THE BASICS OF ERROR ESTIMATION

What is the plan?

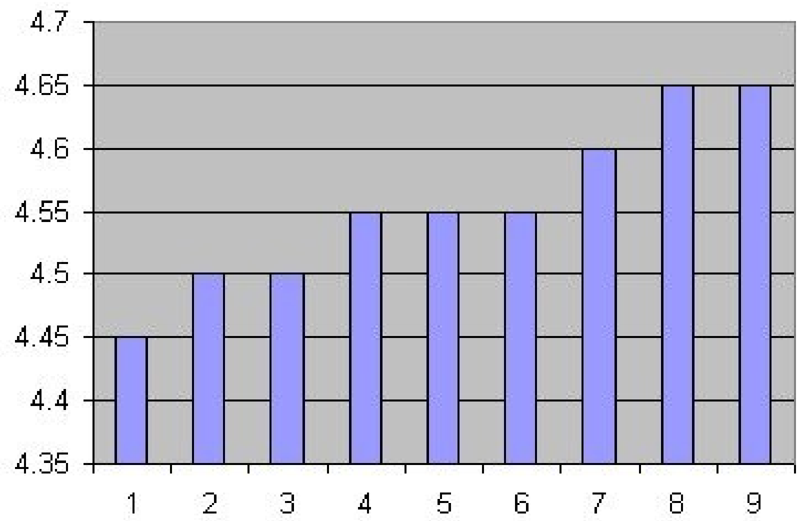
I. Statistical estimations from the observational data

II. Poisson and normal distributions

III. Filtration of the data

Mean values

1) **Mean** or **“average”**: $\langle x \rangle = \frac{1}{N} \sum_i x_i$



Imagine we have collection of nine data points:

4.45, 4.50, 4.50, 4.55, 4.55, 4.55, 4.60, 4.65, 4.65.

Mean: (4.45 + 4.50 + 4.50 + 4.55 + 4.55 + 4.55 + 4.60 + 4.65 + 4.65)/9 = **4.56**

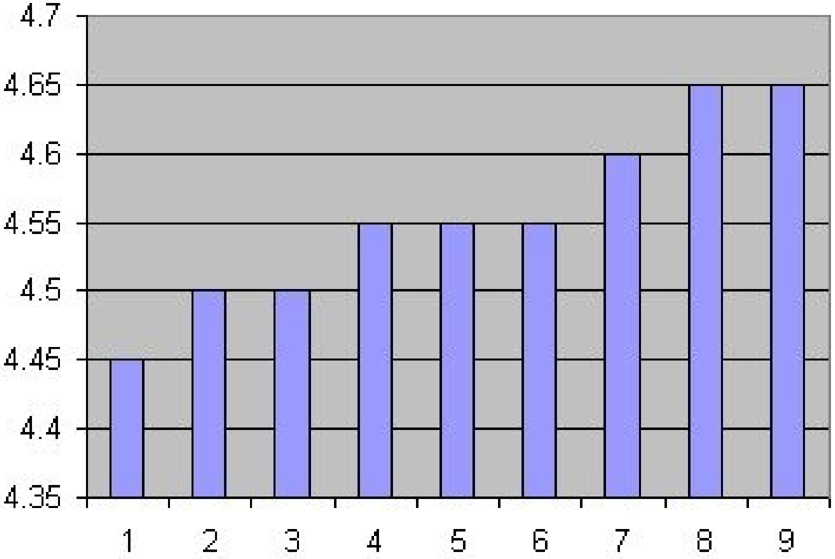
Mean values

2) **Median:**

The individual value from the collection such that

½ the observation are less

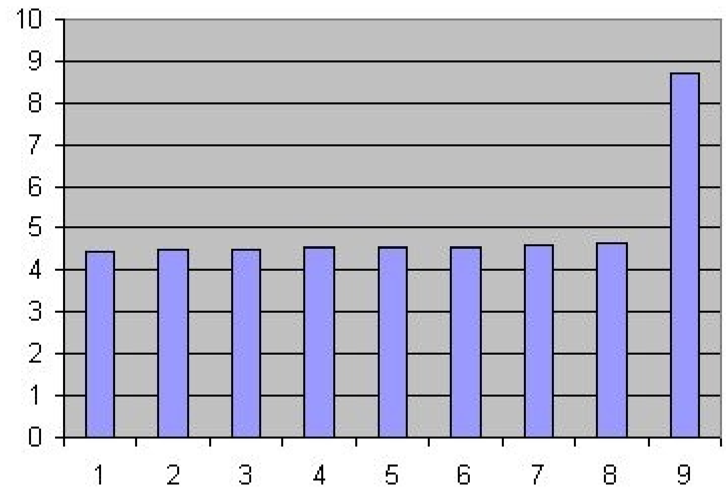
and ½ are greater.



For our data set **median** is **4.55**

Mean values

! Median is unaffected by a single point that is the way out of the main group of points



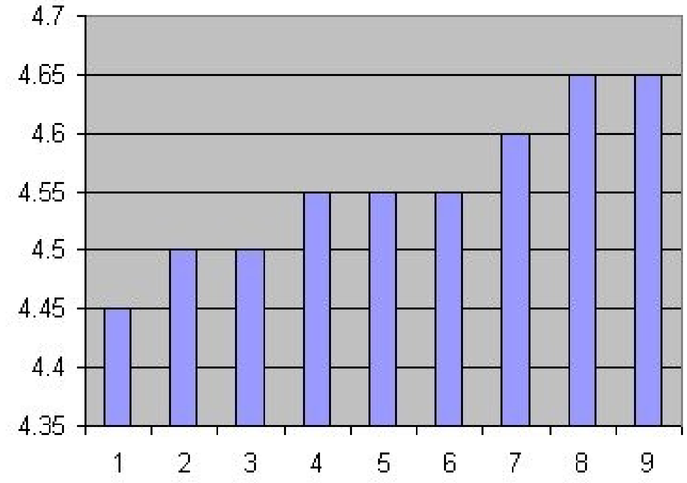
For example we consider a different data set:

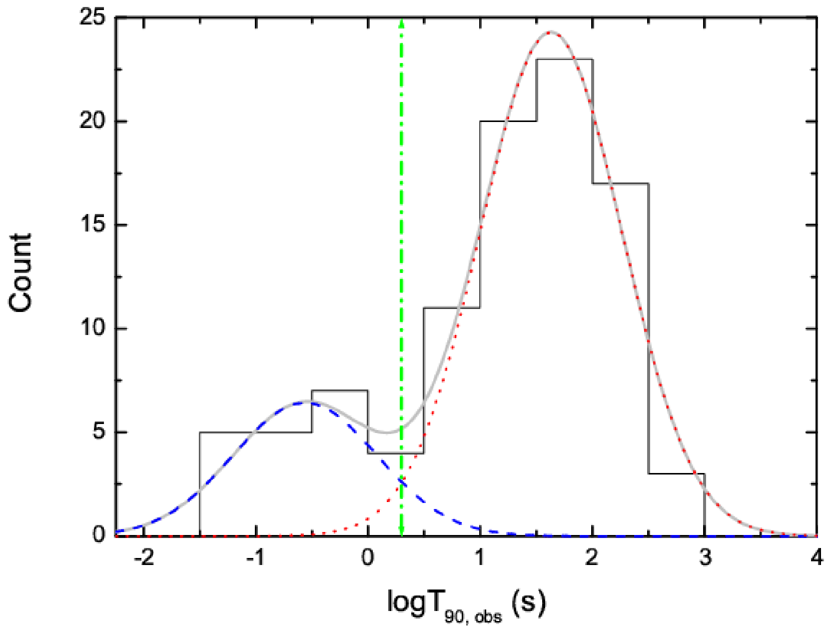
4.45, 4.50, 4.50, 4.55, 4.55, 4.55, 4.60, 4.65, **8.7**

Mean: 5.01 **Median:** 4.55

Mean values

3) **Mode:** The most frequently occurring value





Mean: 4.56 **Median:** 4.55

Mode: 4.55

Sometimes the distribution could be a multimodal

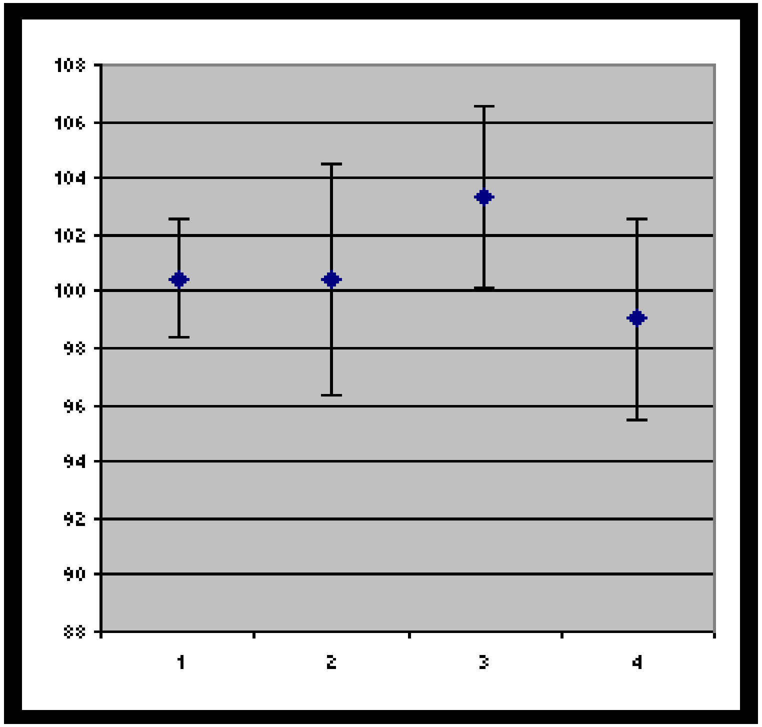
Mean values

4) The weighted mean:

$$\langle x \rangle = \sum_i w_i x_i$$

If we have a deviation for each experimental value we should use the weight:

$$w_i = \frac{1/\sigma_i^2}{\sum_i 1/\sigma_i^2}$$



Error estimations

1) **Simple** deviation: $\Delta_i = x_i - \langle x \rangle$

2) **Mean** deviation: $\langle \Delta \rangle = \frac{1}{N-1} \sum_i |x_i - \langle x \rangle|$

3) **Standard** deviation ("**root mean square** deviation" or **rms** or **sigma**):

$$\sigma_x = \sqrt{\frac{1}{N-1} \sum_i \left(x_i - \langle x \rangle\right)^2}$$

! **rms** is the better estimator than **mean deviation**

Error estimations

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_i \left(x_i - \langle x \rangle\right)^2}$$

! In case of normally distributed data we would expect that

68% of the points will lie within $\pm 1\sigma$

95% of the points will lie within $\pm 2\sigma$

99.7% of the points will lie within $\pm 3\sigma$

! Usually we accept a variation as statistically significant only if it is more than 3σ from the mean

Error estimations

4) If we want to answer the question how reliable is our estimate of the mean we should use **“Standard deviation in the mean”:**

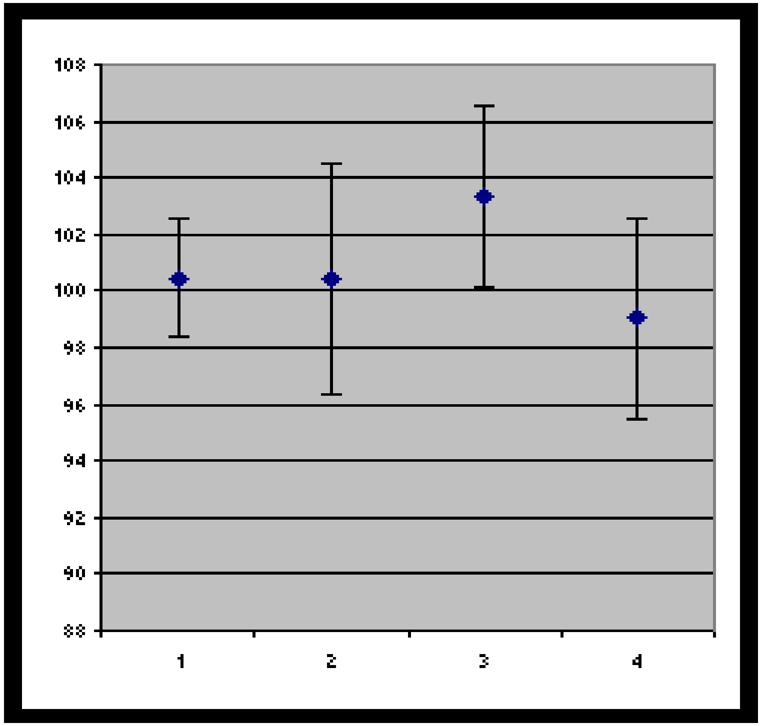
$$\sigma_{\langle x \rangle} = \frac{\sigma_x}{\sqrt{N}} = \sqrt{\frac{\sum_i \left(x_i - \langle x \rangle\right)^2}{N(N-1)}}$$

This is an estimator of quality of the mean and it reflects the improvement gained by averaging several data points.

Error estimations

for example

1	2	3	4
102.7051	96.99768	106.1652	106.7639
93.74577	87.22317	84.87374	92.7521
92.91529	102.4426	107.6497	98.48607
102.2656	112.7647	108.2898	93.23228
110.5028	111.9835	111.2649	110.5721
92.93493	117.3313	117.9858	100.0187
104.8264	78.16412	88.74805	121.3331
102.9943	97.65819	102.8124	96.12005
93.52754	110.9502	107.0592	86.2086
108.0685	89.13299	98.85192	84.94068



$$\langle x \rangle = \frac{1}{N} \sum_i x_i$$

100.4486 100.4649 103.3701 99.04276

$$\sigma_x = \sqrt{\frac{1}{N-1} \sum_i \left(x_i - \langle x \rangle \right)^2}$$

6.660202 12.92402 10.09143 11.23548

$$\sigma_{\langle x \rangle} = \frac{\sigma_x}{\sqrt{N}}$$

2.106141 4.086935 3.191189 3.552972

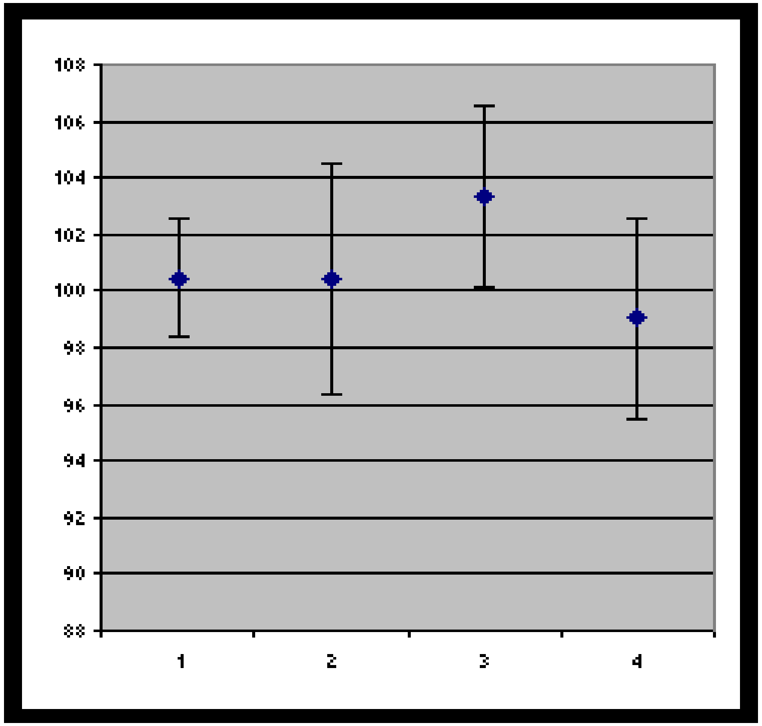
Error estimations

And finally we'll calculate the weighted Mean

$$\langle x \rangle = \frac{1}{\sum_i 1/\sigma_i^2} \sum_i \frac{x_i}{\sigma_i^2}$$

And it's variance

$$\sigma_{\langle x \rangle}^2 = \frac{1}{\sum_i 1/\sigma_i^2}$$



x_i	100.4486	100.4649	103.3701	99.04276
σ_i	2.106141	4.086935	3.191189	3.552972

$\langle x \rangle$ 100.7280

$\sigma_{\langle x \rangle}$ 1.47

Error propagation

Let consider the quantities $x_1, x_2, \ldots, x_n$ with given uncorrelated errors $\sigma_1, \sigma_2, \ldots, \sigma_n$

Now we want to calculate the error for quantity q which is depend on $x_1, x_2, \ldots, x_n$:

$$q = f(x_1, x_2, \ldots, x_n)$$

We can estimate the error of q via

$$\sigma_q^2 = \sum_i \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2$$

Error propagation

Examples. 1D case:

$$\sigma_q = \left|\frac{dq}{dx}\right| \sigma_x$$

We know the frequency ν of the photon with error σ. Now we want to calculate the energy and wavelength of the photon and their errors.

$$E = h\nu \qquad \sigma_E = \left|\frac{dE}{d\nu}\right| \sigma_\nu = h\sigma_\nu$$

$$\lambda = \frac{c}{\nu} \qquad \sigma_\lambda = \left|\frac{d\lambda}{d\nu}\right| \sigma_\nu = \frac{c}{\nu^2}\sigma_\nu$$

Error propagation

Examples. 2D case:

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2$$

We know the fluxes in two X-ray lines and their errors. How to calculate the error of ratio of the fluxes?

$$R = \frac{F_1}{F_2}$$

$$\sigma_R^2 = \left(\frac{\sigma_{F_1}}{F_2}\right)^2 + \left(\frac{\sigma_{F_2}}{F_2^2}\right)^2 F_1^2$$

$$\left(\frac{\sigma_R}{R}\right)^2 = \left(\frac{\sigma_{F_1}}{F_1}\right)^2 + \left(\frac{\sigma_{F_2}}{F_2}\right)^2 \Rightarrow \sigma_R = R\sqrt{\left(\frac{\sigma_{F_1}}{F_1}\right)^2 + \left(\frac{\sigma_{F_2}}{F_2}\right)^2}$$

Error propagation

Examples. ND case:

$$\sigma_q^2 = \sum_i \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2$$

Now we derive the **"standard deviation in the mean"** formula using the general case of uncorrelated error propagation

$$\langle x \rangle = \frac{1}{N}\sum_i x_i \qquad \frac{\partial \langle x \rangle}{\partial x_i} = \frac{1}{N} \qquad \sum_i \sigma_x^2 = N$$

$$\sigma_{\langle x \rangle}^2 = \sum_i \left(\frac{\sigma_x}{N}\right)^2 = \frac{N}{N^2}\sigma_x^2 = \frac{\sigma_x^2}{N}$$

$$\sigma_{\langle x \rangle} = \frac{\sigma_x}{\sqrt{N}}$$

Poisson distribution

I. The Poisson distribution arises when we observe independent random events that are occurring at a constant rate, such the expected number of events is $\lambda>0$.

The Poisson probability for obtaining k (integer positive or zero) such events in the given interval is

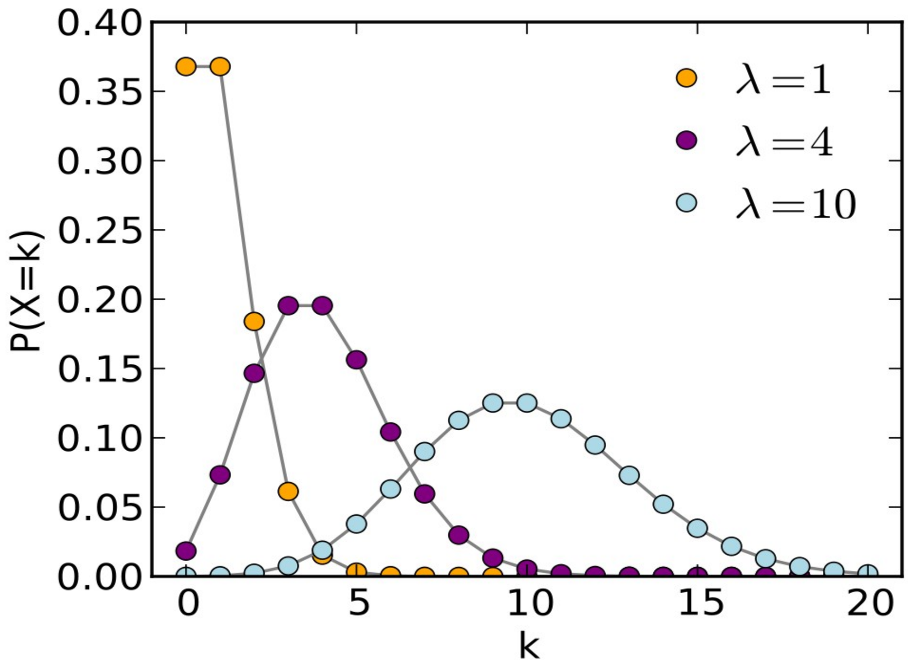
$$p(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Poisson distribution

For example: the flux of cosmic rays reaching the Earth is about 1 per cm^2.

If we have a detector with 20 cm^2 area we might expect on average 3.3 cosmic rays for each 10 seconds intervals.

But in actual experiment we would never clearly observe 3.3. But numbers like 3, 4 or sometimes 1 or 5 or 7.



We could produce a histogram showing how many times we observed exactly k cosmic rays in our 10 second interval. If we divide such histogram by total numbers of time intervals we would derive the Poisson probability distribution for the cosmic rays.

Poisson distribution

! If we have an observational data which is distributed by Poisson distribution (e.g. number of photons in the given interval of time) then we can expect following statistical characteristics:

Mean: $$\langle x \rangle = \lambda$$

Deviation: $$\langle \sigma_x \rangle = \sqrt{\lambda}$$

Poisson distribution

For example: we have in mean *N* photons per second. Statistical variability of this flux would be lie $\pm\sqrt{N}$ interval

We can also calculate **the signal-to-noise ratio** (shot noise) using these relations:

$$\mathrm{SNR} = \frac{\lambda}{\sigma} = \frac{N}{\sqrt{N}} = \sqrt{N}$$

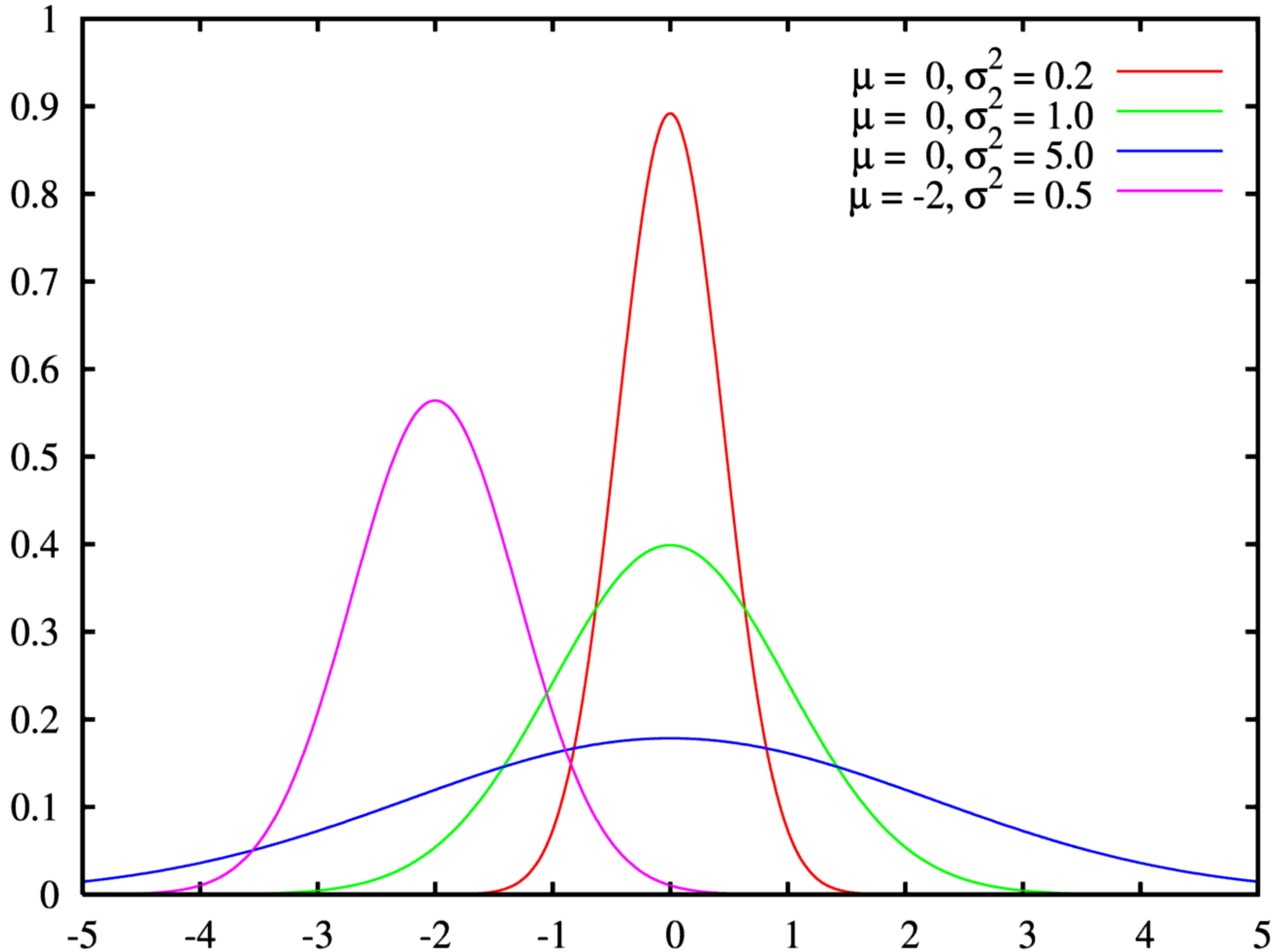
Normal distribution

II. A **normal distribution** in a variate $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and variance σ^2 is a statistic distribution with probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

It is called also **Gaussian distribution** or **“bell curve”** because of the shape

Normal distribution



Normal distribution

! Random variates with unknown distributions are often assumed to be normal, especially in physics and astronomy.

it is often a good approximation due to a result known as the central limit theorem.

This theorem states that the mean of any set of variates with any distribution having a finite mean and variance tends to the normal distribution.

Normal distribution

! The formulas for the mean and standard deviation

$$\langle x \rangle = \frac{1}{N}\sum_i x_i \qquad \sigma_x = \sqrt{\frac{1}{N-1}\sum_i \left(x_i - \langle x \rangle\right)^2}$$

supposes that the data is normally distributed

Normal distribution

- In case of normally distributed data we would expect that

  68% of the points will lie within $\pm 1\sigma$

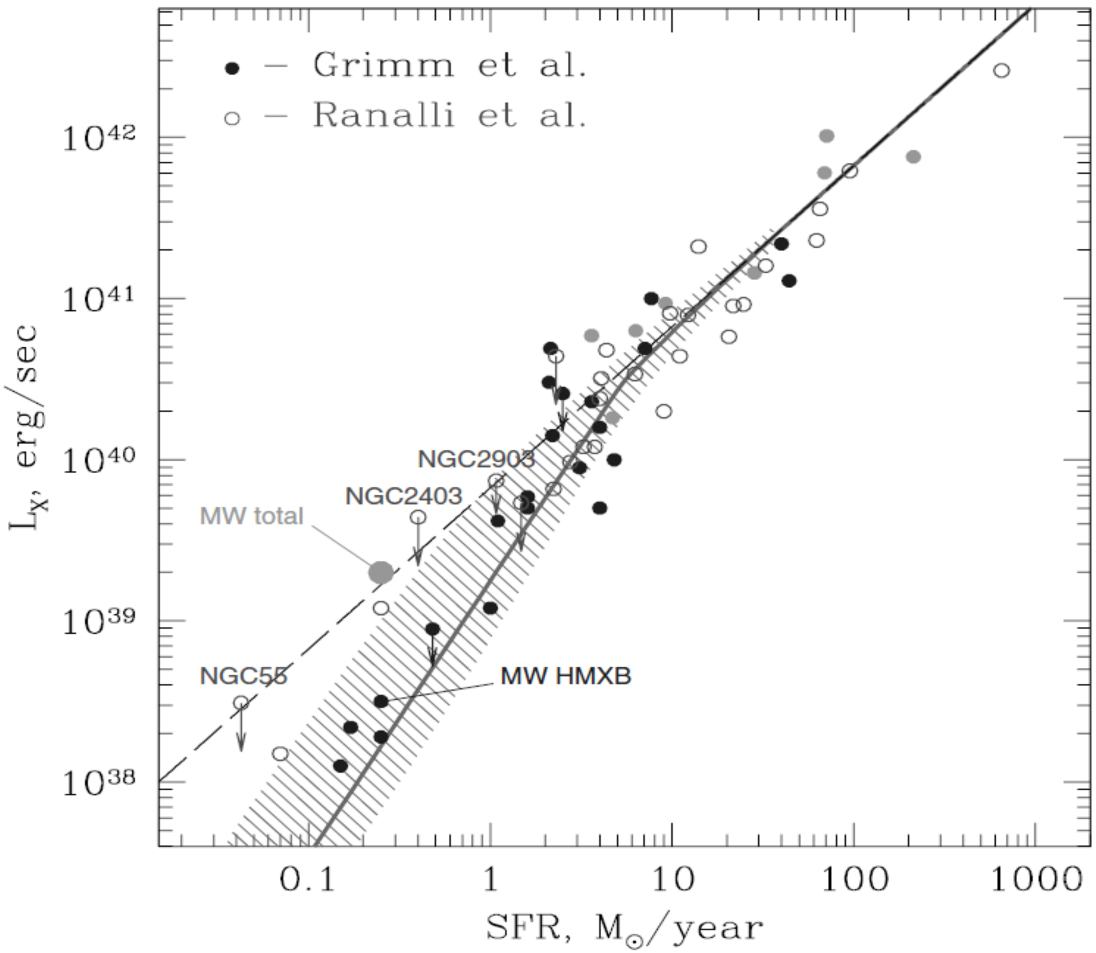
  95% of the points will lie within $\pm 2\sigma$

  99.7% of the points will lie within $\pm 3\sigma$

- Usually we accept a variation as statistically significant only if it is more than 3σ from the mean

Connection between distributions

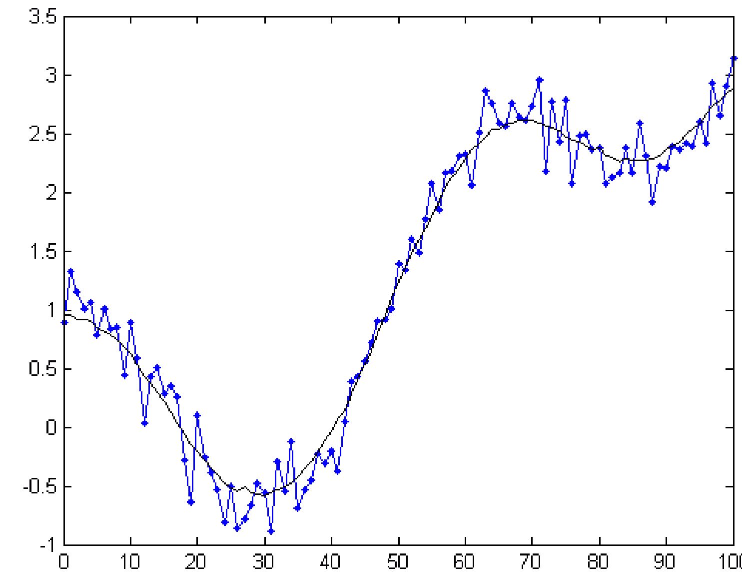
- Poisson distribution becomes normal when $\lambda \to \infty$



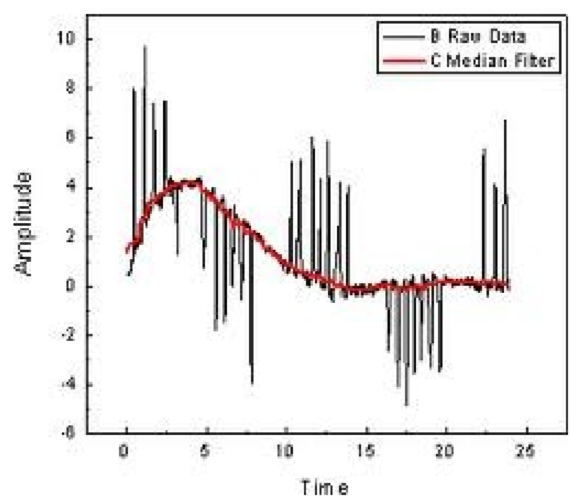
Filtering the data

I. Moving average

$$y[i] = \frac{1}{N} \sum_{j=0}^{N-1} x[i-j]$$

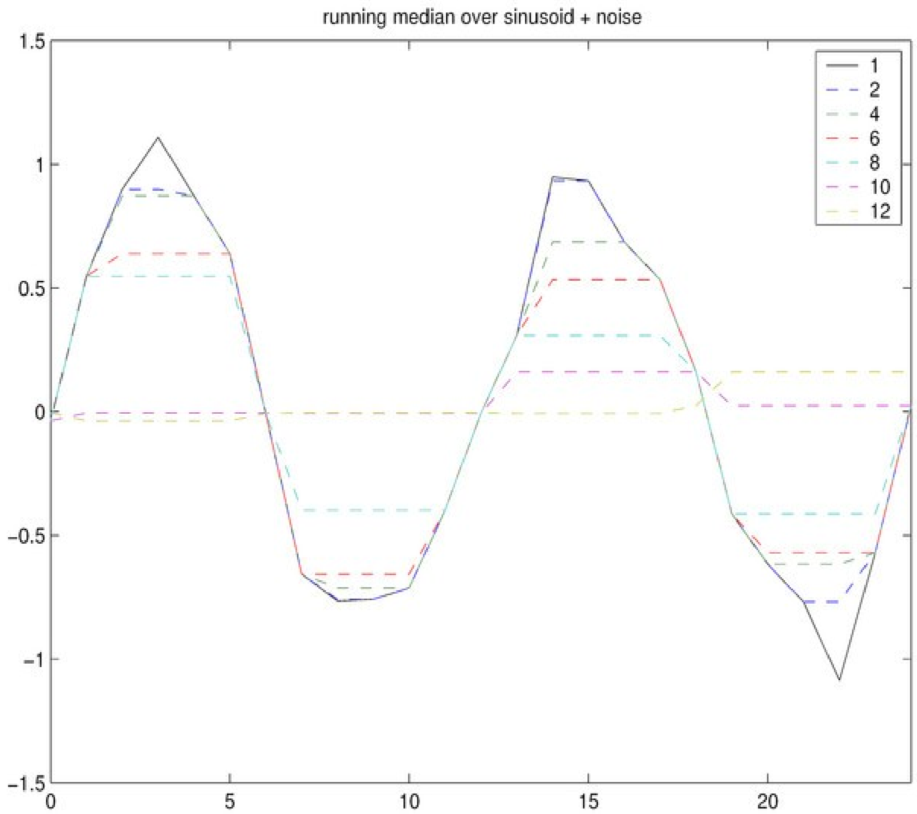


II. Median



Median and mean

Median over sinusoid



Moving average over sinusoid