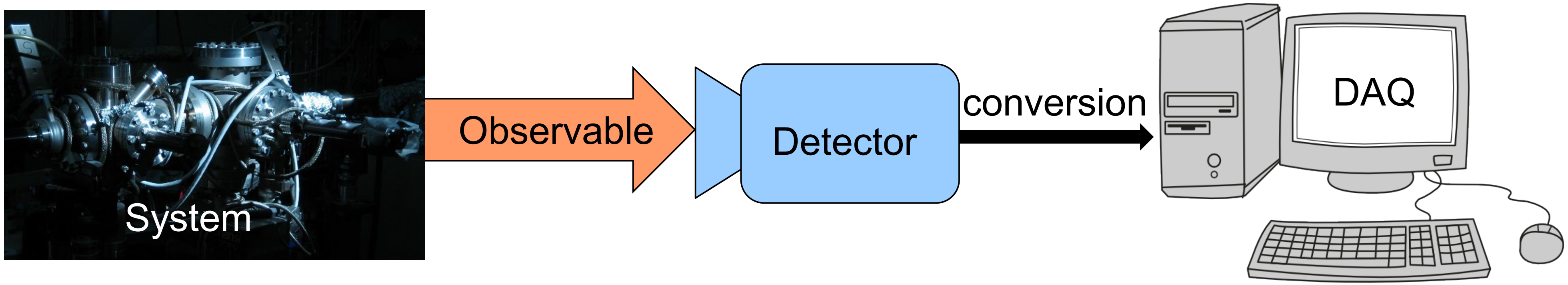
Measurement uncertainties / Error analysis

based on

- Slides by Kirill Bolotin and Martin Weinelt
- P. R. Bevington, and D. K. Robinson, "*Data Reduction and Error Analysis for the Physical Sciences*"
- P. Möhrke, and B.-U. Runge, "*Arbeiten mit Messdaten*"
- Check out also: T. Kampfrath, Grundlagen der Mess- und Labortechnik, Fri 2-4pm, Large Lect Hall

Always report measurement results with uncertainties!

Errors and uncertainties in experimental data



- Environmental
- Instrumental
- Observational
- Theoretical

Wind, magnetic field, temperature

Instrument accuracy limit
Fluctuating signal

Misused instrument

Error in equation
Effect not accounted for

Outline

- Types of errors and uncertainties
- Characterization of distributions
- Probability distributions

- Uncertainty analysis
- How to report and compare results
- Data fitting

Sources of uncertainty

Systematic error

- Reproducible inaccuracy
- Faulty instrumentation
- Neglected effect

➢ Getting rid of requires physics understanding!

  ➢ Improving measurement equipment

  ➢ Instrument calibration!

➢ **Statistical analysis DOES NOT help!**

Random fluctuations

- Statistical variations
- Limited precision
- Fluctuating results

➢ Repeating experiments required!

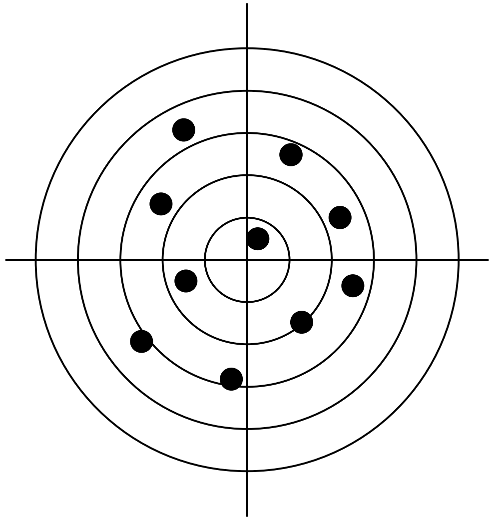
➢ **Statistical analysis helps!**

➢ ***Goal*: Estimate uncertainty to extract maximum of information!**

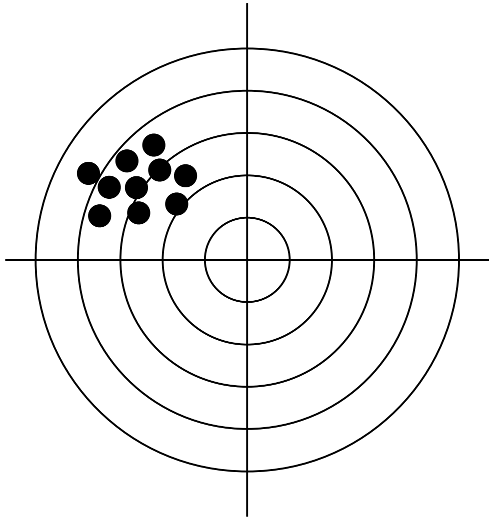
Sources of uncertainty

Accuracy: a measure how close a result is to the true value

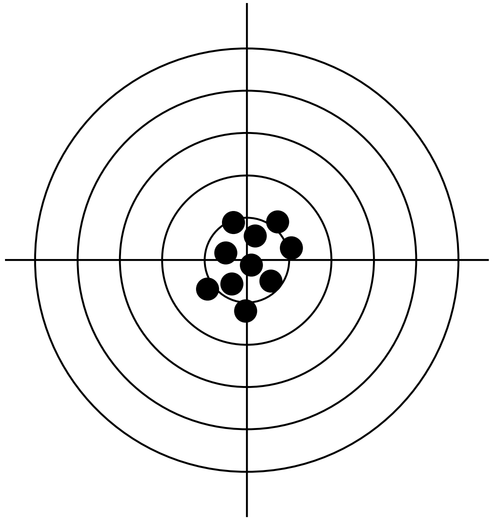
Precision: a measure how well a result has been determined without reference to the true value



accurate but imprecise



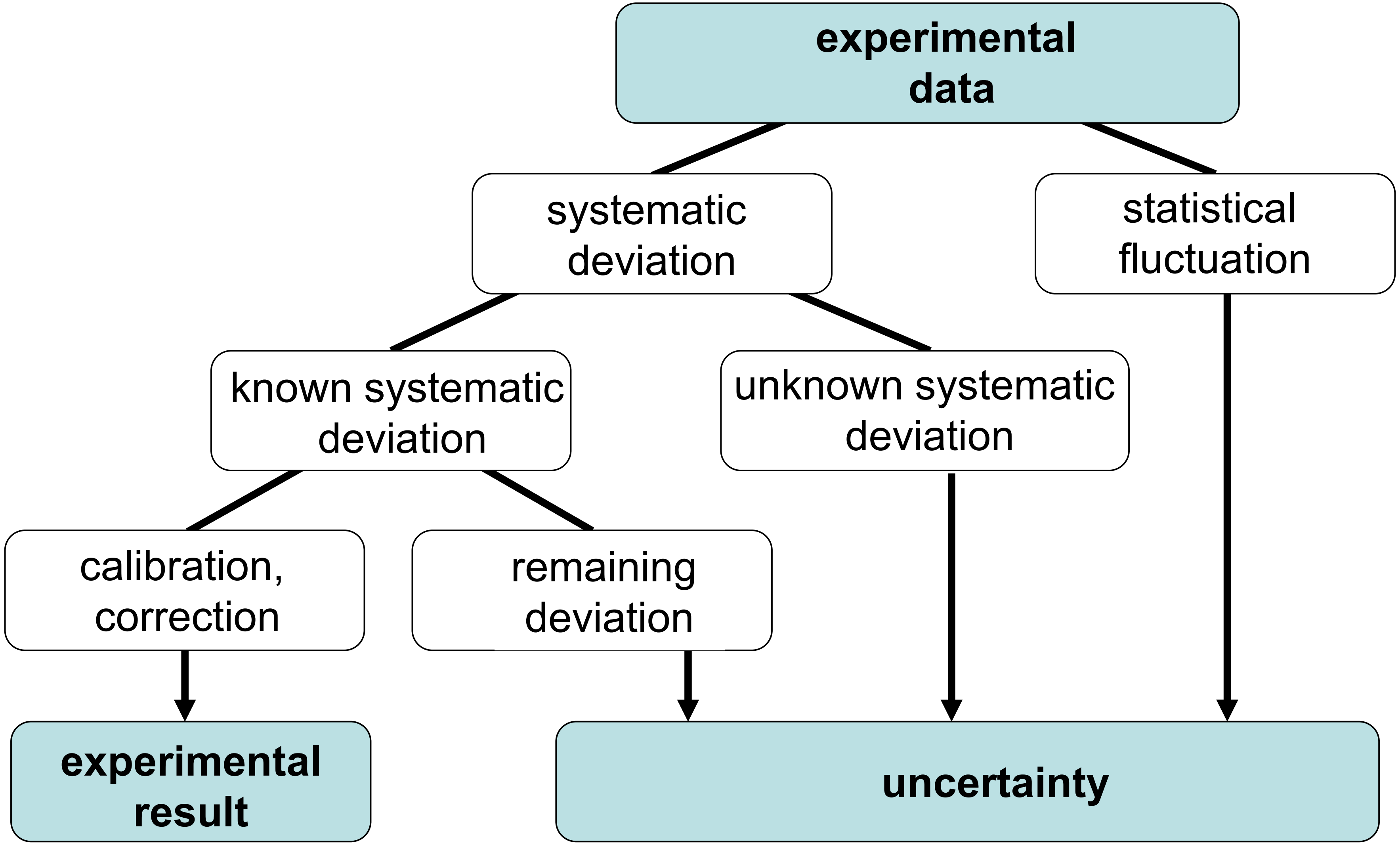
precise but inaccurate



accurate and **precise**

Accuracy is mostly related to systematic deviations, *precision* mostly to statistics.

Work flow



modified from M. Hemla, QZ. 41, 1156 (1996)

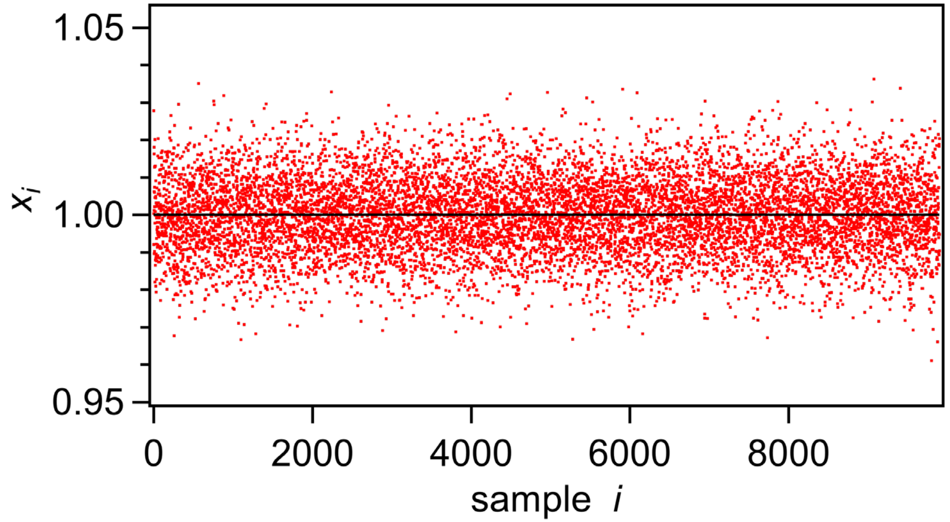
What follows deals (mostly) with statistical uncertainties!

How to make a more precise measurement?
How to estimate the measurement uncertainty?

Characterization of a distribution

Single measurements x_i group around the "true" value μ

For an infinite number of measurements:
The distribution yields the parent distribution of x_i



Mean of the parent distribution

$$\mu = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i$$

Variance of the parent distribution (**standard deviation** σ)

$$\sigma^2 := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Characterization of a distribution

Transfer of events x_i into histogram:
Probability/frequency $P(x_j)$ of possible values x_j

Discrete distribution

$$\mu = \langle x \rangle = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i = \sum_j x_j \, P(x_j)$$

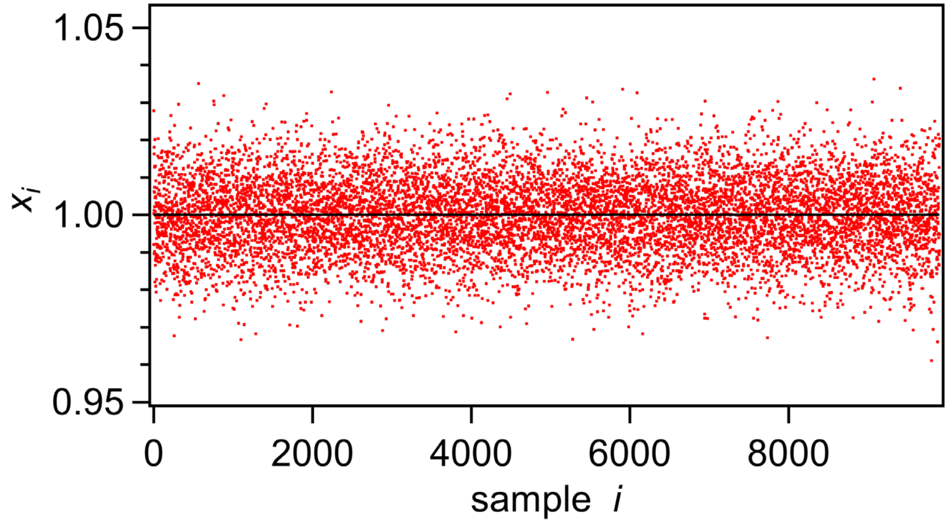
Continuous distribution

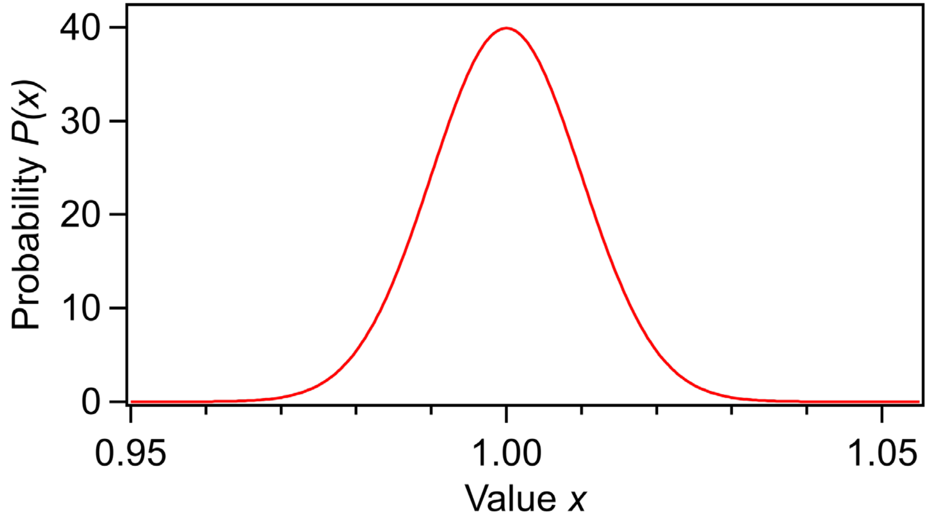
$$\mu = \int_{-\infty}^{+\infty} x' \; P(x')\mathrm{d}x'$$

Normalization

$$\int_{-\infty}^{+\infty} P(x')\mathrm{d}x' = 1$$

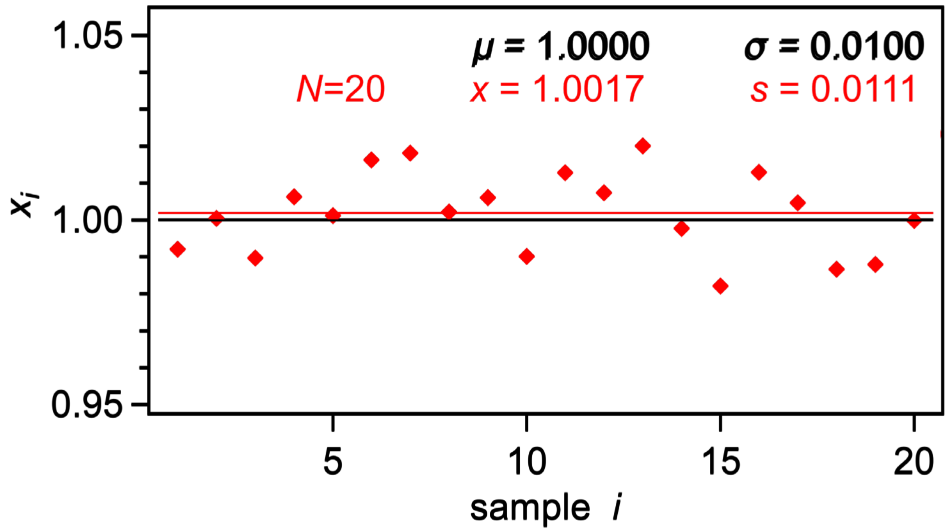
Problem: Parent distribution is in most cases not known!





Characterization of a sample distribution

For finite number of measurements:
The measurements are sample distributions
of the parent distribution



Mean of a sample distribution:

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

Variance of a sample distribution

$$s^2 := \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Standard uncertainty of a sample distribution
(standard deviation of the mean)

$$u(x) := \sigma_{\bar{x}} := \frac{s}{\sqrt{N}} = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Probability distributions

What is the propobability of observing x in an experiment.

To answer this question we need to understand the parent distribution. Most common are:

Binomial distribution $$P_B(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Poisson distribution $$P_B(x;\mu) = \frac{\mu^x}{x!} e^{-\mu}$$

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

Lorentzian distribution $$P_L(x;\mu,\Gamma) = \frac{1}{\pi}\frac{\Gamma/2}{(x-\mu)^2+(\Gamma/2)^2}$$

Voigt = convolution of Lorentzian and Gaussian

Binomial distribution

Probability of observing in a random experiment ***x*** successes out of ***n*** tries when the probability of success is ***p***

$$P_B(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!}\, p^x (1-p)^{n-x}$$

Example: **toss *n* coins**, P_B = probability to find *x* coins with head up

p: probablity head up

$1 - p$: probability head down (= tail up)

here: $p = ½$

Example: **roll n dices**, find *x* dices with a certain number, $p(x) = 1/6$, $(1 – p) = 5/6$

Binomial distribution (Dice rolling)

Example: Probability of rolling a “6” in 10 attempts

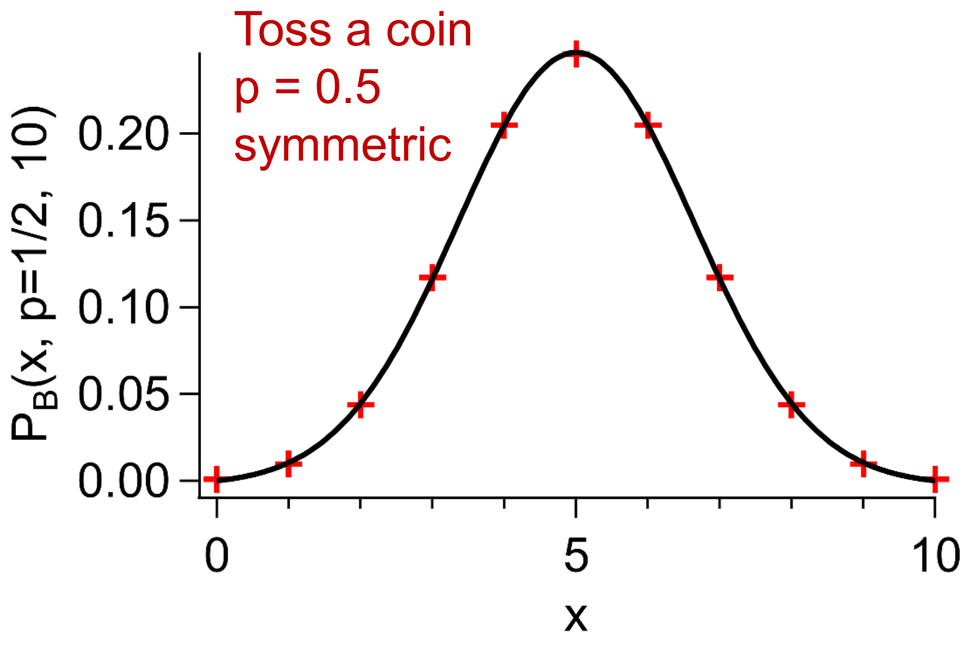
Probability of observing in a random experiment ***x*** successes out of ***n*** tries when the probability of success is ***p***

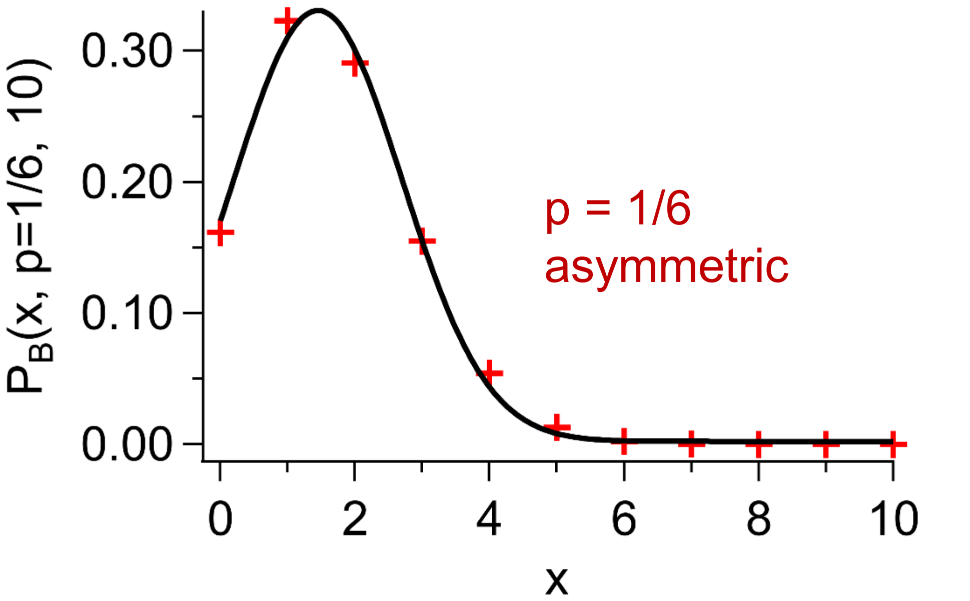
$$P_B(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$= \frac{n!}{x!(n-x)!}\, p^x (1-p)^{n-x}$$

Mean: $\mu = np$

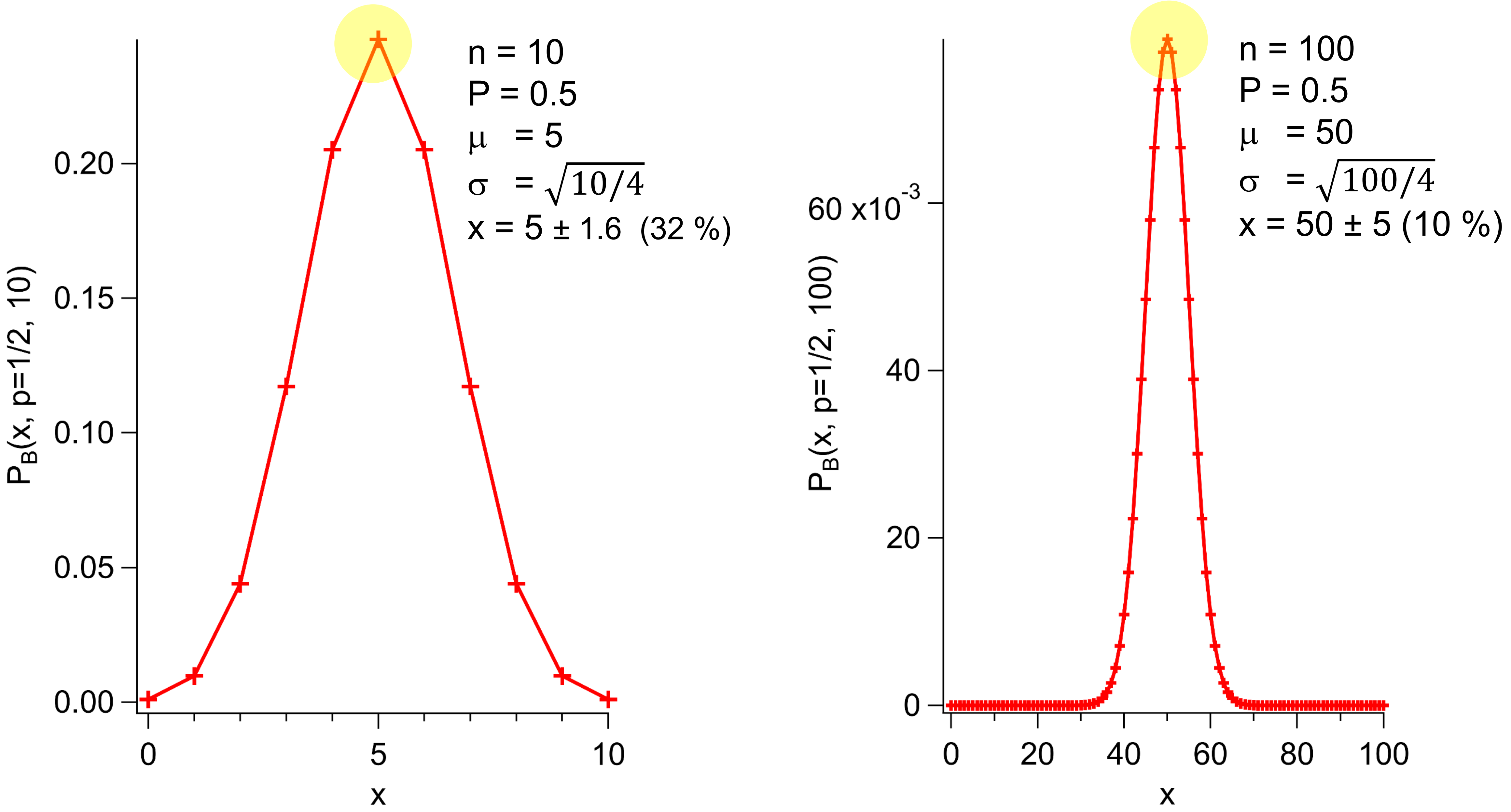
Variance: $\sigma^2 = np(1-p)$

10 dices, p = 1/6 , $\mu(6)$ = 10/6=1.67, σ^2 = 50/36, σ = 1.17





For large *n* and small steps Δ*x*, evaluation becomes impractical !



Poisson distribution: Counting statistics

For small probabilities of success, $p \ll 1$, the Binomial distribution can be approximated by

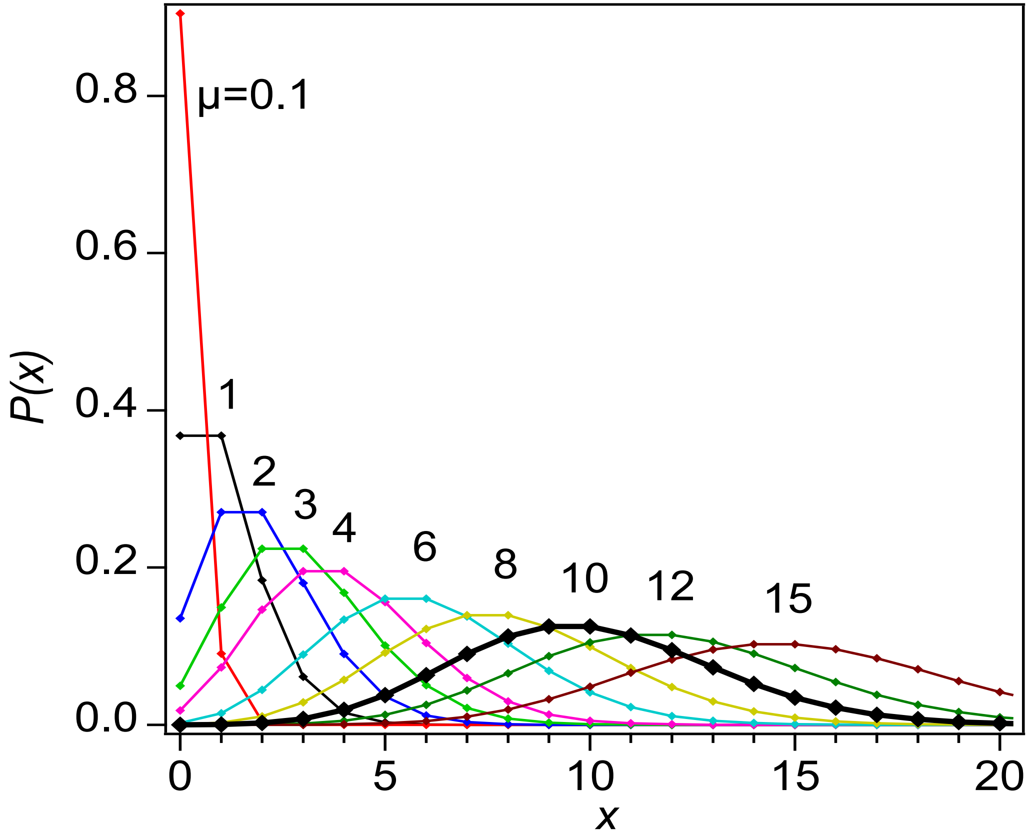
the **Poisson distribution**: $$\boxed{P_{\mathrm{P}}(x;\mu) = \frac{\mu^x}{x!}\,\mathrm{e}^{-\mu}}$$

Poisson distribution only depends on μ

Mean $$\overline{x} = \sum_{x=0}^{\infty} x\,\frac{\mu^x}{x!}e^{-\mu} = \mu$$

Variance $$\sigma^2 = \sum_{x=0}^{\infty} (x-\mu)^2\,\frac{\mu^x}{x!}e^{-\mu} = \mu$$

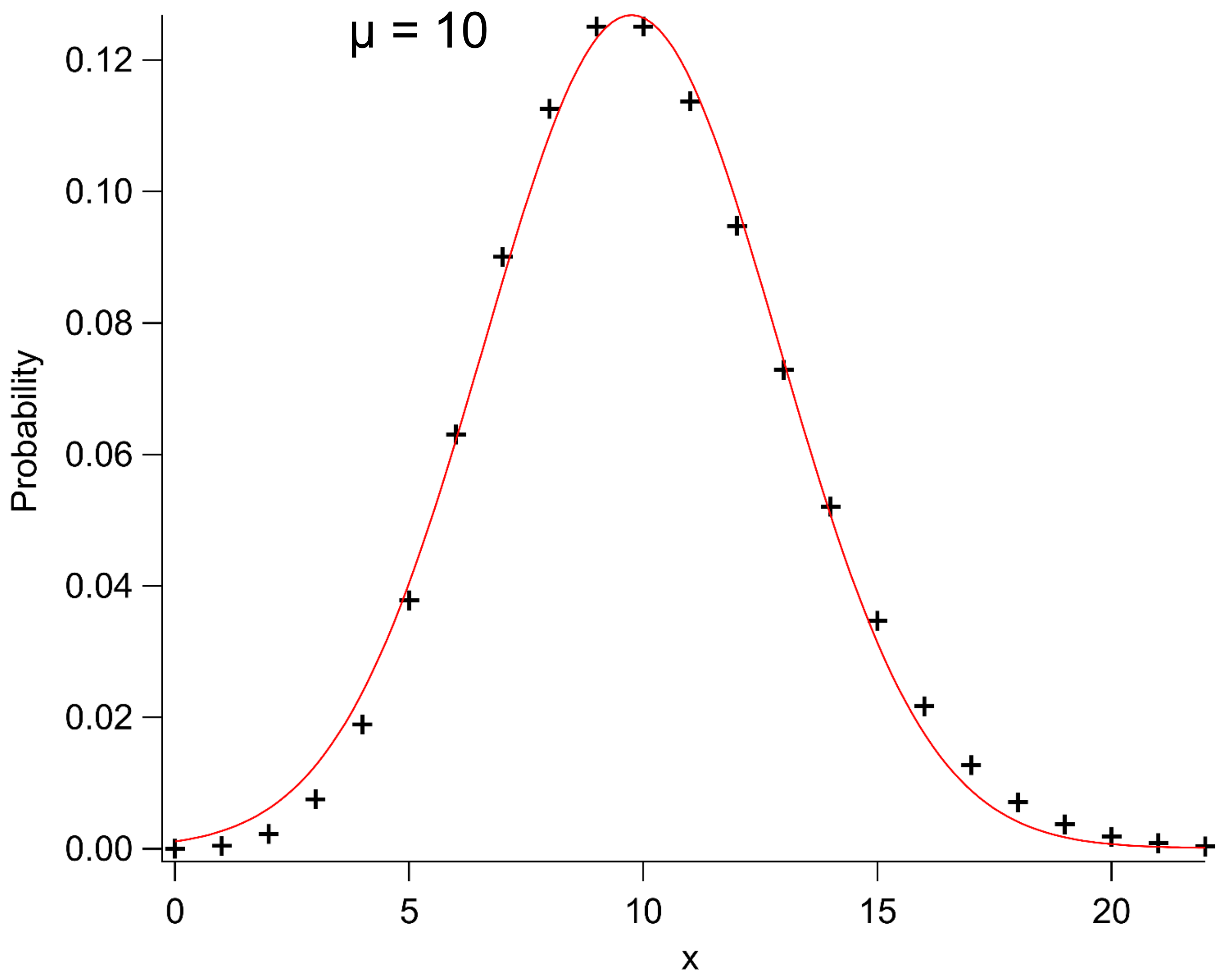
$$\boxed{\frac{\sigma}{\mu} = \frac{1}{\sqrt{\mu}}}$$



Outline

- Recap of previous seminar
- $\boxed{\text{Probability distribution: Gaussian and Lorentzian}}$
- Error propagation
- Data fitting
- How to report and present results

Poisson vs. Gaussian distribution



For μ >10, the Poisson distribution (+) is well approximated by a Gaussian distribution.

Gaussian distribution

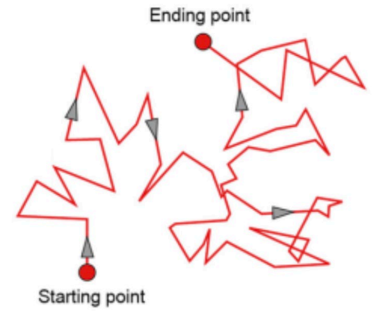
For $\mu \gg 1$ the Poisson distribution is nicely approximated by a **Gaussian distribution.**

"There are several derivations of the Gaussian distribution from first principles, none of them as convincing as the fact that the distribution is **reasonable**, that it has a fairly **simple analytic form**, and that it is accepted by convention and experimentation to be the **most likely distribution for most experiments**.

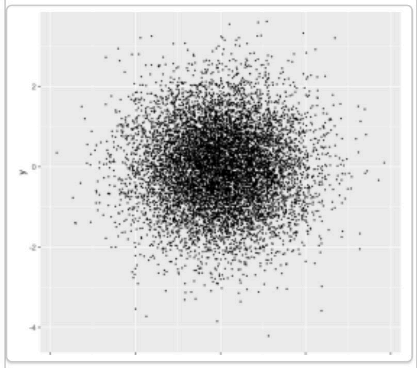
In addition, it has the satisfying characteristic that the **most probable estimate of the mean *μ*** from a random sample of observations *x* is the **average** of those observations *x*."

(From: P. R. Bevington, and D. K. Robinson, "Data Reduction and Error Analysis for the Physical Sciences")

Examples:

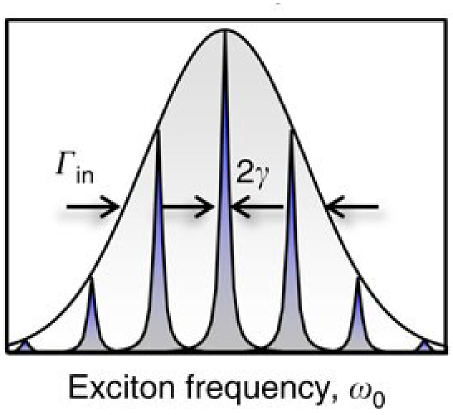


Brownian motion



10,000 points in Brownian motion

Inhomo-geneous broadening

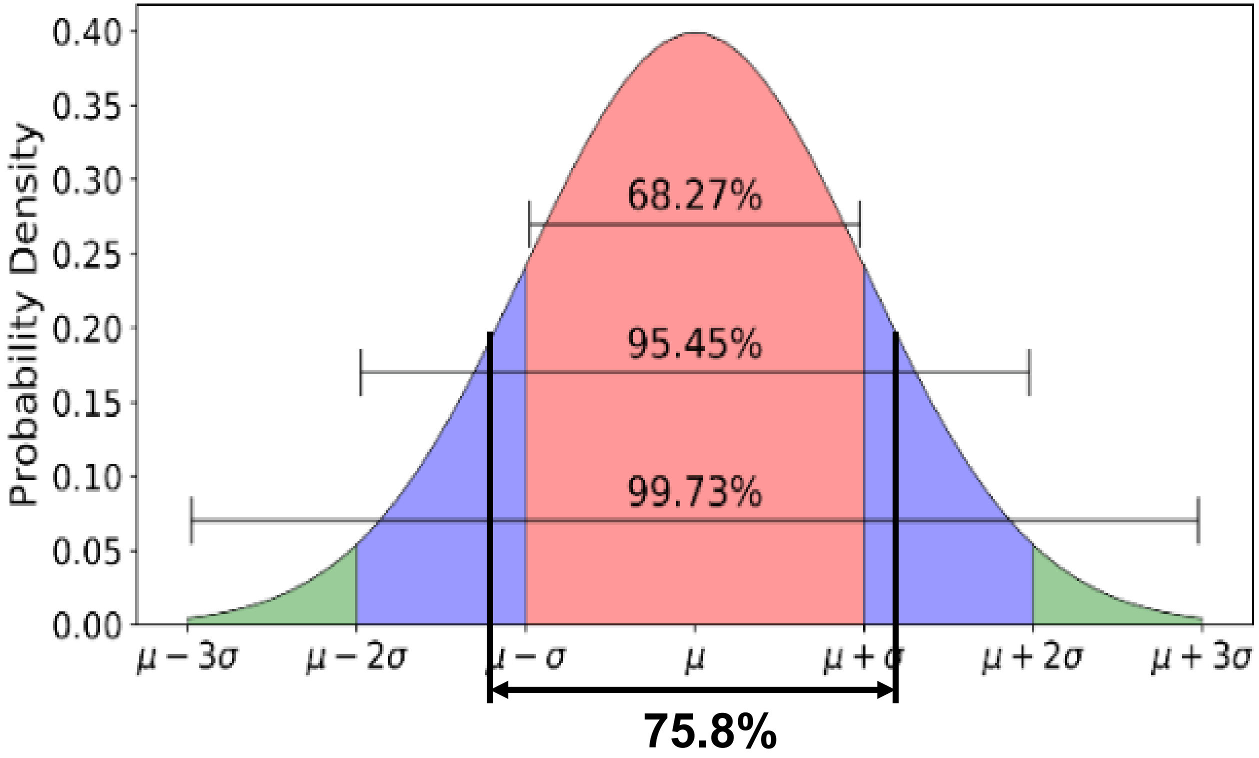


Moody et al. Nat. Commun. 6, 8315 (2015)

Gaussian distribution

Probability of one measurement being within [-σ,+σ]

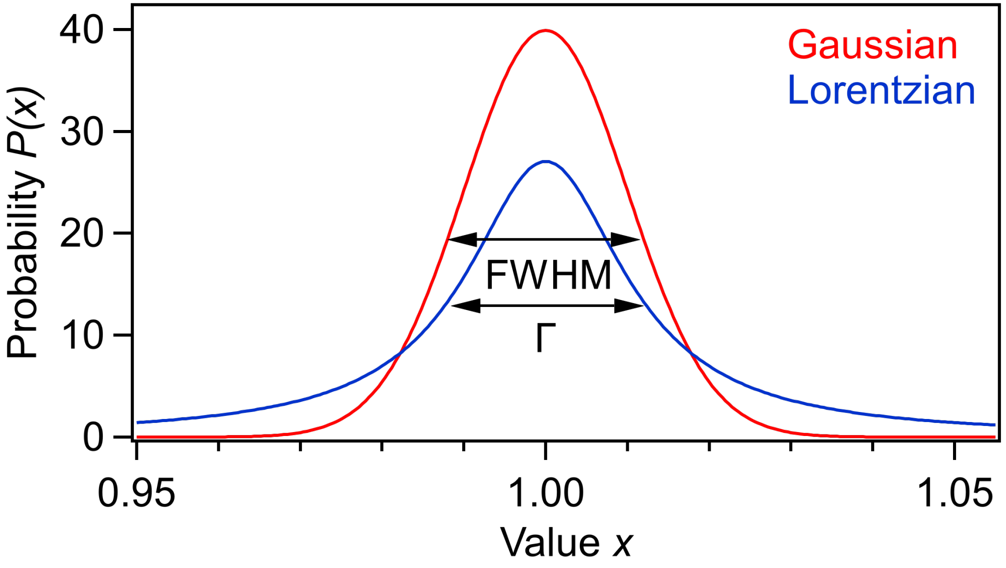
$$P_G(x \text{ within } \sigma) = \int_{\mu-\sigma}^{\mu+\sigma} dx \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$



$$\mathrm{FWHM} = 2\,\sigma\sqrt{2\ln(2)} \approx 2.354\,\sigma$$

Lorentzian distribution (Cauchy distribution)

$$P_L(x;\mu,\Gamma) = \frac{1}{\pi}\frac{\Gamma/2}{(x-\mu)^2+(\Gamma/2)^2}$$

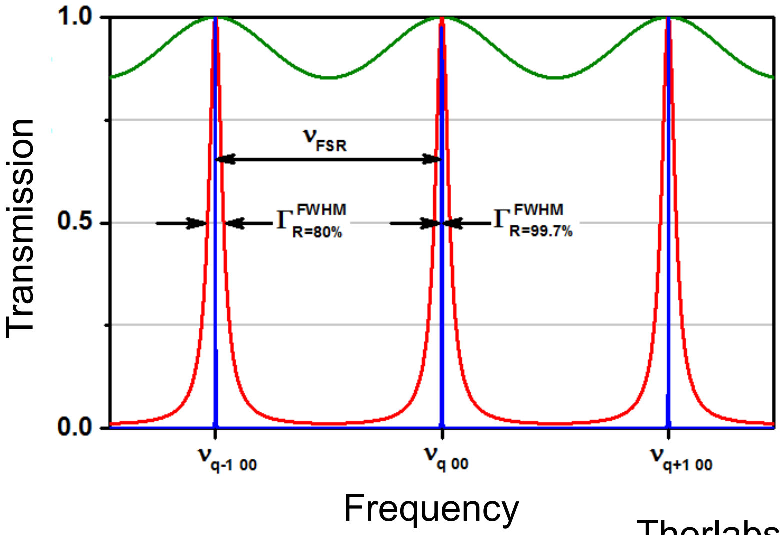


Standard deviation not defined, because of strong “wings”.

Lorentzian distribution:
Fourier transform of exponential decay

➔ Width of frequency distribution is related to damping:

- Finite lifetime of excited states
- Lineshape of laser cavity modes due to losses (HeNe laser, Fabry-Perot)

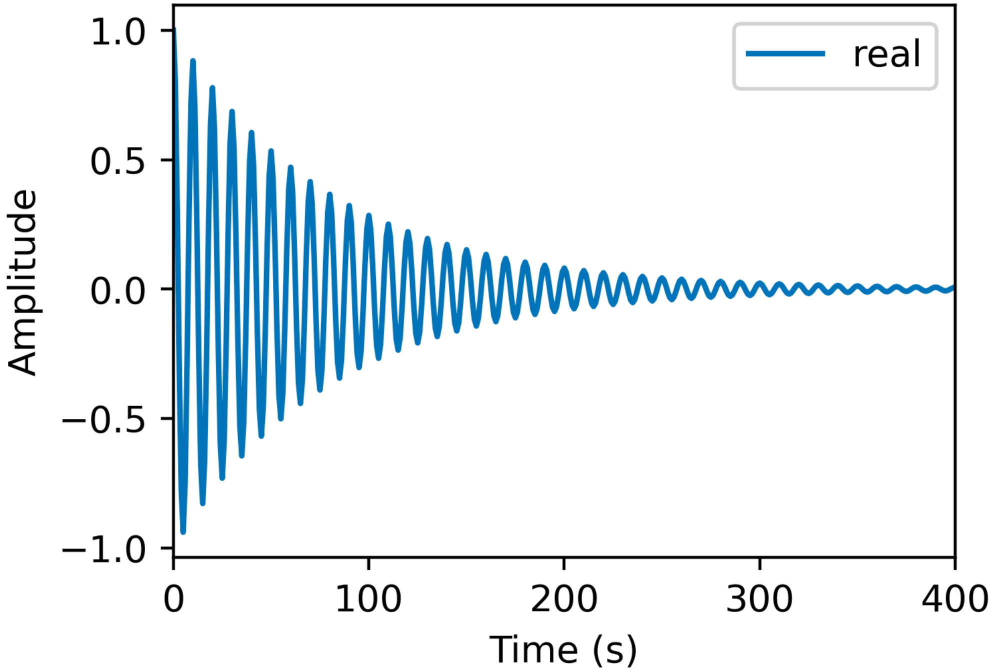


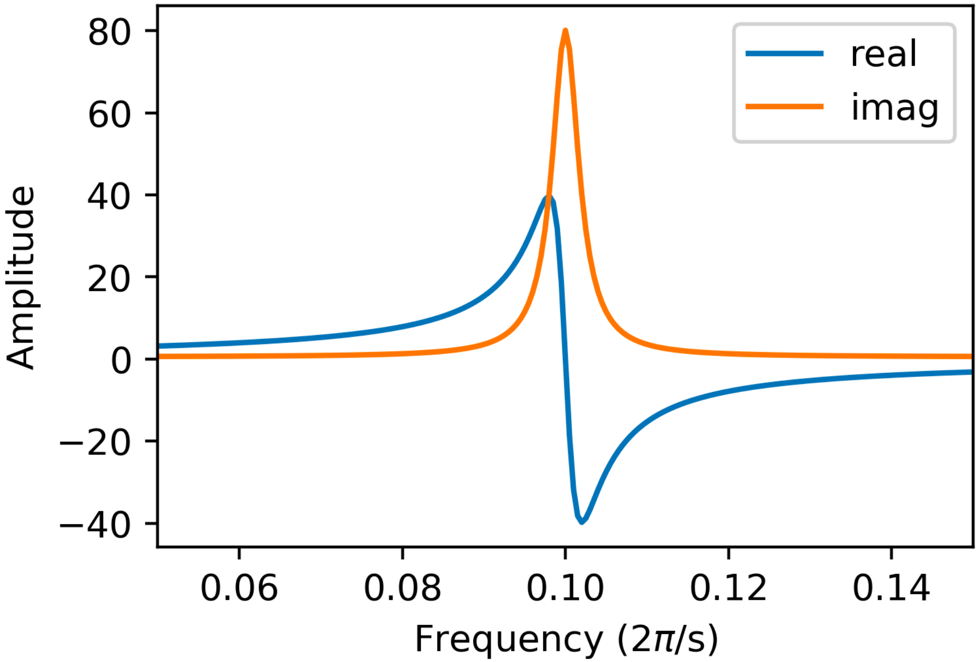
Thorlabs: Fabry-Perot tutorial

Damped oscillator (light-wave emission, dipole radiation)

$$f(t) = e^{i\omega_0 t} e^{-\gamma t}$$

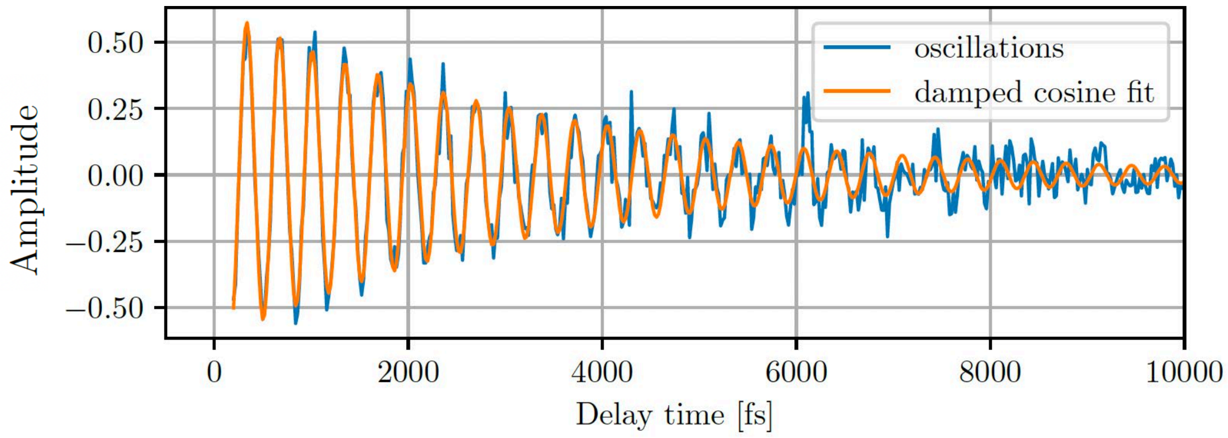
$$\mathcal{F}[f(t), \omega] \propto \frac{\omega_0 - \omega}{(\omega_0 - \omega)^2 + \gamma^2} + i \frac{\gamma}{(\omega_0 - \omega)^2 + \gamma^2}$$

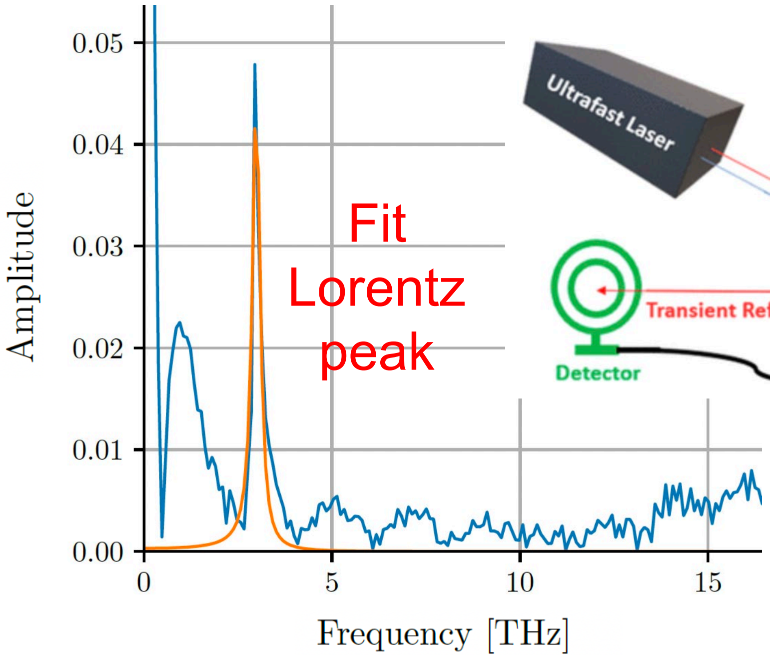


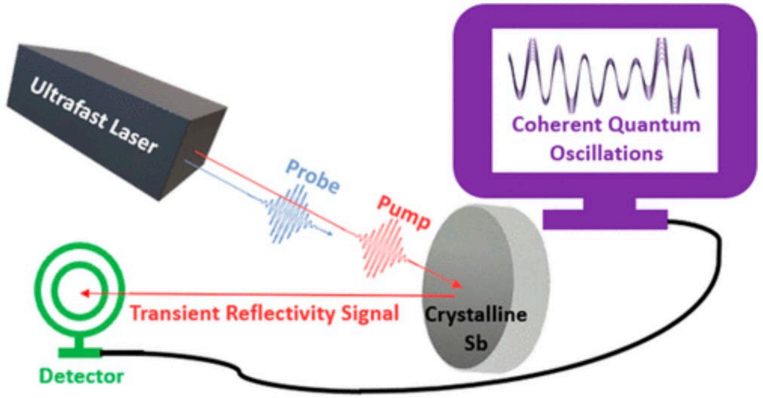


Examples of Lorentz vs Gauss Line Shape

Ma16: Coherent phonons



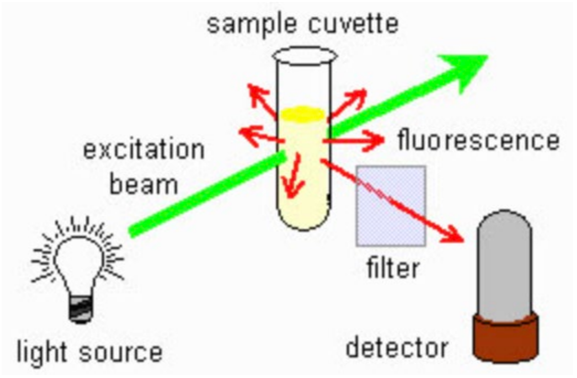


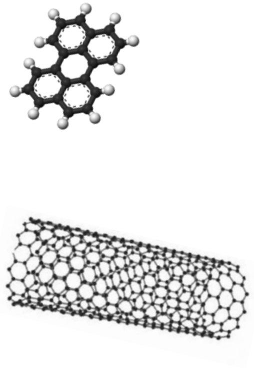


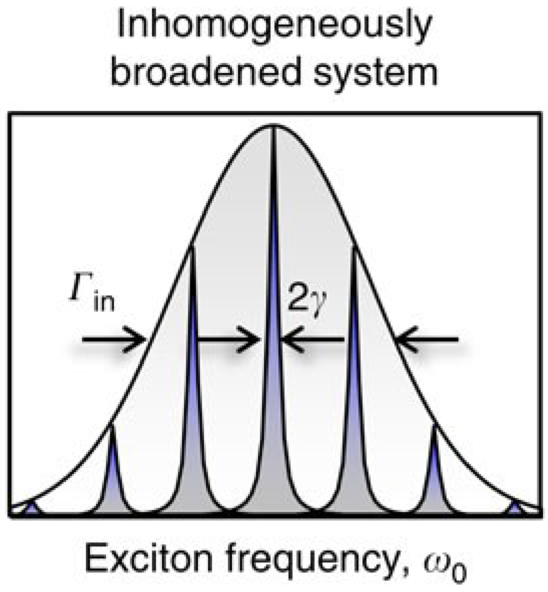
Fit Lorentz peak

Molesky et al.
J. Chem. Educ.
100, 342 (2023)

Ma9: Photoluminescence







Fit Gauss or Voigt

Moody et al. Nat. Commun. 6, 8315 (2015)

Outline

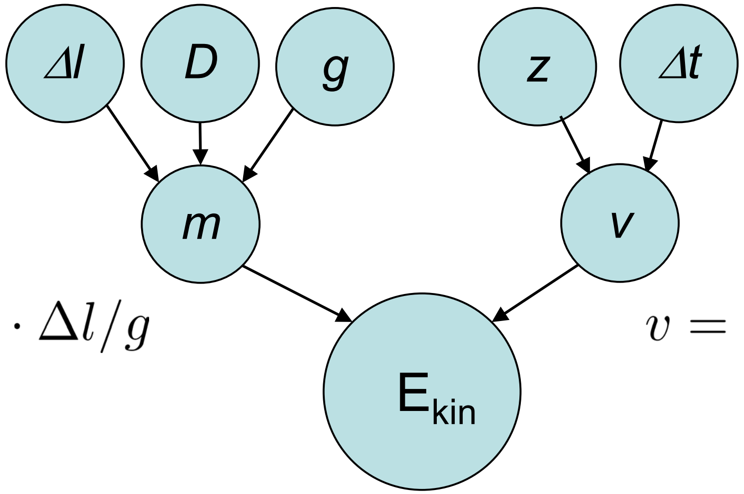
- Recap of previous seminar
- Probability distribution: Gaussian and Lorentzian
- Error propagation
- Data fitting
- How to report and present results

Models

Functional dependences of quantities

$$x = f(u, v, \dots)$$

result model function measured quantities



$$m = D \cdot \Delta l / g$$

$$v = \Delta z / \Delta t$$

$$E_{kin} = \frac{1}{2} m v^2$$

- Test model by systematically varying parameters
- Measuring $f(u,v, \dots; a,b,\dots)$
  Determine model parameters *a,b*,… by a fit
- What is the uncertainty of the result?

Models

$$x = f(u, v, \ldots)$$

result

model function

measured quantities

Data evaluation procedures:

a.) Calculate individual $x_i = f(u_i, v_i, \ldots)$ ➢ form histogram $h(x_j)|\mathrm{d}x_j|$

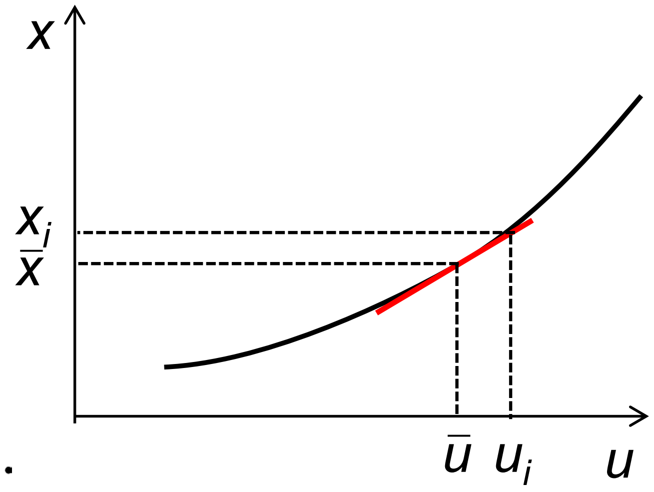
➢ If u_i and v_i belong together

b.) Calculate means and take most probable values $\bar{x} = f(\bar{u}, \bar{v}, \ldots)$

➢ If u_i and v_i from independent measurements

➢ Error propagation?

Propagation of uncertainty



For measured quantities *u, v,* …

$$x = f(u, v, ...)$$

Assuming that the deviations in u_i, v_i,… are small, deviations in x_i can be expressed as

Taylor expansion $$(x_i - \overline{x}) = \left.\frac{\partial f}{\partial u}\right|_{\overline{u}} (u_i - \overline{u}) + \left.\frac{\partial f}{\partial v}\right|_{\overline{v}} (v_i - \overline{v}) + \ldots$$

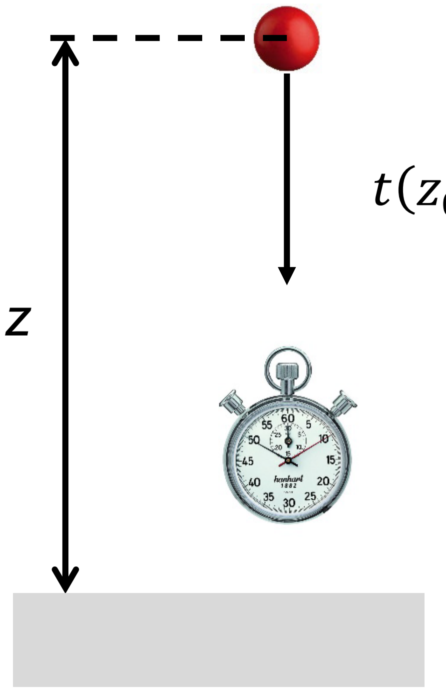
(derivatives evaluated for variables fixed at mean)

Variance: $$s_f^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2 = \frac{1}{N-1}\sum_i \left[(u_i - \bar{u})\left(\frac{\partial f}{\partial u}\right) + (v_i - \bar{v})\left(\frac{\partial f}{\partial v}\right) + ...\right]^2$$

$$s_f^2 = s_u^2\left(\frac{\partial f}{\partial u}\right)^2 + s_v^2\left(\frac{\partial f}{\partial v}\right)^2 + 2s_{uv}^2\left(\frac{\partial f}{\partial u}\right)\left(\frac{\partial f}{\partial v}\right) + ...$$

Propagation of uncertainty

$$s_f^2 = s_u^2 \left(\frac{\partial f}{\partial u}\right)^2 + s_v^2 \left(\frac{\partial f}{\partial v}\right)^2 + 2 s_{uv}^2 \left(\frac{\partial f}{\partial u}\right)\left(\frac{\partial f}{\partial v}\right) + \ldots$$



$$t(z_0, t_0) = \sqrt{\frac{2(z + z_0)}{g}} + t_0$$

$$\frac{\partial t}{\partial z_0} < 0, \qquad \frac{\partial t}{\partial t_0} > 0$$

Covariance: characterizing the correlation between deviations in different variables

$$s_{uv}^2 = \frac{1}{N-1} \sum (u_i - \bar{u})(v_i - \bar{v})$$

- can be positive or negative
- vanishes for uncorrelated quantities

Propagation of standard uncertainty

$$\sigma_{\bar{x}}^2 = \frac{s_f^2}{N} = \sigma_{\bar{u}}^2 \left(\frac{\partial f}{\partial u}\right)^2 + \sigma_{\bar{v}}^2 \left(\frac{\partial f}{\partial v}\right)^2 + \cancel{2\sigma_{\bar{u}\bar{v}}^2 \left(\frac{\partial f}{\partial u}\right)\left(\frac{\partial f}{\partial v}\right)} + \ldots$$

for uncorrelated errors
because

$$s_{uv}^2 = \frac{1}{N-1} \sum (u_i - \bar{u})(v_i - \bar{v}) = 0$$

Often it is sufficient to use just the biggest term!

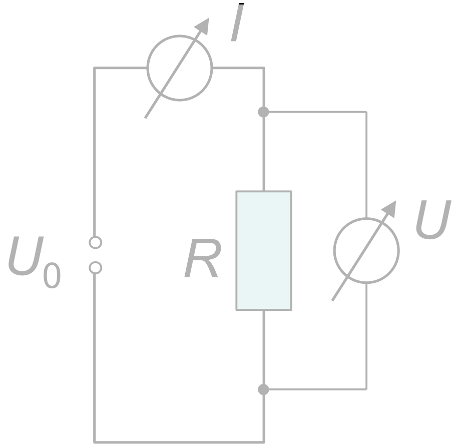
Propagation of uncertainty: *examples*

$$\sigma_{\bar{x}}^2 = \frac{s_f^2}{N} = \sigma_{\bar{u}}^2 \left(\frac{\partial f}{\partial u}\right)^2 + \sigma_{\bar{v}}^2 \left(\frac{\partial f}{\partial v}\right)^2$$

Resistance measurement:

Measure current *I* and voltage *U* on resistor to determine resistance *R*

$$R = \frac{U}{I} \qquad \sigma_R \approx \sqrt{\sigma_U^2 \left(\frac{1}{I}\right)^2 + \sigma_I^2 \left(\frac{U}{I^2}\right)^2}$$



Conversion of wavelength to energy:

$$E\ [\mathrm{eV}] = \frac{hc}{\lambda} \approx \frac{1240}{\lambda[\mathrm{nm}]}\,\mathrm{eV} \qquad \lambda = (635 \pm 6)\ \mathrm{nm}$$

$$\rightarrow E = (? \pm ?)\ \mathrm{eV}$$

Outline

- Recap of previous seminar
- Probability distribution: Gaussian and Lorentzian
- Error propagation
- $\boxed{\text{Data fitting}}$
- How to report and present results

χ^2 test

Comparing the assumed probability distribution function P to a measured histogram $h(x_j)$:

$$\chi^2 = \sum_{j=1}^{n} \frac{[h(x_j) - NP(x_j)]^2}{\sigma_j^2(h)}$$

$n = 9\,\text{bins}$ $n_c = 3$ Gaussian: $A, \overline{x}, \sigma$

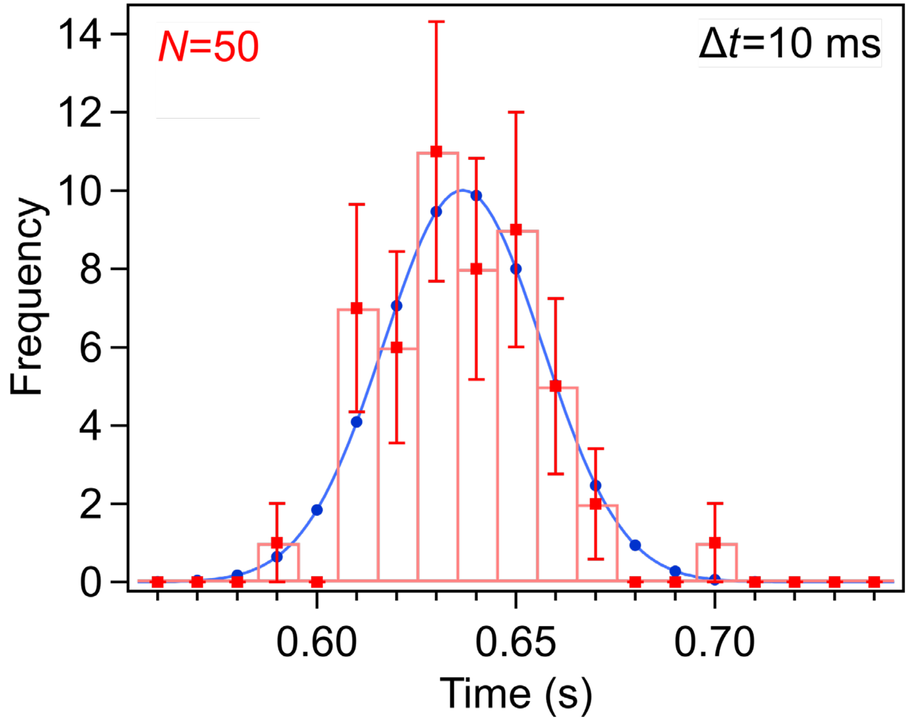
Expected for pure counting statistics:

$$\langle \chi^2 \rangle = \nu = n - n_c$$

data/bins # constraints

Reduced chi-squared:

$$\langle \chi_\nu^2 \rangle = \left\langle \frac{\chi^2}{\nu} \right\rangle = 1$$



χ^2 = 3.27

ν = 9 – 3 = 6

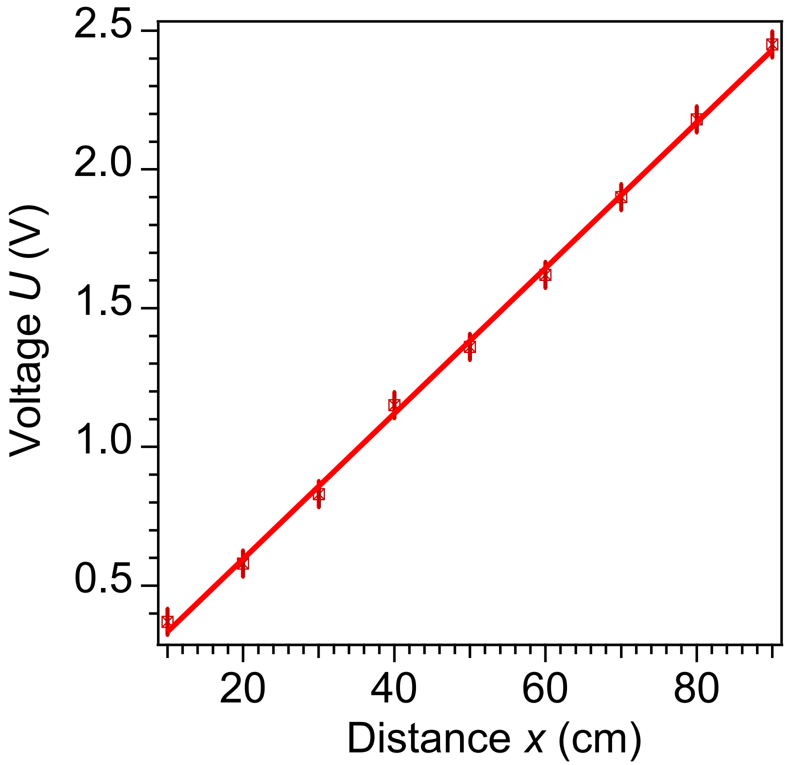
χ_ν^2 = 0.55

Least-squares fit to a line

Fit function:

$$\boxed{y(x) = a + bx}$$

Data set of (x_i, y_i) with uncertainties $\sigma_{yi} = \sigma_i$



Assuming Gaussian uncertainties, the most probable parameters *a*, *b* are obtained at minimum of

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - y(x_i)]^2}{\sigma_i^2} = \sum_{i=1}^{N} \left[\frac{1}{\sigma_i}(y_i - a - bx_i)\right]^2 \approx \nu$$

good fit

Reduced chi-square

$$\chi_\nu^2 = \frac{\chi^2}{\nu} \approx 1$$ with $\nu = N - N_c$ (here N_c= 2)

χ^2 vs R^2

Chi-square

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - y(x_i)]^2}{\sigma_i^2}$$

Reduced chi-square

$$\chi_\nu^2 = \frac{\chi^2}{\nu} \approx 1$$

- *Takes uncertainties into account*
- Test whether data follow particular distribution

R squared

(coefficient of determination)

Fit function

$$R^2 = 1 - \frac{\sum_{i=1}^{N} [y_i - y(x_i)]^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \approx 1$$

for good fit

Mean of data

- *Only from residuals*
- *Measure for the variation of the data* compared to model

Least-squares fit to a line: Linear regression

Analytical solution:

$$a = \frac{1}{\Delta}\left(\boxed{\sum \frac{x_i^2}{\sigma_i^2}}\boxed{\sum \frac{y_i}{\sigma_i^2}} - \boxed{\sum \frac{x_i}{\sigma_i^2}}\boxed{\sum \frac{x_i y_i}{\sigma_i^2}}\right)$$

$$b = \frac{1}{\Delta}\left(\boxed{\sum \frac{1}{\sigma_i^2}}\sum \frac{x_i y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2}\sum \frac{y_i}{\sigma_i^2}\right)$$

with $\Delta = \sum \frac{1}{\sigma_i^2}\sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i}\right)^2$

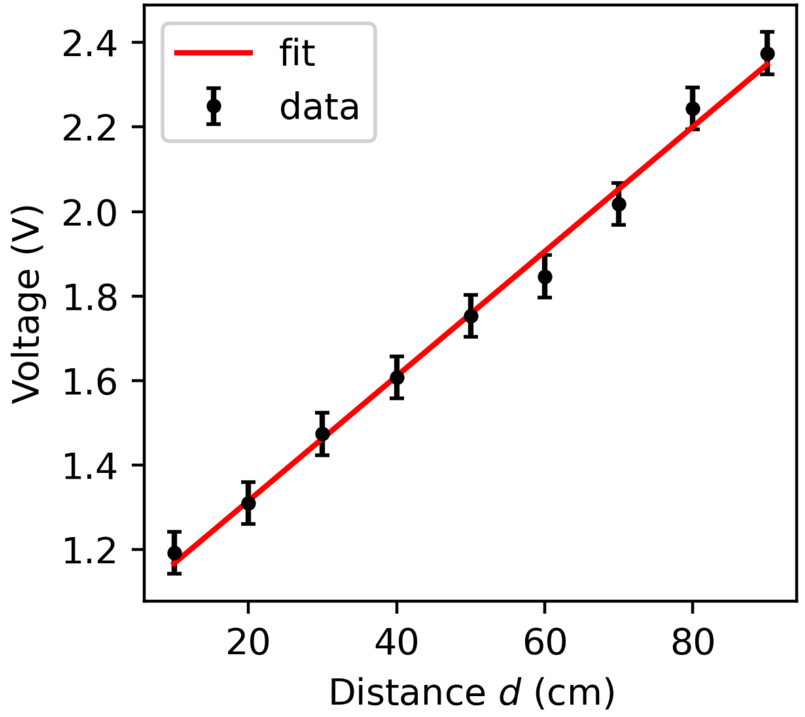
Variance of fit parameters: $\sigma_a^2 = \frac{1}{\Delta}\sum \frac{x_i^2}{\sigma_i^2}$ $\quad \sigma_b^2 = \frac{1}{\Delta}\sum \frac{1}{\sigma_i^2}$

Estimation of uniform variance $\sigma'^2 \simeq s^2 = \frac{1}{N-2}\sum (y_i - a - bx_i)^2$

Example 1: Linear fit

Potential drop along a wire

$$U(d) = RI\, d/d_0$$



σ_U = 0.05 V

R^2 = 0.994 → good fit

χ^2 = 3.21 for N=9, ν=7

- *χ^2 is too small.*
- *Standard deviation was overestimated*

Corrected estimate:

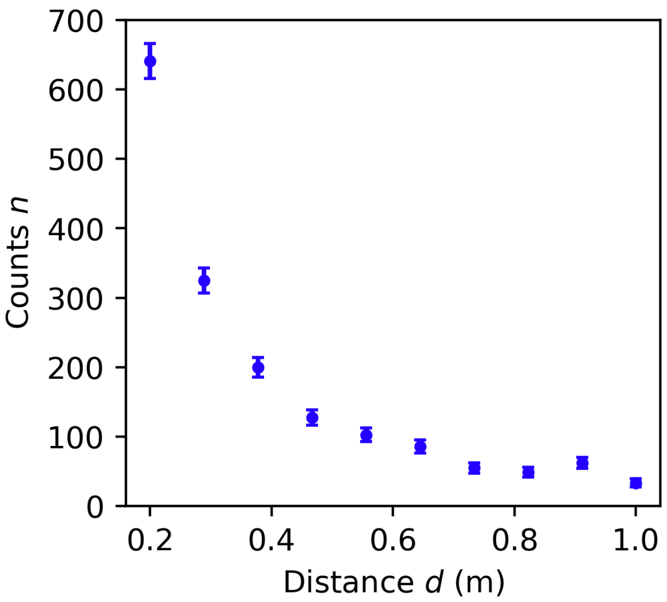
$\sigma' \approx \sigma\sqrt{\chi_\nu}$ = 0.033 V

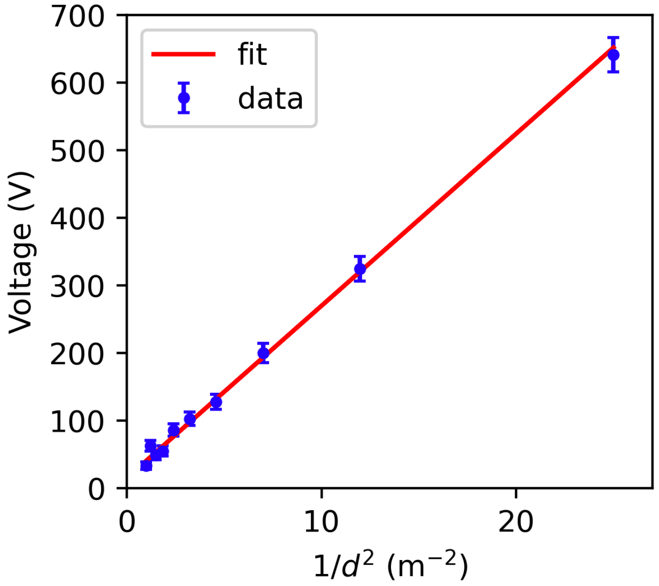
Note: Standard deviation concerns only the **random errors.**

Total uncertainty due to **systematic errors** might be larger!

Example 2: Linearize data

Radioactive source at distance *d*





R^2 = 0.998

χ^2 = 9.3 for *N*=10, ν=8

➢ Estimation of σ_n fits well.

$$I(d) = I_0/d^2 + I_{\mathrm{dark}}$$

⇨ Fit allows separation of signal and background

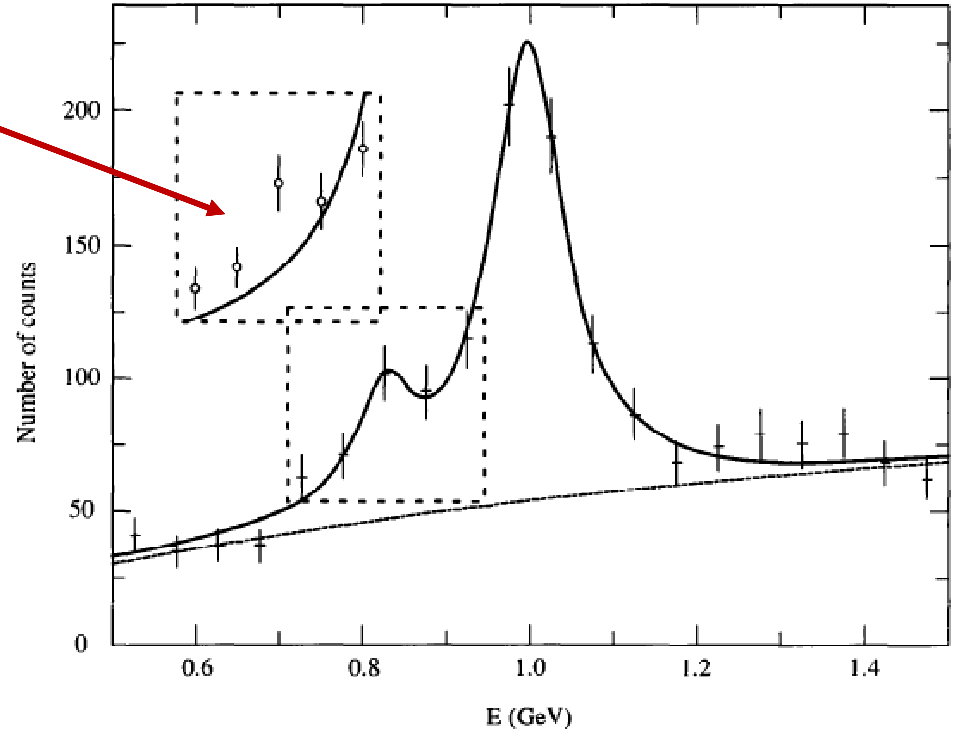
$$\sigma_n = \sqrt{n}$$

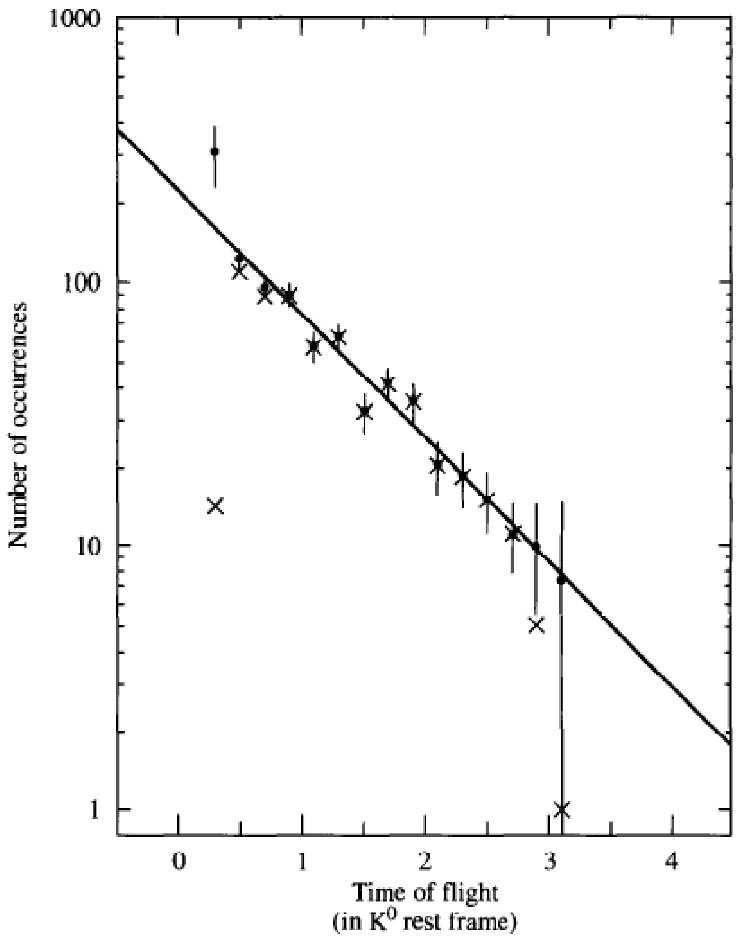
Non-linear fitting

If fitting function is non-linear in parameters *a*, *b*, …

- try to transform to linear fit
- numerical least-mean-squares fit

Fit with one peak





“A fit can never be better than your model function.”

Outline

Freie Universität Berlin

- Recap of previous seminar
- Probability distribution: Gaussian and Lorentzian
- Error propagation
- Data fitting
- How to report and present results

Instrumental

Fluctuations in reading of physical instrument
(limited instrumental precision)

Often independent of the actual value

e.g. length, mass, voltage, current…

- Refer to instrumentation manuals or other test measurements (*external estimate*)

- Estimate of standard deviation from repeated measurements (*internal estimate*)

Statistical uncertainties

Counting experiments:
x represents the count rate in a detector

Statistical fluctuations for finite sample

- Uncertainty can be deduced from data

- Poisson statistics! $\sigma = \sqrt{\mu}$

How to report measurement results and uncertainties

- for $N = 1$ measurement, report instrumental uncertainty
- for $N > 1$ measurement, report $\boxed{x = \bar{x} \pm \sigma_{\bar{x}}}$ with $\sigma_{\bar{x}} = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{N} (x_i - \bar{x})^2}$

(or from fit)

- For fit also report reduced chi-squared and R^2 value.
- Indicate estimated *systematic* uncertainty in addition.
- Report uncertainty with 2 significant digits if intermediate result, and 1 digit for final result.
- Report result with same precision as uncertainty.

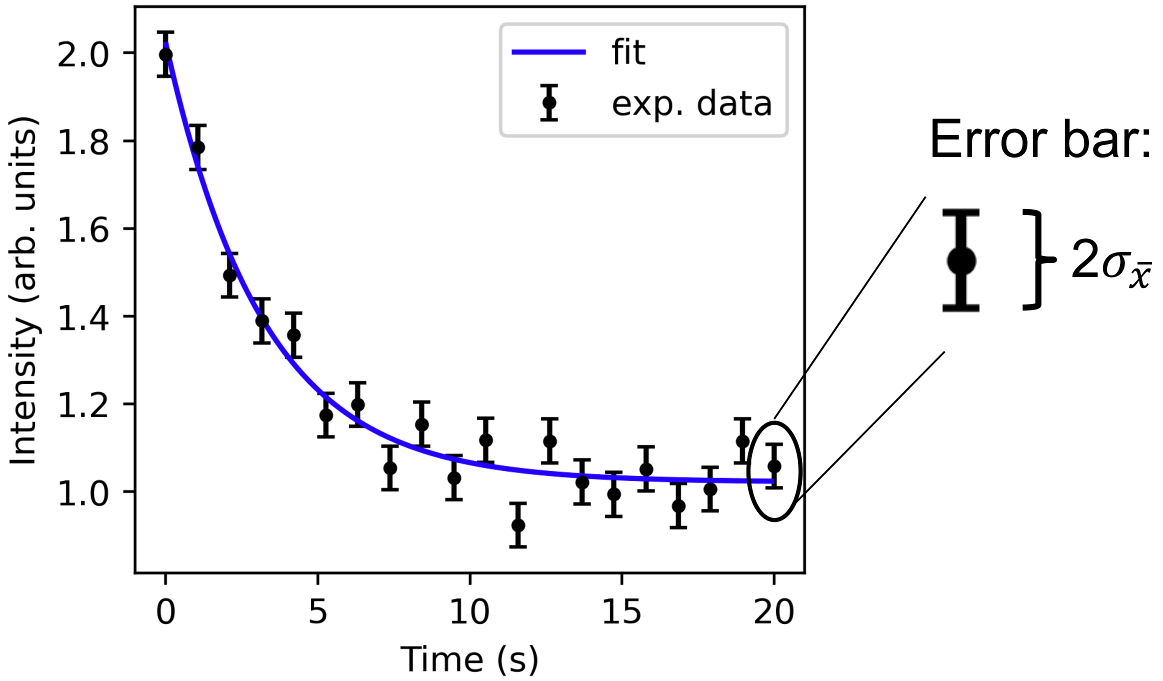
How to state measurement results and uncertainties

Exercise: State the quantity with uncertainty – convert to scientific notation if helpful

quantity	standard uncertainty	
50.003 m	0.1m	(50.0 ± 0.1) m
53612 s	300 s	$(5.36 \pm 0.03)*10^4$ s
632.497 nm	0.153 nm	$(632.5 \quad \pm 0.2)$ nm
367134 C	783 C	$(3.671 \pm 0.008)*10^5$ C

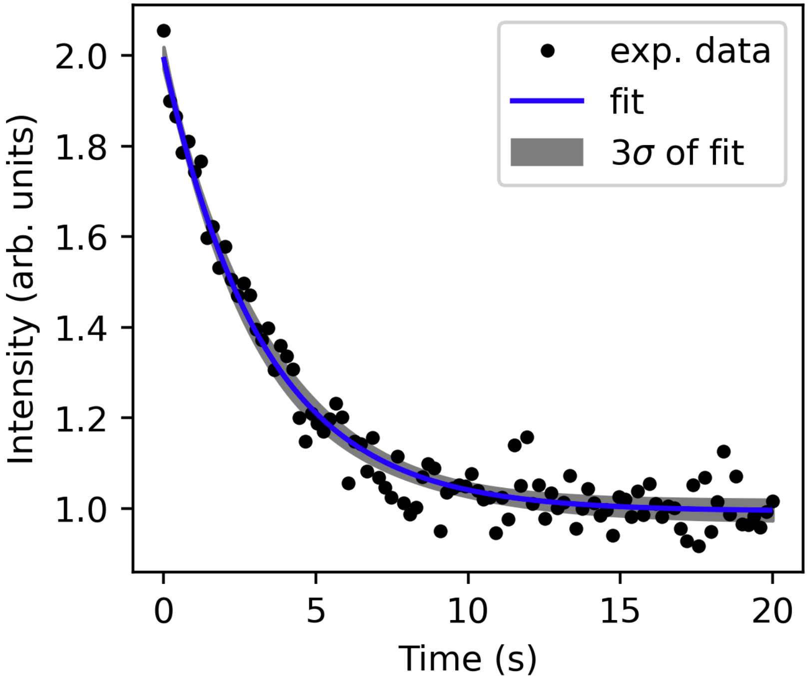
How to plot uncertainties of measured data

For small data set:



- *Show error bars*
- *Error bars show two standard deviations*

For large data set:



- *Plot uncertainty of fit*
- *Or use binning*

Take home…

- Keep good experimental records
- Try to measure in several ways
- Calibrate instruments, read their manuals
- Check rounding errors
- Be mindful of largest uncertainty source
- The uncertainty of a fit result does not consider systematic deviations
- **Never report measurement results without uncertainties**

Literature

- P. R. Bevington, and D. K. Robinson, “*Data Reduction and Error Analysis for the Physical Sciences*”
- P. Möhrke, and B.-U. Runge, “*Arbeiten mit Messdaten*”
- “*Guide to uncertainty propagation and Error analysis*” Stony Brook U.
- “*A beginner guide to uncertainty of measurement*” by Stephanie Bell
- Slides and python code examples will be uploaded to website

Examples with Python