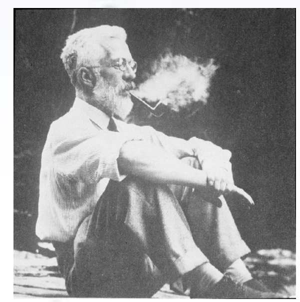
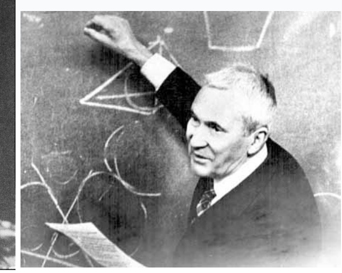
Applied Statistics

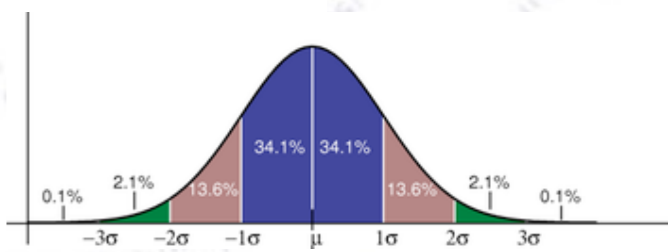
Mean and Width





Troels C. Petersen (NBI)



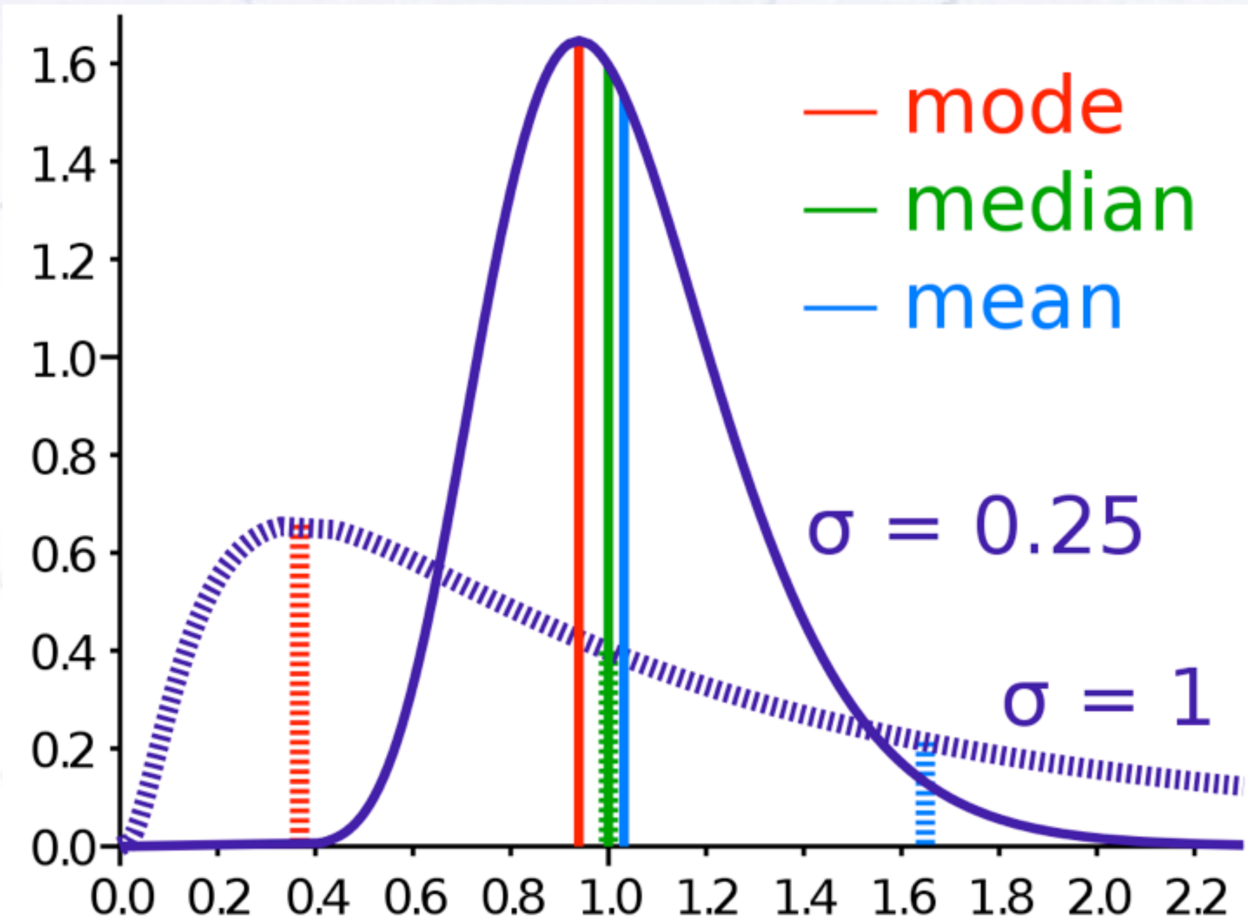
"Statistics is merely a quantisation of common sense"

Defining the mean

There are several ways of defining "a typical" value from a dataset:

a) Arithmetic mean   b) Mode (most probably)   c) Median (half below, half above)

d) Geometric mean   e) Harmonic mean   f) Truncated mean (robustness)



Mean and Width

It turns out, that the best estimator for the **mean** is (as you all know):

$$\hat{\mu} = \frac{1}{N}\sum_i x_i = \bar{x}$$

For the **width** of the distribution (a.k.a. **standard deviation** or **RMS**) it is:

$$\hat{\sigma} = \sqrt{\frac{1}{N}\sum_i (x_i - \mu)^2}$$

Note the “hat”, which means “estimator”. It is sometimes dropped...

Mean and Width

It turns out, that the best estimator for the **mean** is (as you all know):

$$\hat{\mu} = \frac{1}{N} \sum_i x_i = \bar{x}$$

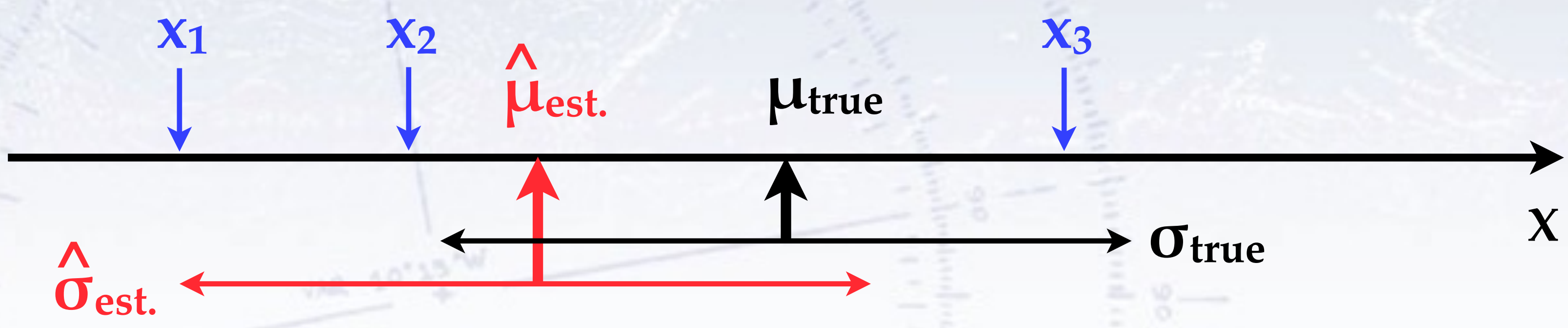
For the **width** of the distribution (a.k.a. **standard deviation** or **RMS**) it is:

$$\hat{s} = \sqrt{\frac{1}{N-1} \sum_i (x_i - \bar{x})^2}$$

Note the “hat”, which means “estimator”. It is sometimes dropped...

Why not "just" the naive RMS?

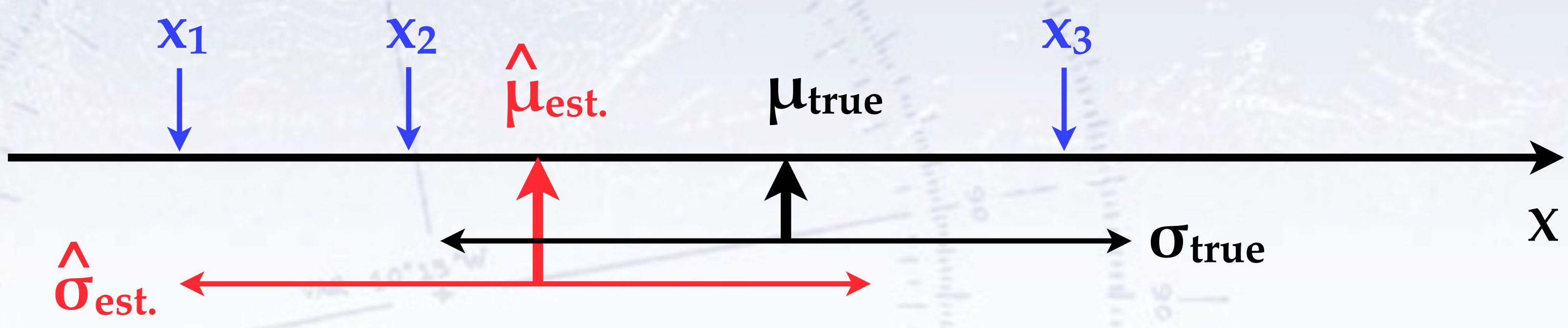
Imagine taking 3 independent measurements, and then the mean and RMS:



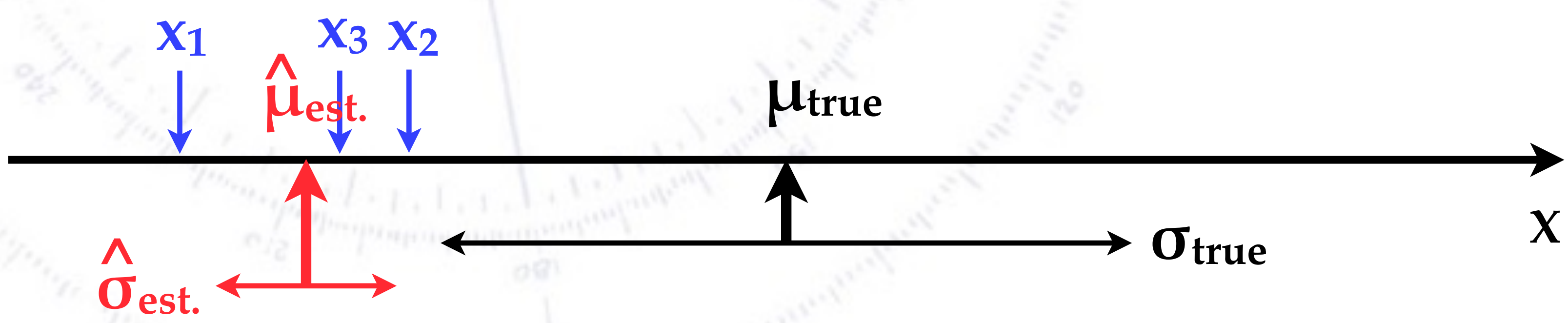
Above, all went well, because measurements were nicely distributed on both sides of the mean, and spread out according to RMS.

Why not "just" the naive RMS?

Imagine taking 3 independent measurements, and then the mean and RMS:



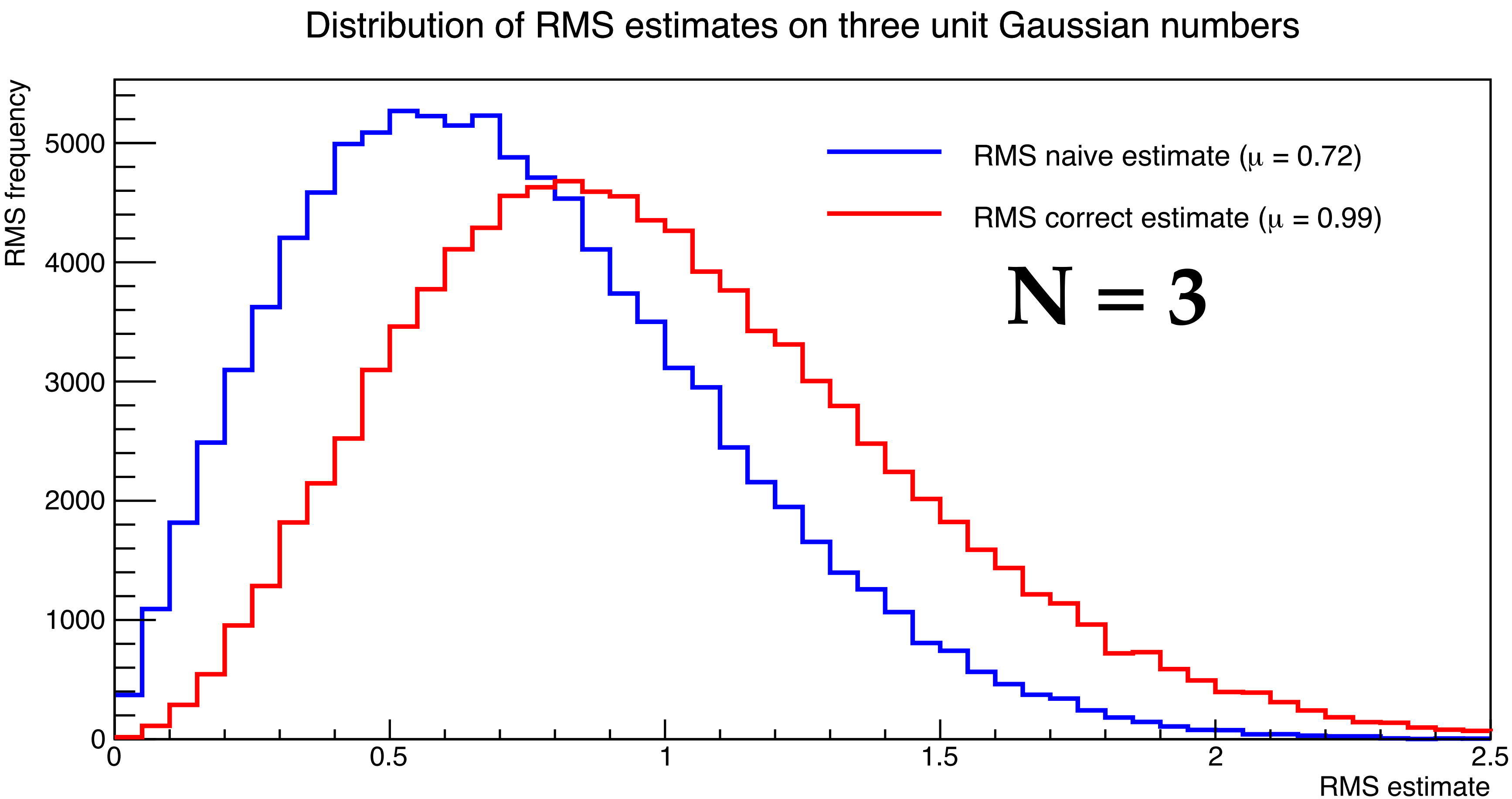
Above, all went well, because measurements were nicely distributed on both sides of the mean, and spread out according to RMS.



However, now the mean is off (not terribly so) and the RMS way off (terribly so!). If we had used the true mean in the formula, it would not have been a problem.

How incorrect is the naive RMS?

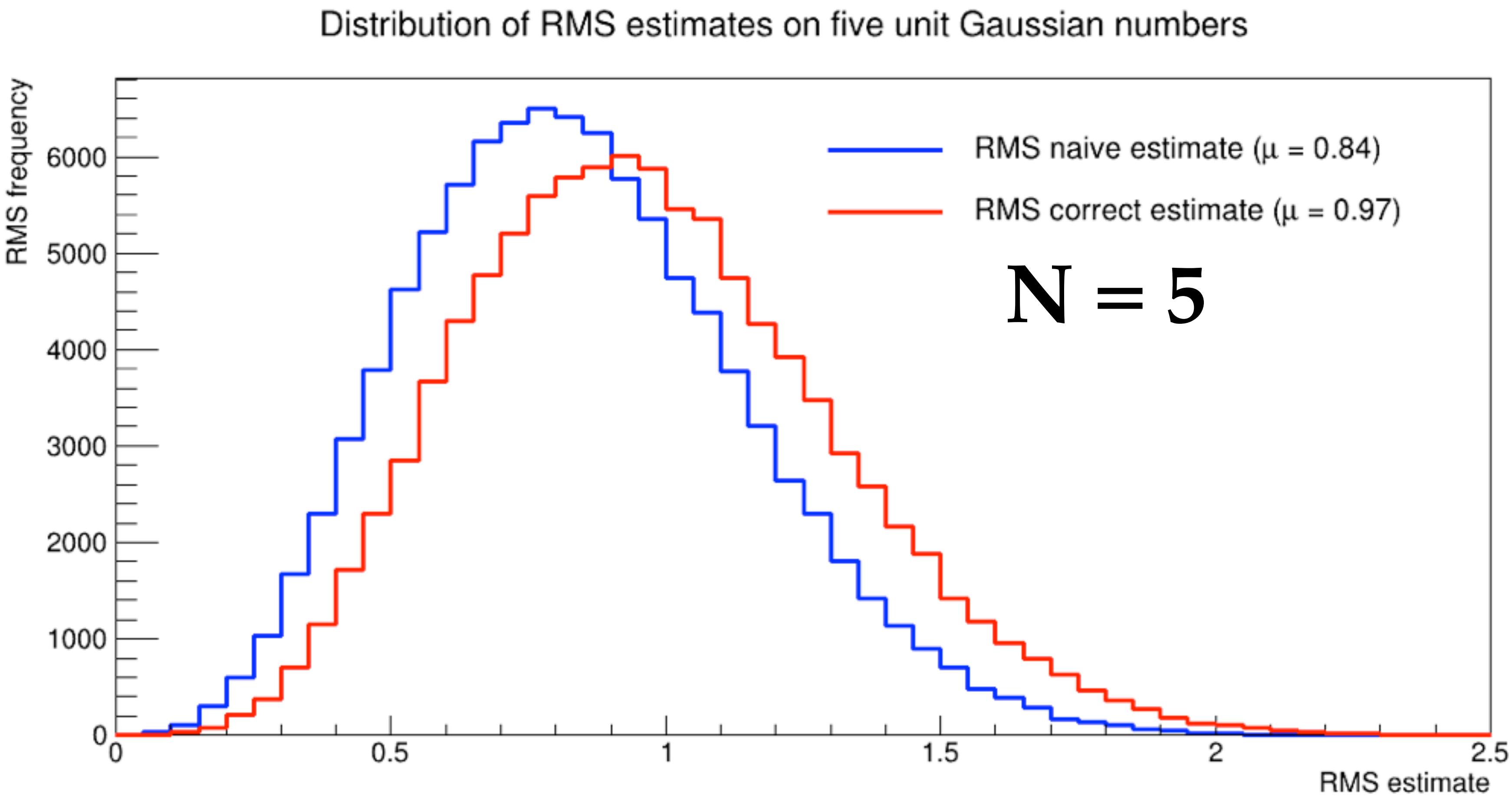
Such questions can most easily be answered by a small simulation…
Produce N=5 numbers from a unit Gaussian, and calculate the RMS estimate:



So, the "naive" RMS underestimates the uncertainty a bit...

How incorrect is the naive RMS?

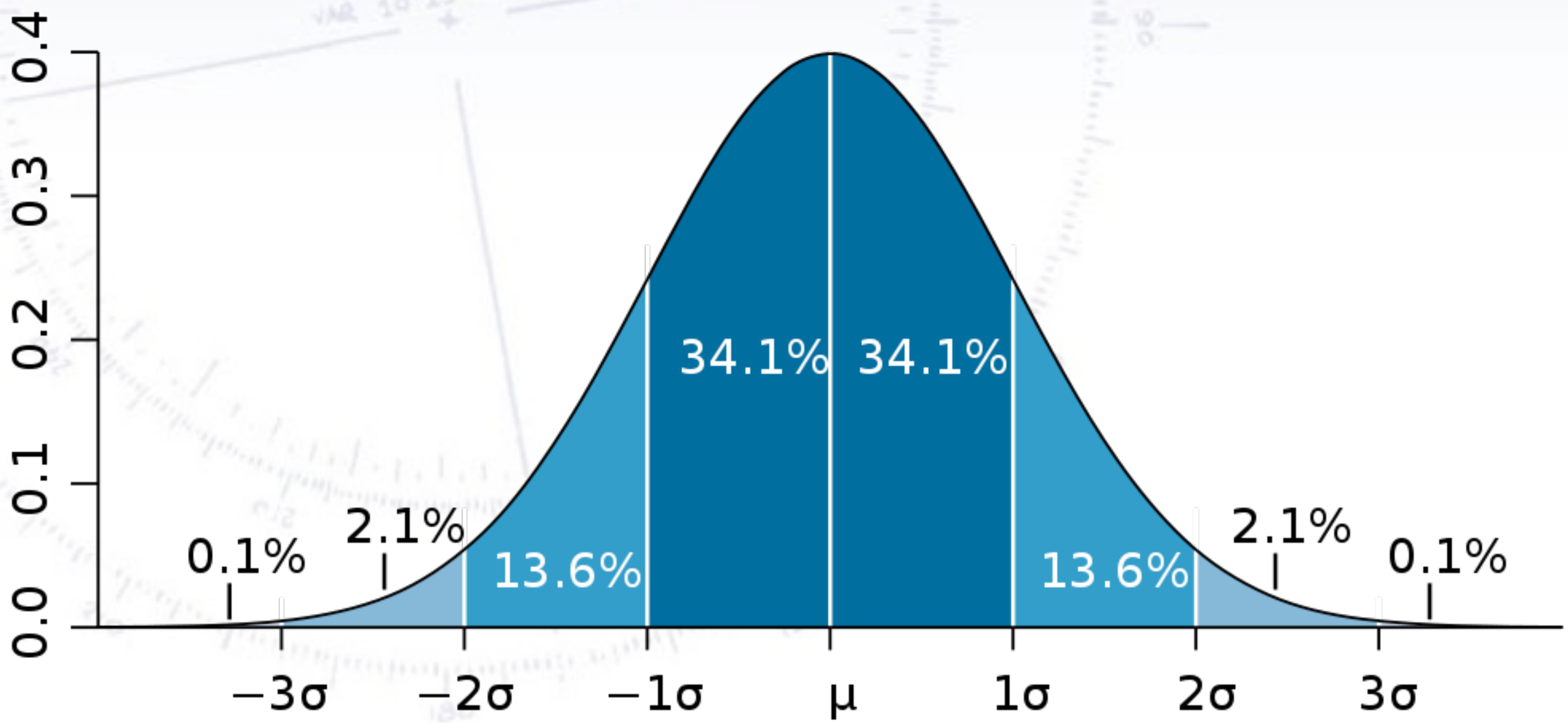
Such questions can most easily be answered by a small simulation…
Produce N=5 numbers from a unit Gaussian, and calculate the RMS estimate:



So, the "naive" RMS underestimates the uncertainty a bit...

Relation between RMS and Gaussian width...

When a distribution is Gaussian, **the RMS corresponds to the Gaussian width σ**:



Mean and Width

What is the **uncertainty on the mean?** And how quickly does it improve with more data?

Mean and Width

What is the **uncertainty on the mean?** And how quickly does it improve with more data?

$$\hat{\sigma}_\mu = \hat{\sigma}/\sqrt{N}$$

Mean and Width

What is the **uncertainty on the mean?** And how quickly does it improve with more data?

$$\hat{\sigma}_{\mu} = \hat{\sigma}/\sqrt{N}$$

<u>Example:</u>

Cavendish Experiment

(measurement of Earth's density)

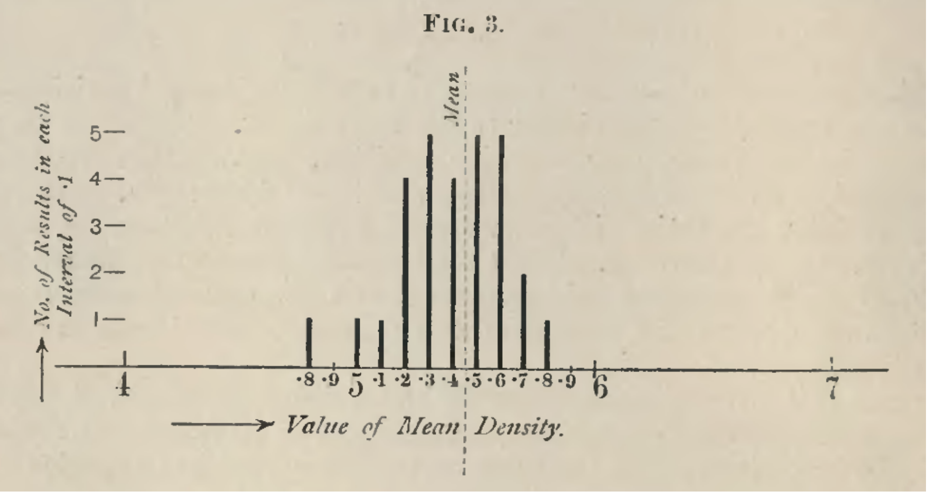
N = 29

mu = 5.42

sigma = 0.333

sigma(mu) = 0.06

Earth density = 5.42 ± 0.06



Mean and Width

What is the **uncertainty on the mean?** And how quickly does it improve with more data?

Please, commit to memory now!

$$\hat{\sigma}_\mu = \hat{\sigma} / \sqrt{N}$$

<u>Example:</u>

Cavendish Experiment

(measurement of Earth's density)

N = 29

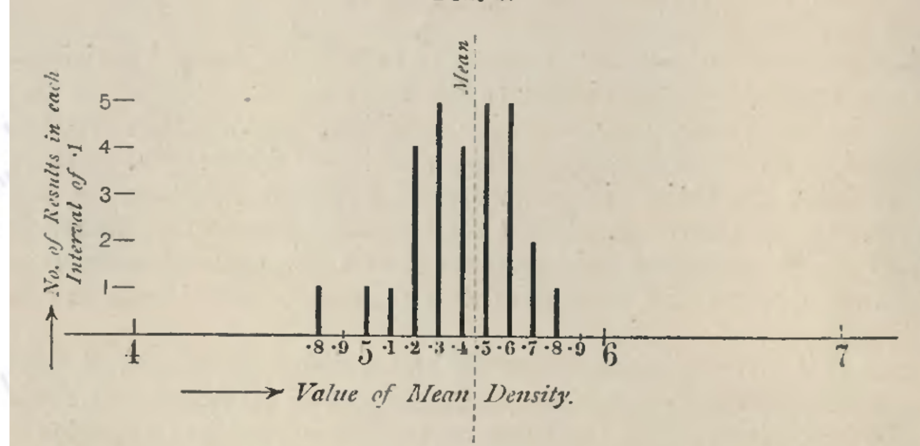
mu = 5.42

sigma = 0.333

sigma(mu) = 0.06

Earth density = 5.42 ± 0.06

Fig. 3.



Weighted Mean

What if we are given data, which has different uncertainties?
How to average these, and what is the uncertainty on the average?

$$\hat{\mu} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}$$

For measurements with varying uncertainty, there is no meaningful RMS!
The uncertainty on the mean is:

$$\hat{\sigma}_\mu = \sqrt{\frac{1}{\sum 1/\sigma_i^2}}$$

Can be understood intuitively, if two persons combine 1 vs. 4 measurements

Weighted Mean

What if we are given data, which has different uncertainties?
How to average these, and what is the uncertainty on the average?

$$\hat{\ } \qquad \sum x_i/\sigma_i^2$$

Note that when doing a weighted mean, one should check if the measurements agree with each other!
This can be done with a ChiSquare test.

For measur ... RMS!
The uncerta ...

$$\hat{\sigma}_\mu = \sqrt{\frac{1}{\sum 1/\sigma_i^2}}$$

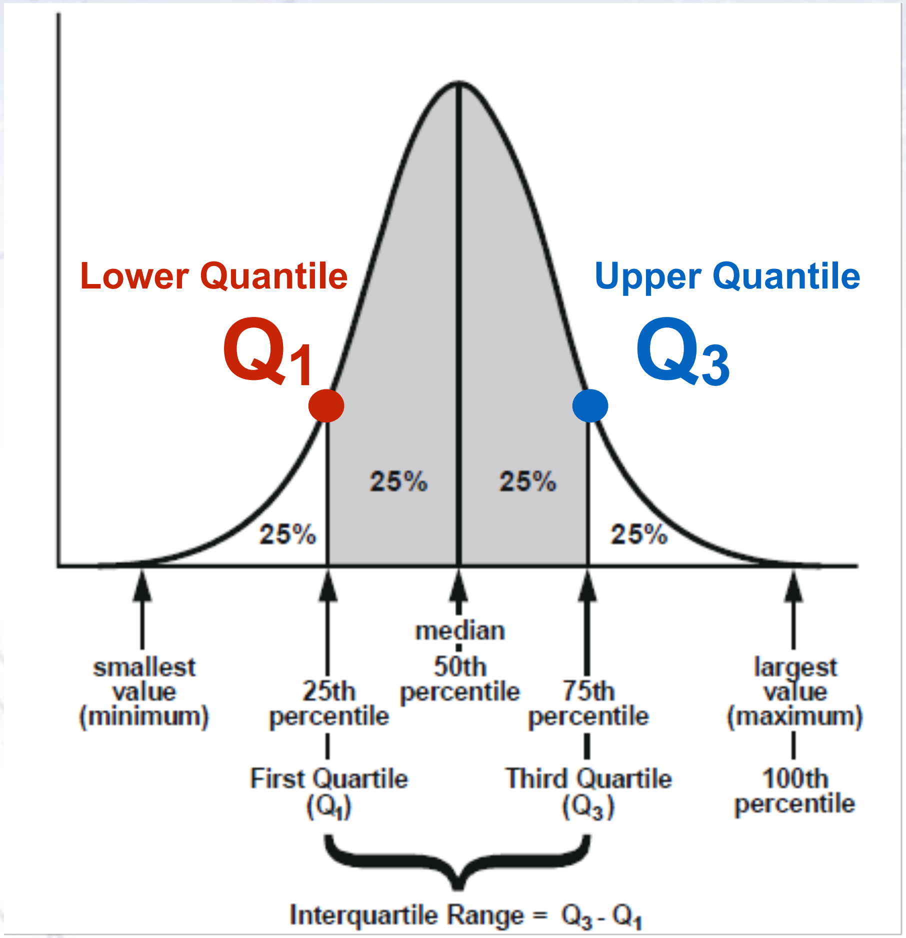
Can be understood intuitively, if two persons combine 1 vs. 4 measurements

Resolution using InterQuantile Range

A useful measure of resolution is the InterQuantile Range (IQR), as this is not affected by long tails.

IQR measures **statistical dispersion**, calculated as the difference

$$\mathrm{IQR} = Q_3 - Q_1$$



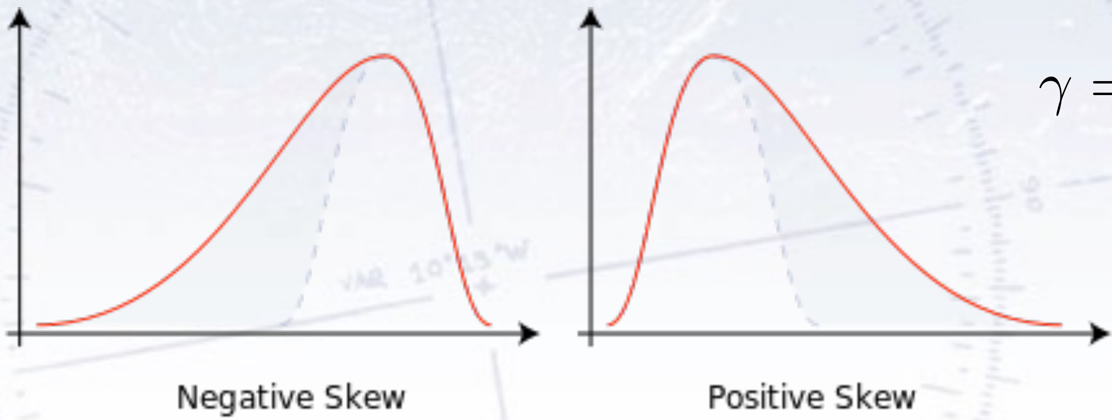
The InterQuantile Efficiency (IQE) is defined as:

$$\mathrm{IQE} = \mathrm{IQR} / 1.349$$

The factor $1.349 = 2\,\Phi^{-1}(0.75)$ ensures that IQR = 1 for a unit Gaussian.

Skewness and Kurtosis

Higher moments reveal something about a distributions asymmetry and tails:



$$\gamma = \frac{\frac{1}{N}\sum_i (x_i - \bar{x})^3}{(\frac{1}{N}\sum_i (x_i - \bar{x})^2)^3/2}$$

$$\kappa = \frac{\frac{1}{N}\sum_i (x_i - \bar{x})^4}{(\frac{1}{N}\sum_i (x_i - \bar{x})^2)^2} - 3$$