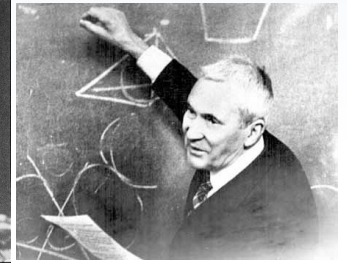
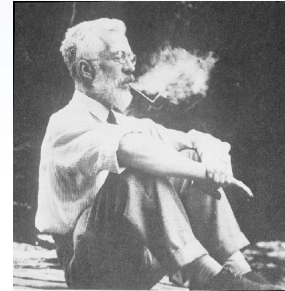
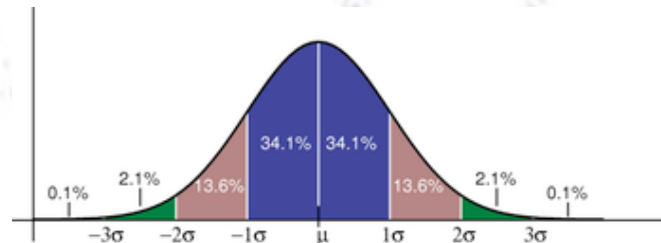


# Applied Statistics

## Mean and Width



Troels C. Petersen (NBI)



*"Statistics is merely a quantisation of common sense"*

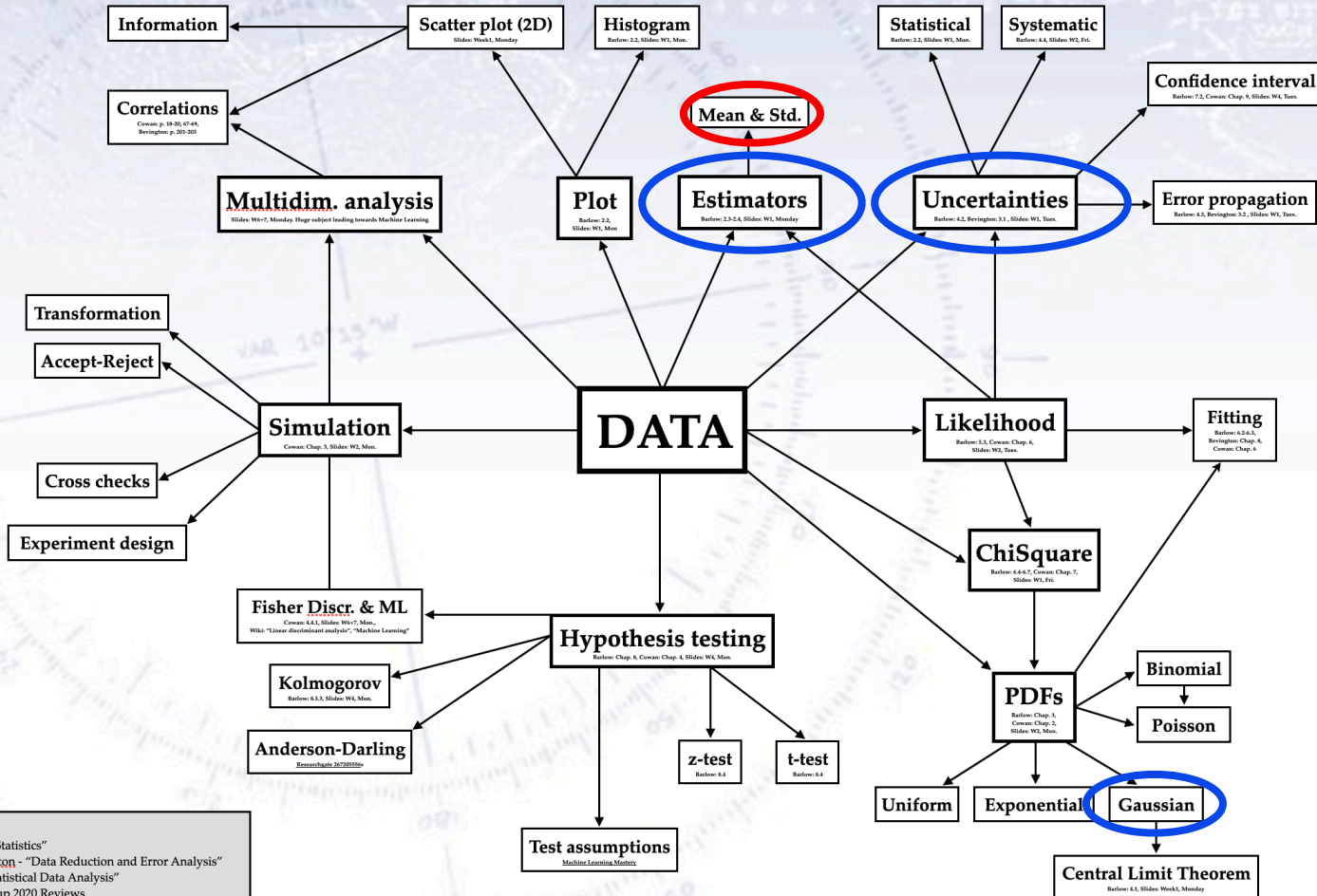
# Mean & Width

## Applied Statistics

## Describe data (Quantify & Visualise)

## Overview of subjects

Version 1.2, 6. Nov. 2020



Simulate data (Design & Cross Check)

Model data (Predict & Understand)

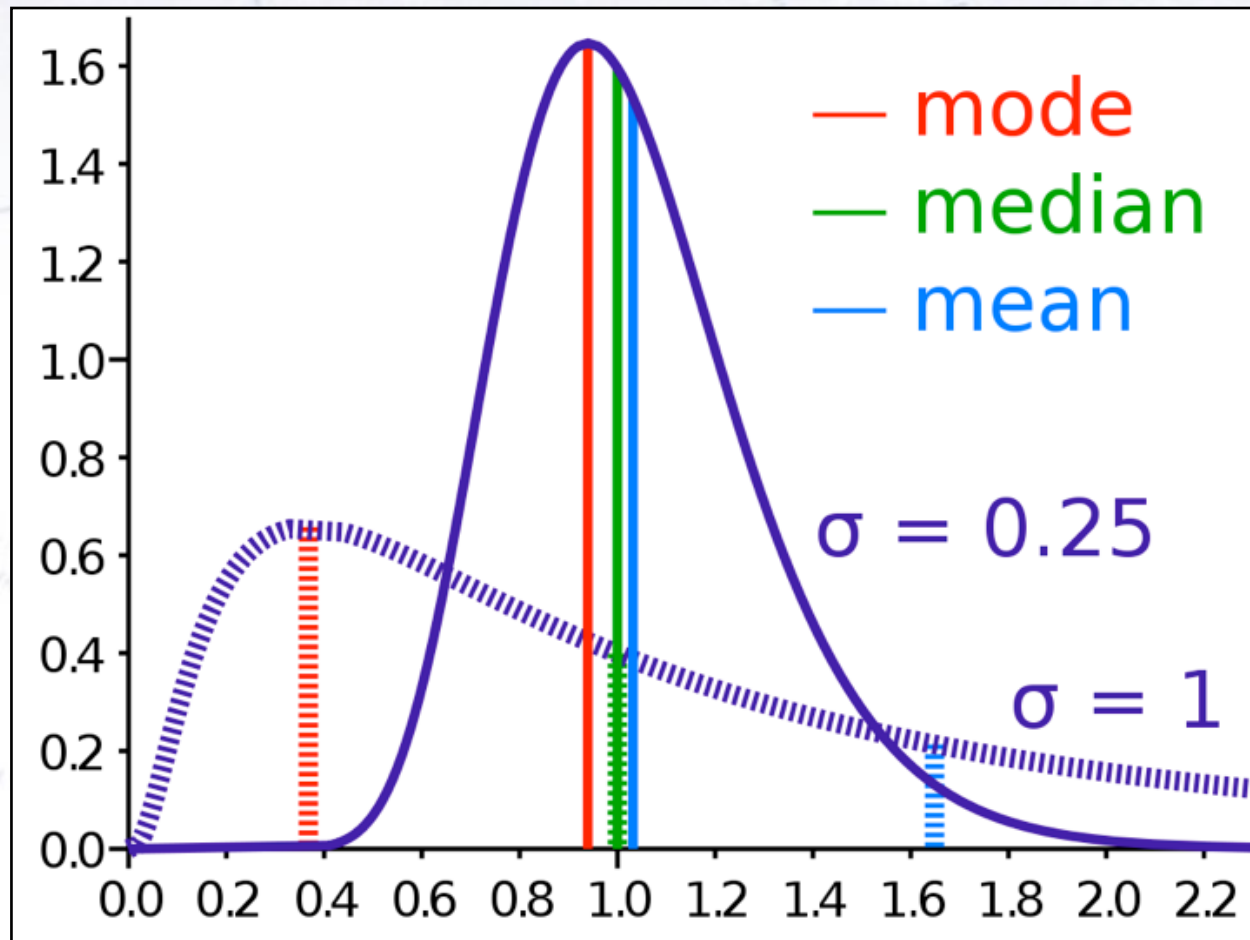
References:  
 Barlow: R. J. Barlow - "Statistics"  
 Bevington: P. H. Bevington - "Data Reduction and Error Analysis"  
 Cowan: G. Cowan - "Statistical Data Analysis"  
 PDG: Particle Data Group 2020 Reviews  
 Slides: T. C. Petersen - "Applied Statistics 2020" course (W = Week)  
 Wiki: Good reference for ALL subjects (only specified when essential)  
 SciPy: SciPy Statistical Functions and (very brief) documentation

## Test hypotheses on data (Decide)

# 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):

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

The second (central) moment of the data is called the **variance**, defined as:

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

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 **standard deviation (Std)**, a.k.a. **width** or **RMSE**, 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):

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

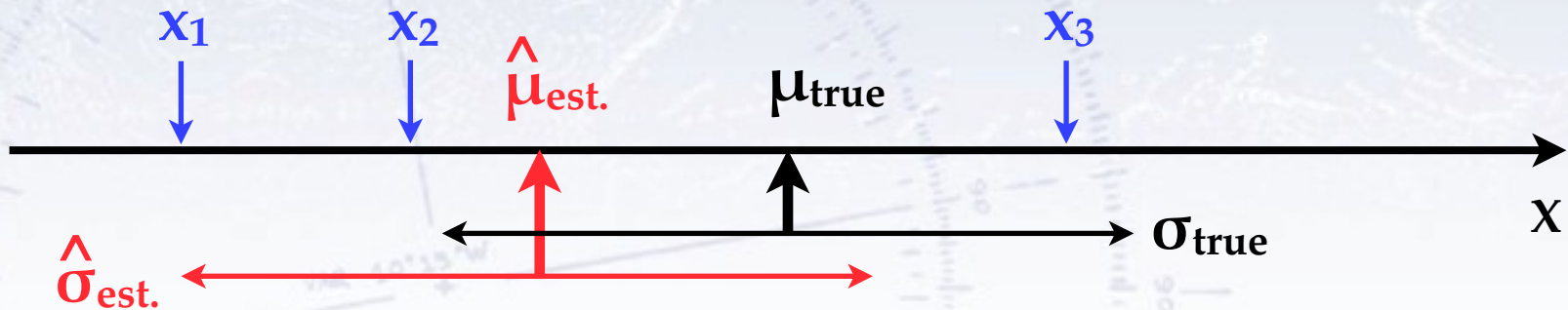
For the **standard deviation (Std)**, a.k.a. **width** or **RMSE**, it is:

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

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

# Why not “just” the naive SD?

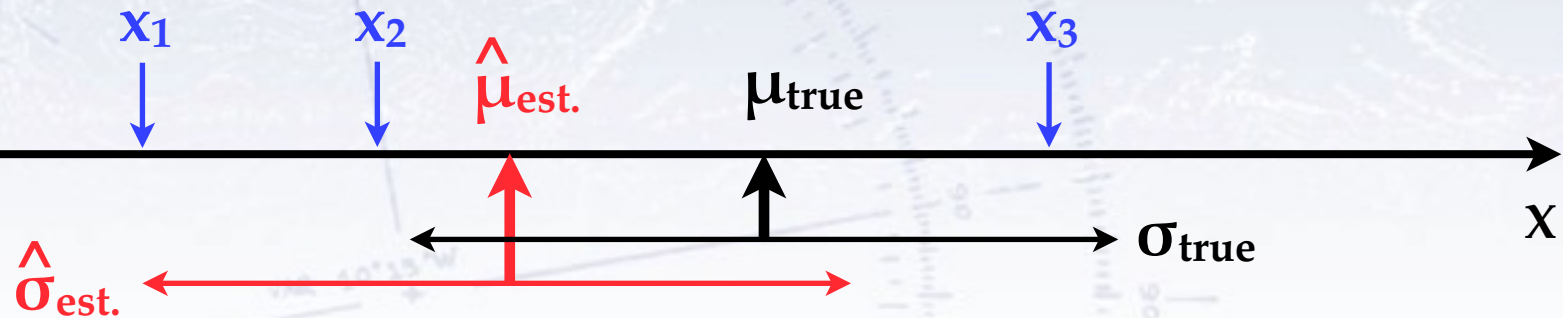
Imagine taking 3 independent measurements, and then the mean and SD:



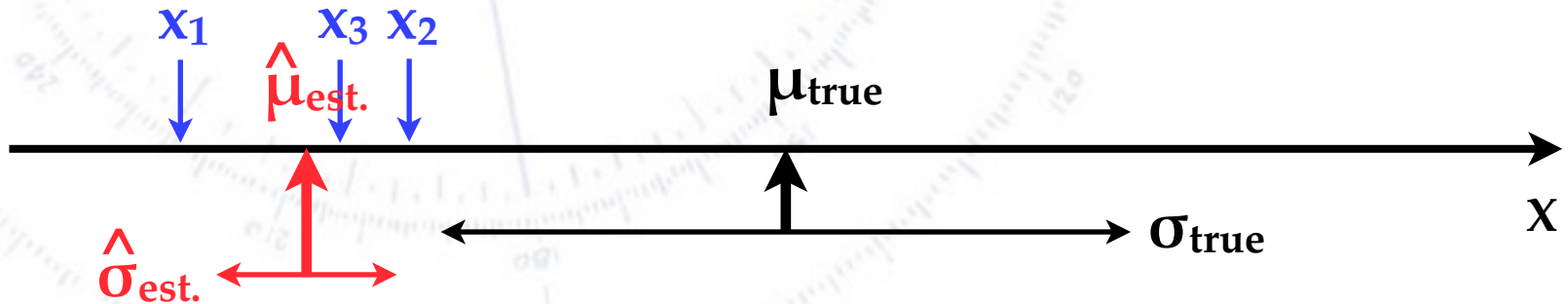
Above, all went well, because measurements were nicely distributed on both sides of the mean, and spread out according to SD.

# Why not “just” the naive SD?

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



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



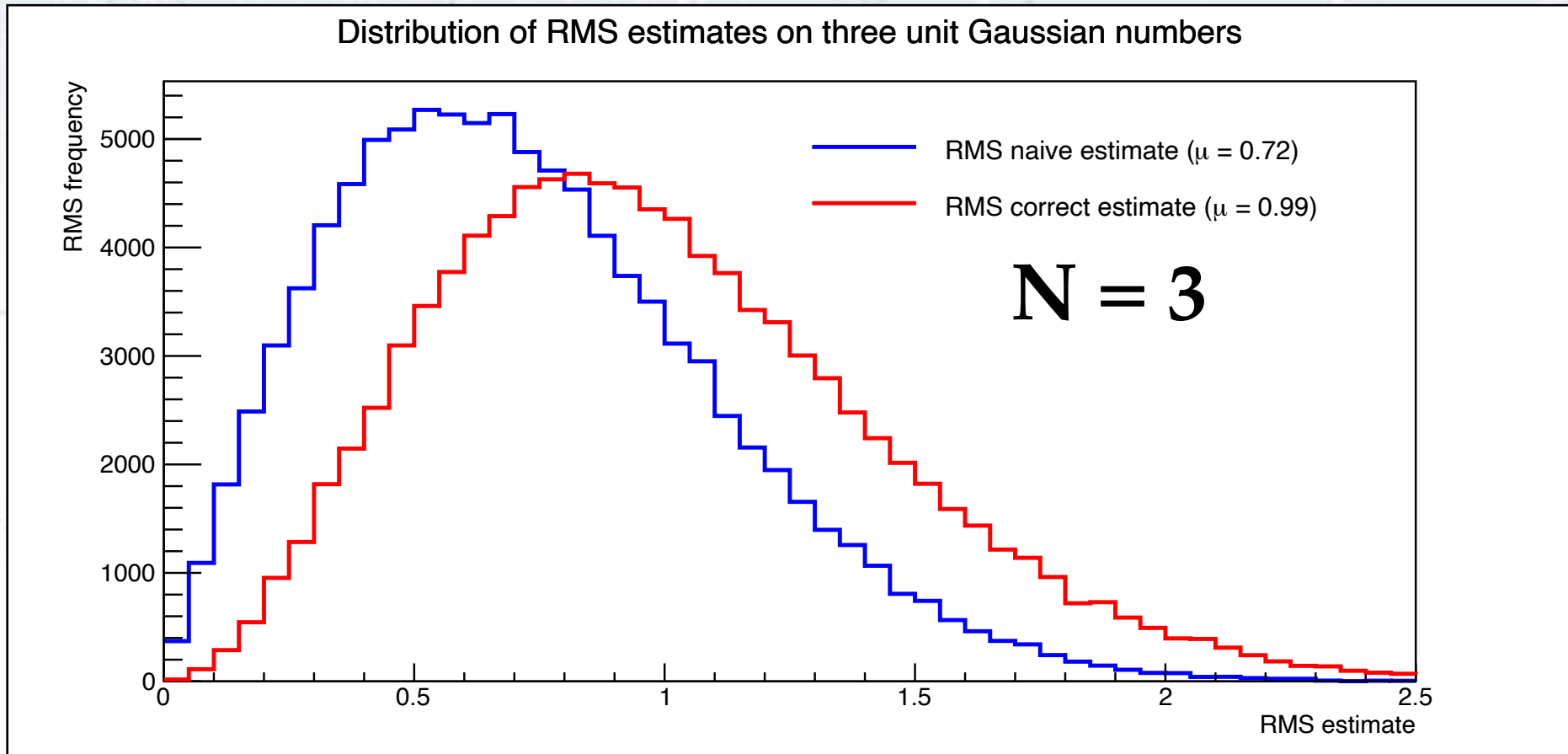
However, now the mean is off (not terribly so) and the SD way off (terribly so!). If we had used the true mean in the formula, it would not have been a problem.



# How incorrect is the naive SD?

Such questions can most easily be answered by a small simulation...

Produce  $N=3$  numbers from a unit Gaussian, and calculate the SD estimate:

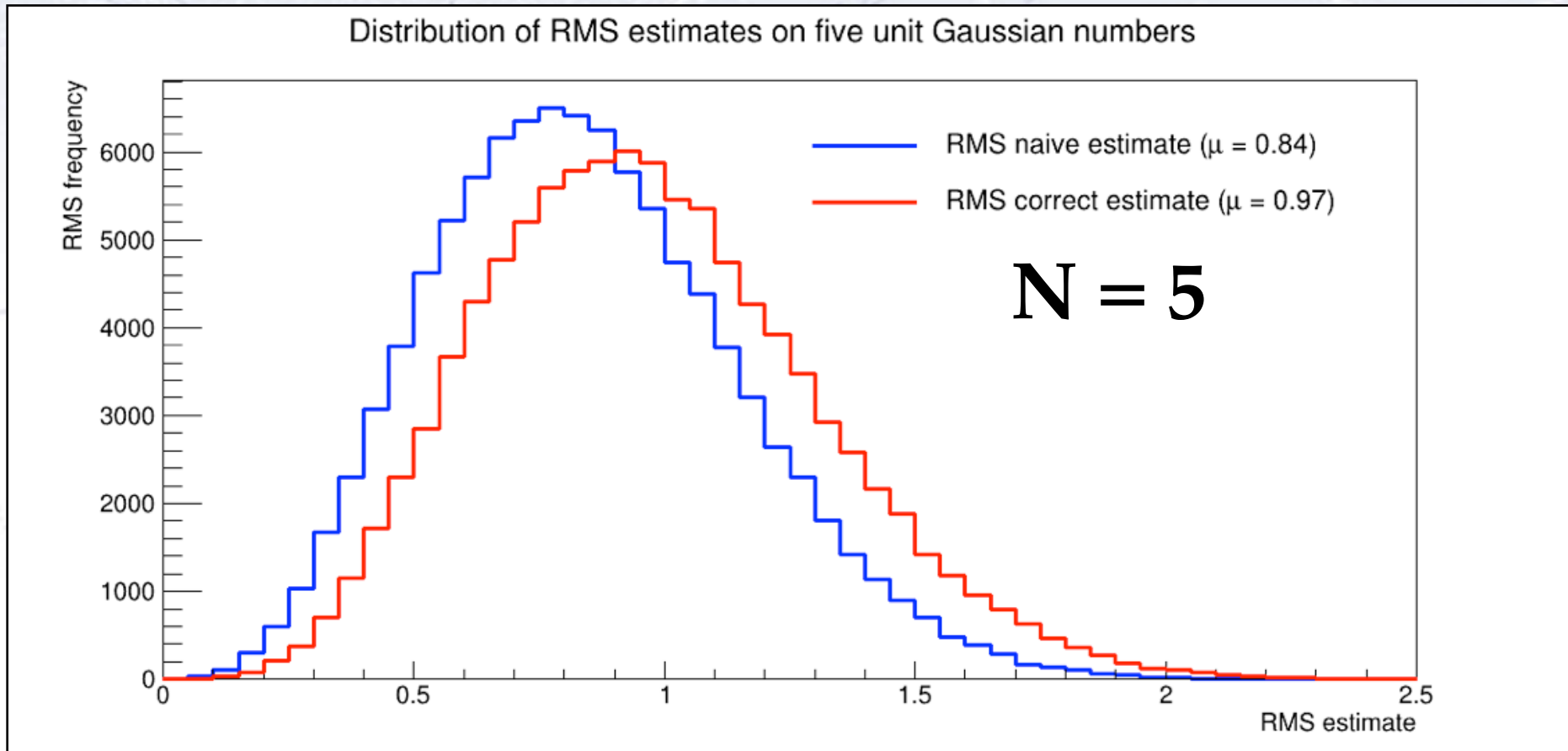


So, the “naive” SD underestimates the uncertainty significantly...

# How incorrect is the naive SD?

Such questions can most easily be answered by a small simulation...

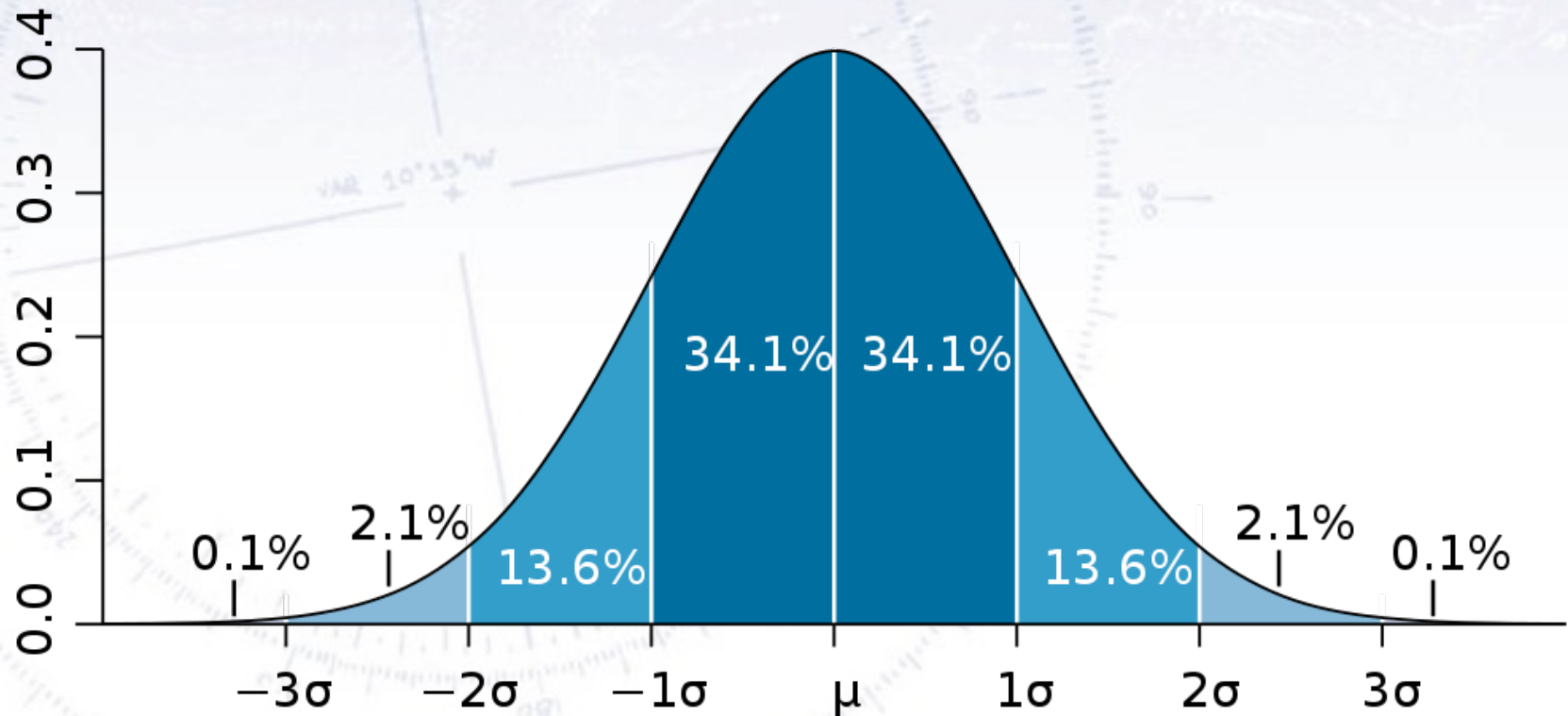
Produce  $N=5$  numbers from a unit Gaussian, and calculate the SD estimate:



Here, the “naive” SD underestimates the uncertainty a bit...

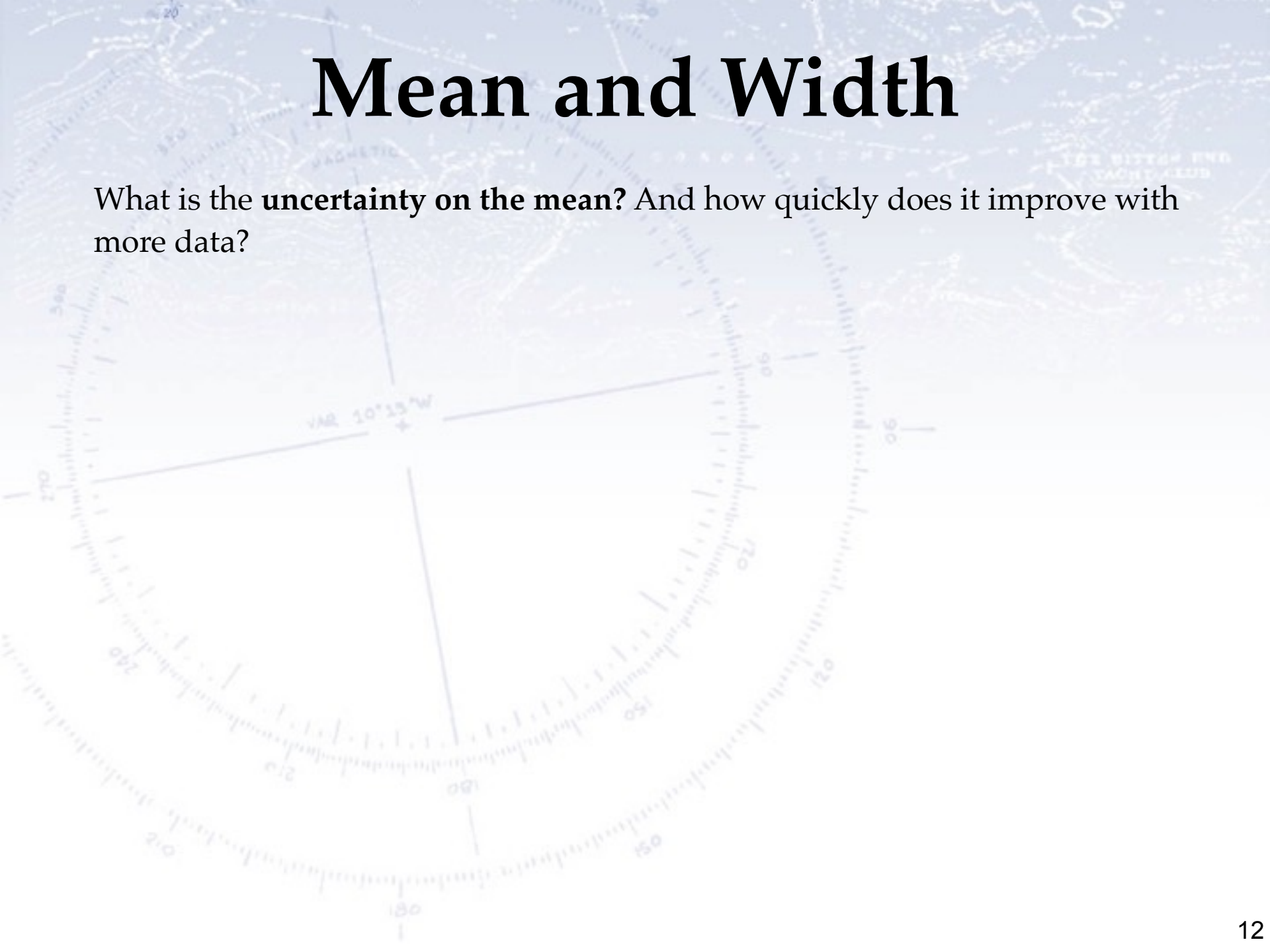
# SD and Gaussian $\sigma$ relation

When a distribution is Gaussian, the Std. corresponds to the Gaussian width  $\sigma$ :



# 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?

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

Example:

**Cavendish Experiment**  
(measurement of Earth's density)

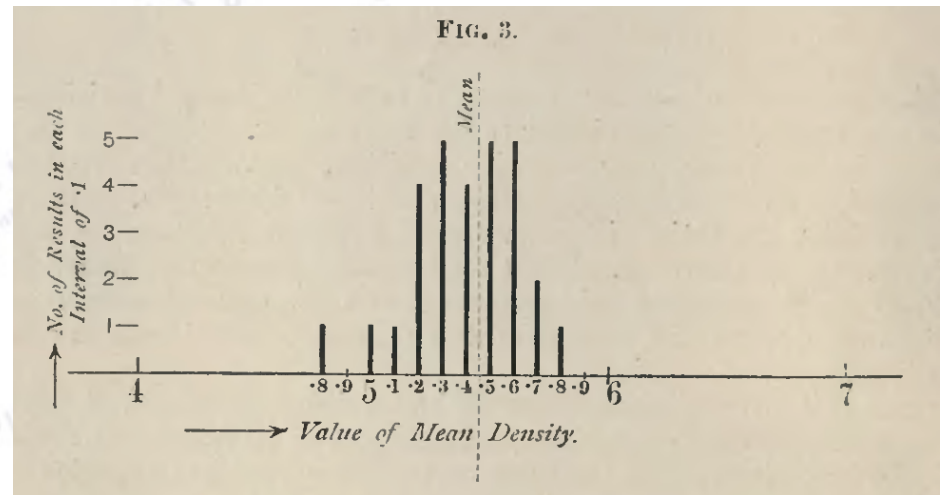
$N = 29$

$\mu = 5.42$

$\sigma = 0.333$

$\sigma(\mu) = 0.06$

**Earth density =  $5.42 \pm 0.06$**



# Mean and Width

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

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

Example:

Cavendish Experiment

(measurement of Earth's density)

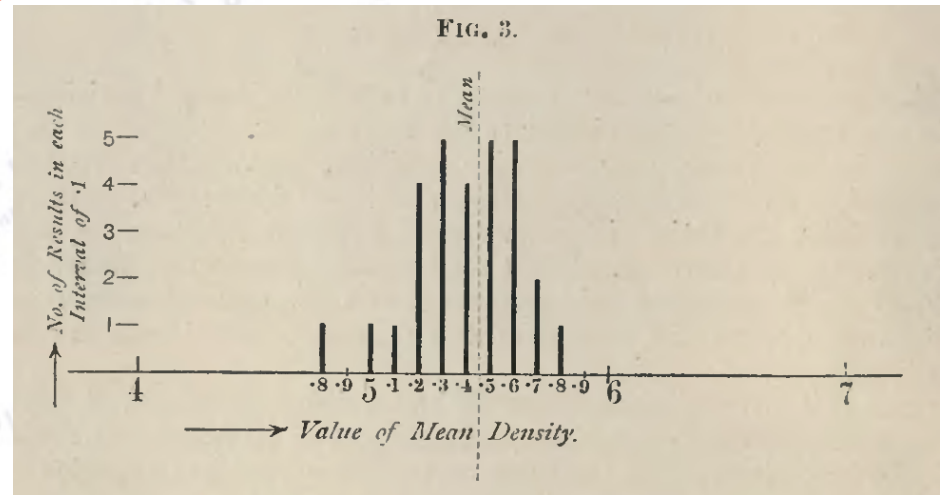
$N = 29$

$\mu = 5.42$

$\sigma = 0.333$

$\sigma(\mu) = 0.06$

Earth density =  $5.42 \pm 0.06$



# 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 SD!

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

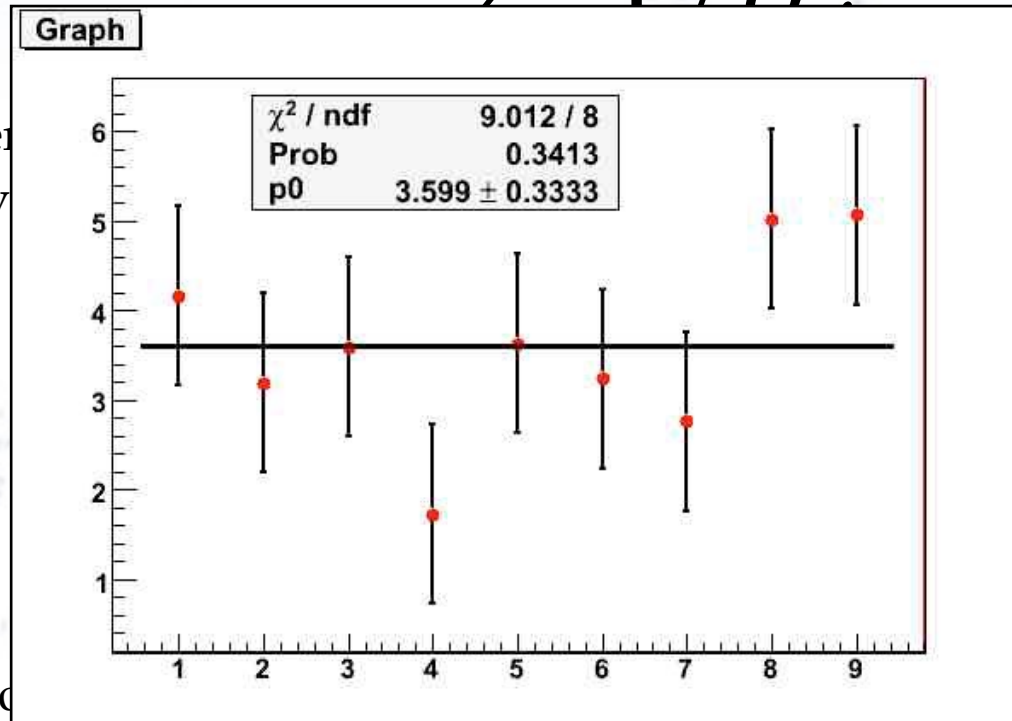
What if we are given data which has different uncertainties?

How to average?

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 measurements with different SD!  
The uncertainty



For measurements with different SD!

Can be understood

measurements

# Resolution using InterQuantile Range

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

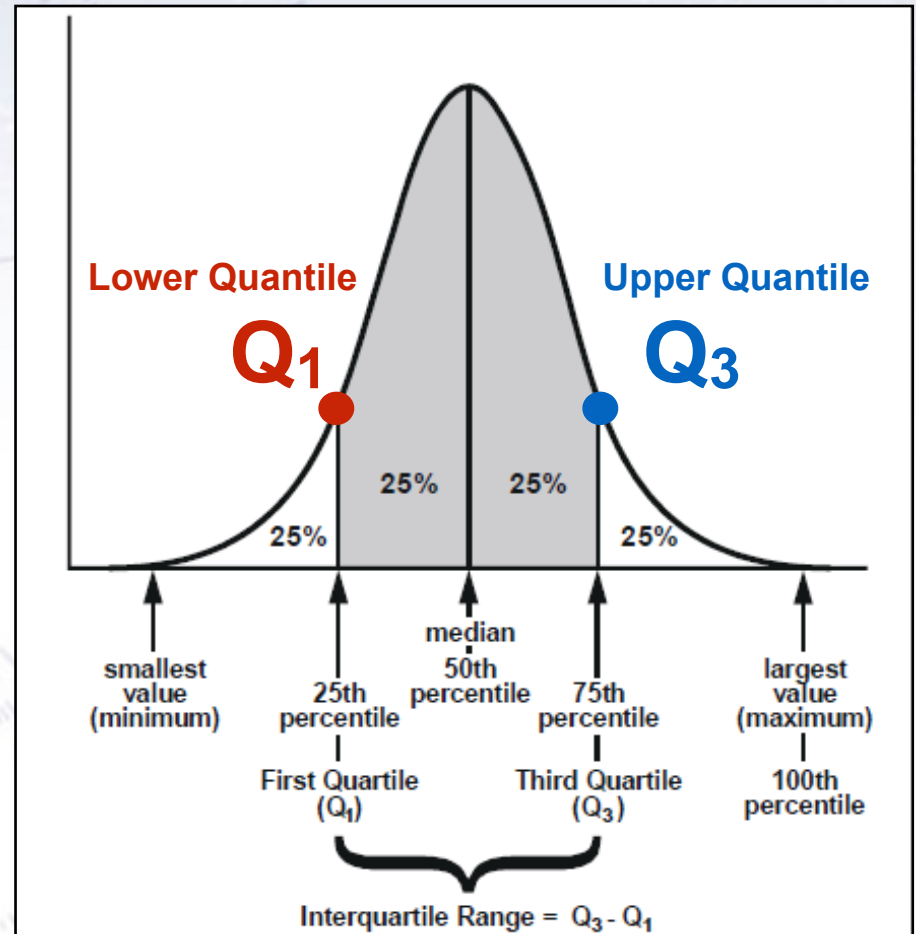
IQR measures **statistical dispersion**, calculated as the difference

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

The InterQuantile Efficiency (IQE) is defined as:

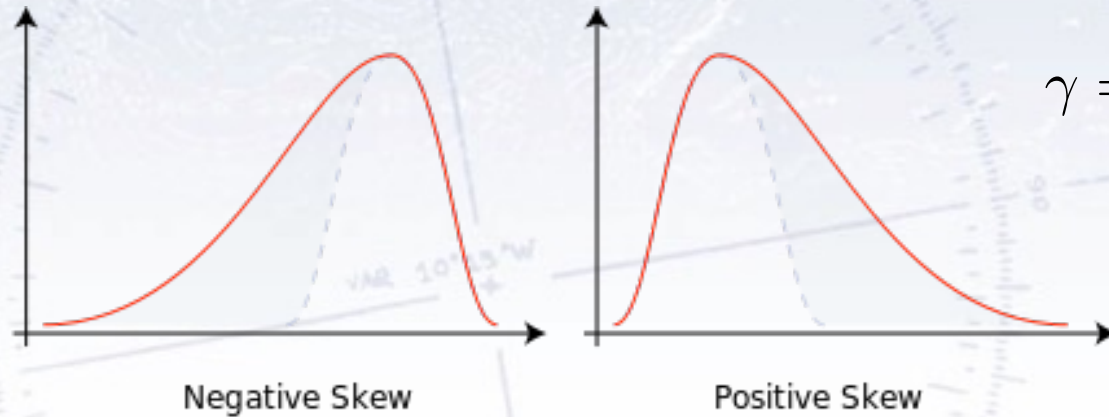
$$\text{IQE} = \text{IQR} / 1.349$$

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



# Skewness and Kurtosis

Higher moments reveal something about a distributions asymmetry and tails:



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

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

**LEPTOKURTIC**  
(thicker tails)

**MESOKURTIC**  
(normal tails)

**PLATYKURTIC**  
(thinner tails)

