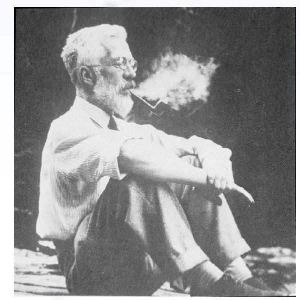
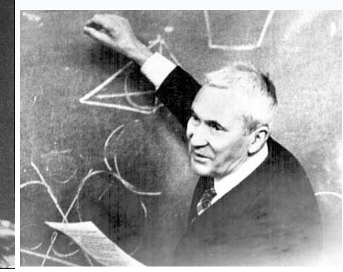
Applied Statistics

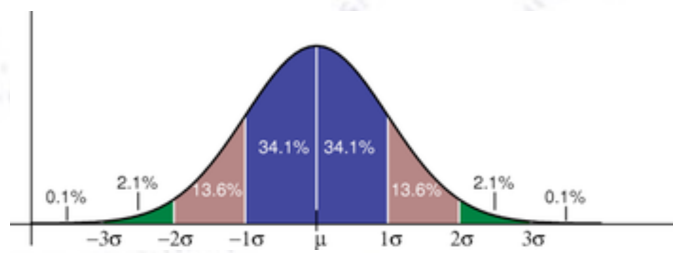
Basic Statistics





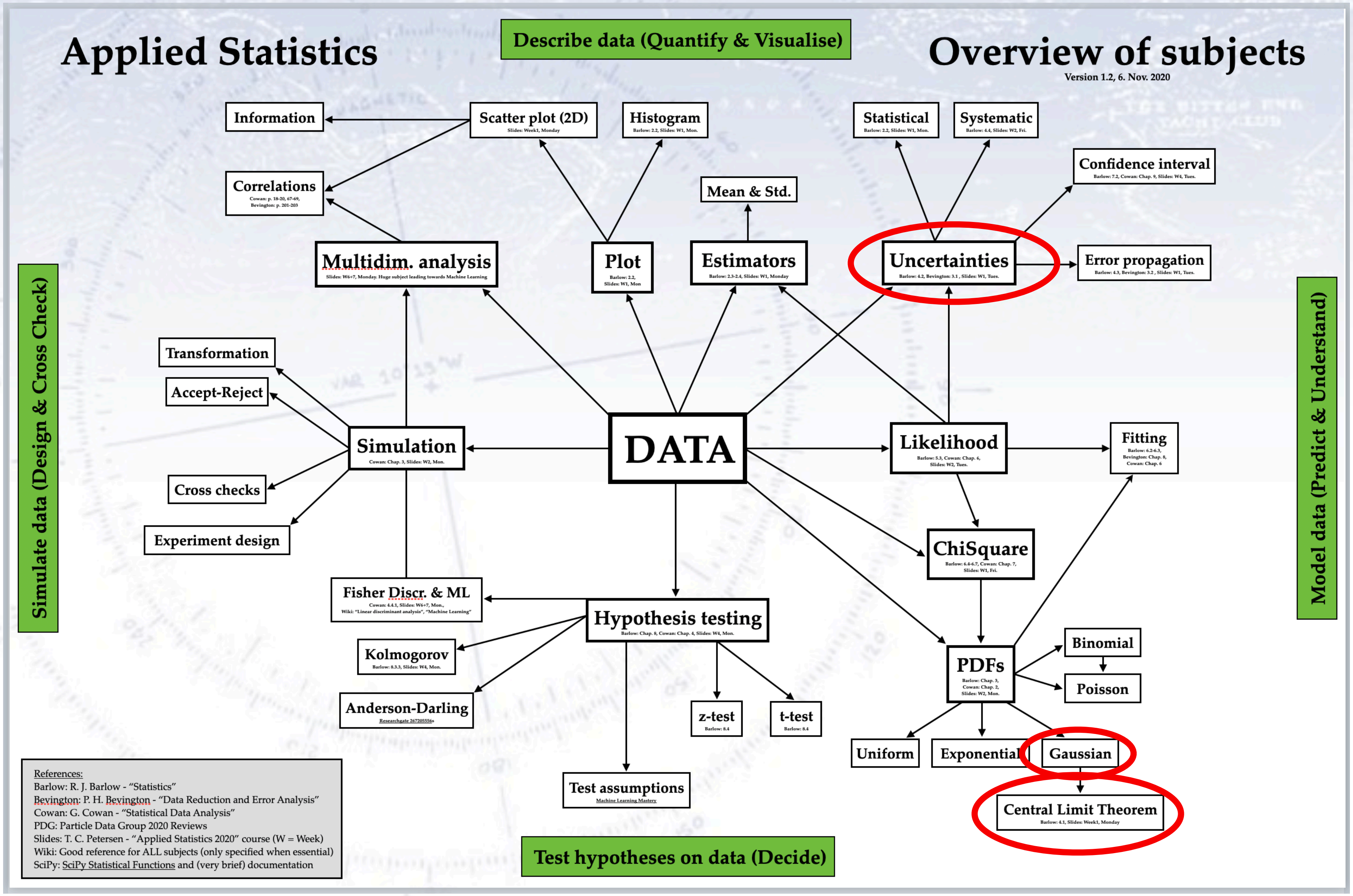
Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

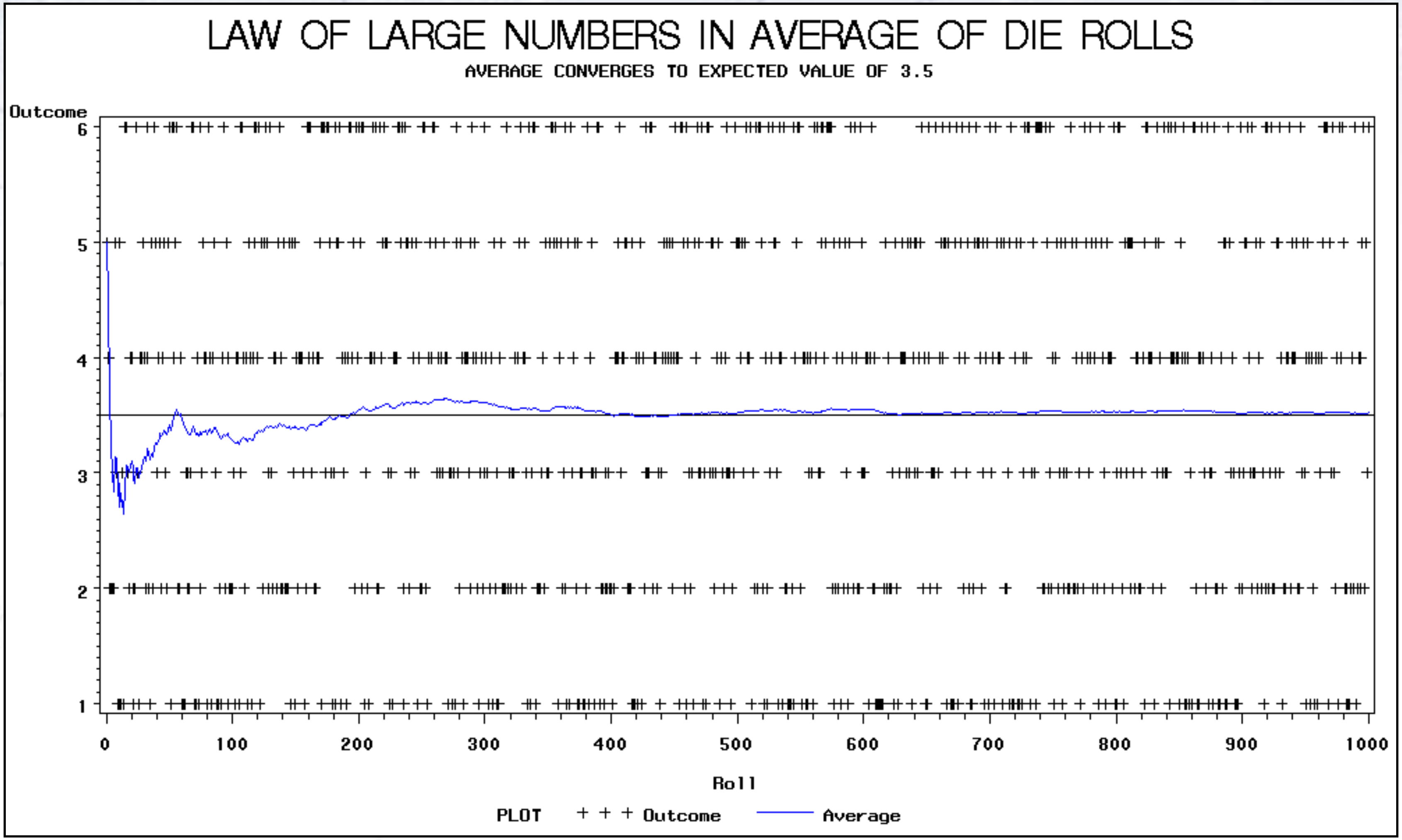
Central Limit Theorem

Central Limit Theorem



Law of large numbers

When rolling a normal die and averaging the outcome, it is no surprise that this converges towards 3.5… with enough rolls, you can get as close as you want!



Adding random numbers

If each of you chose a random number from your own favorit distribution*, and we added all these numbers, repeating this many times…

What would you expect?

* OK - to be nice to me, you agree to have similar RMSs in these distributions!

Adding random numbers

If each of you chose a random number from your own favorit distribution*, and we added all these numbers, repeating this many times…

What would you expect?

Gaussian!!!

…by the central limit theorem!

* OK - to be nice to me, you agree to have similar RMSs in these distributions!

Adding random numbers

If each of you chose a random number from your own favorite distribution*, and we added all these numbers, repeating this many times…

Gaussian!!!

the central limit theorem!

Central Limit Theorem:

The sum of N *independent* continuous random variables x_i with means μ_i and variances σ_i^2 becomes a Gaussian random variable with mean $\mu = \Sigma_i \mu_i$ and variance $\sigma^2 = \Sigma_i \sigma_i^2$ in the limit that N approaches infinity.

Central Limit Theorem

<u>Central Limit Theorem:</u>

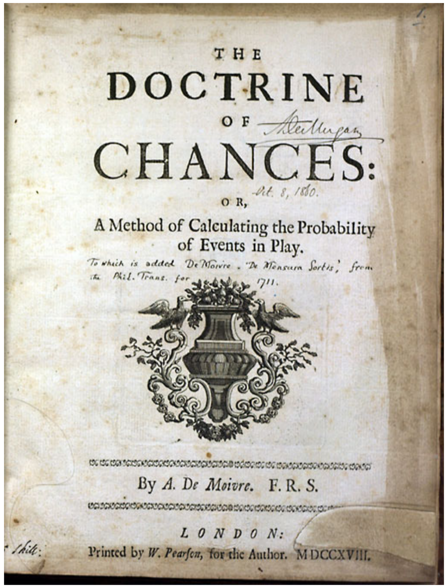
The sum of N *independent* continuous random variables x_i with means μ_i and variances σ_i^2 becomes a Gaussian random variable with mean $\mu = \Sigma_i \mu_i$ and variance $\sigma^2 = \Sigma_i \sigma_i^2$ in the limit that N approaches infinity.

The Central Limit Theorem holds under fairly general conditions, which means that the Gaussian distribution takes a central role in statistics...

The Gaussian is "the unit" of distributions!

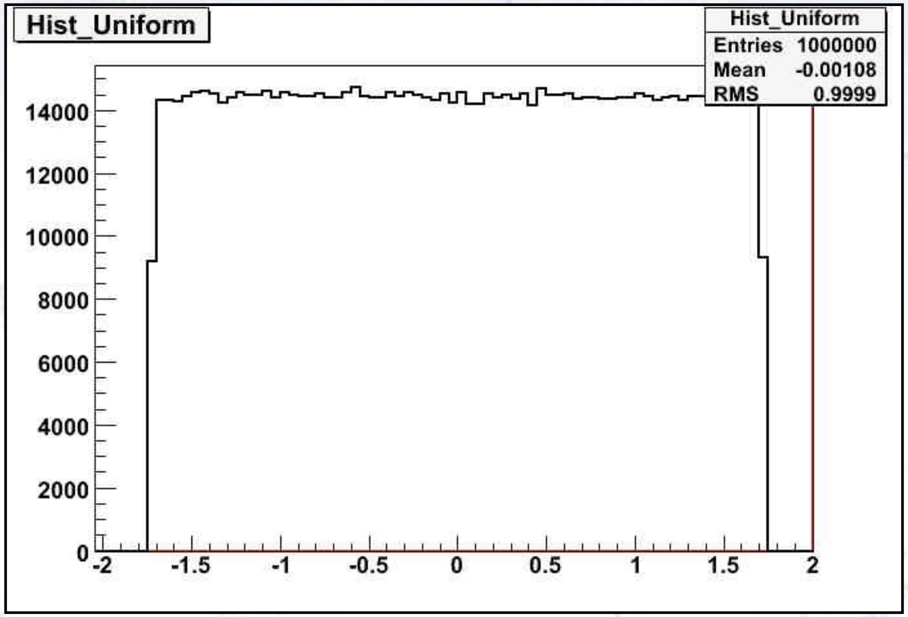
Since measurements are often affected by many small effects, uncertainties tend to be Gaussian (until otherwise proven!).

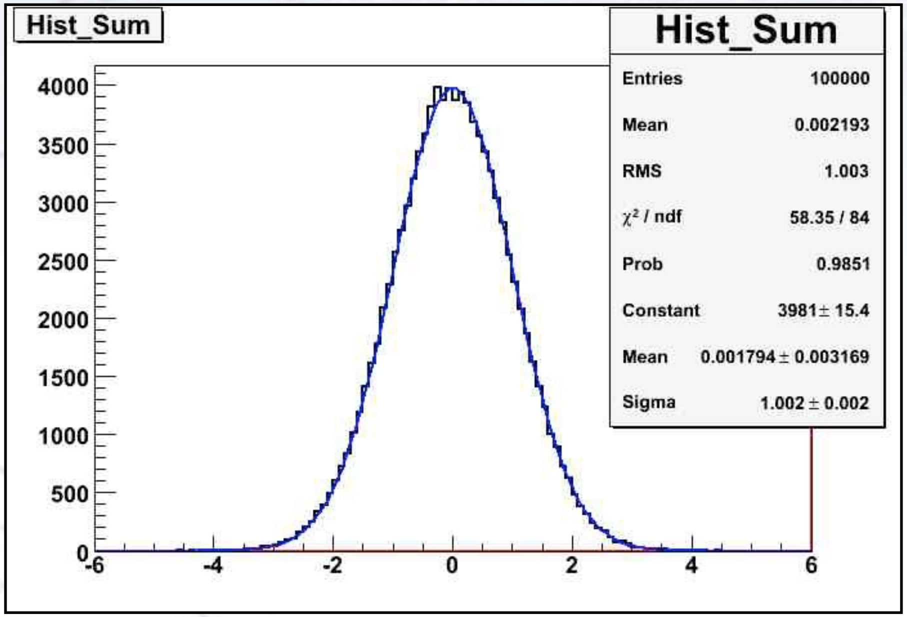
Statistical rules often require Gaussian uncertainties, and so **the central limit theorem is your new good friend..**



Example of Central Limit Theorem

Take the sum of 100 uniform numbers!
Repeat 100000 times to see what distribution the sum has…



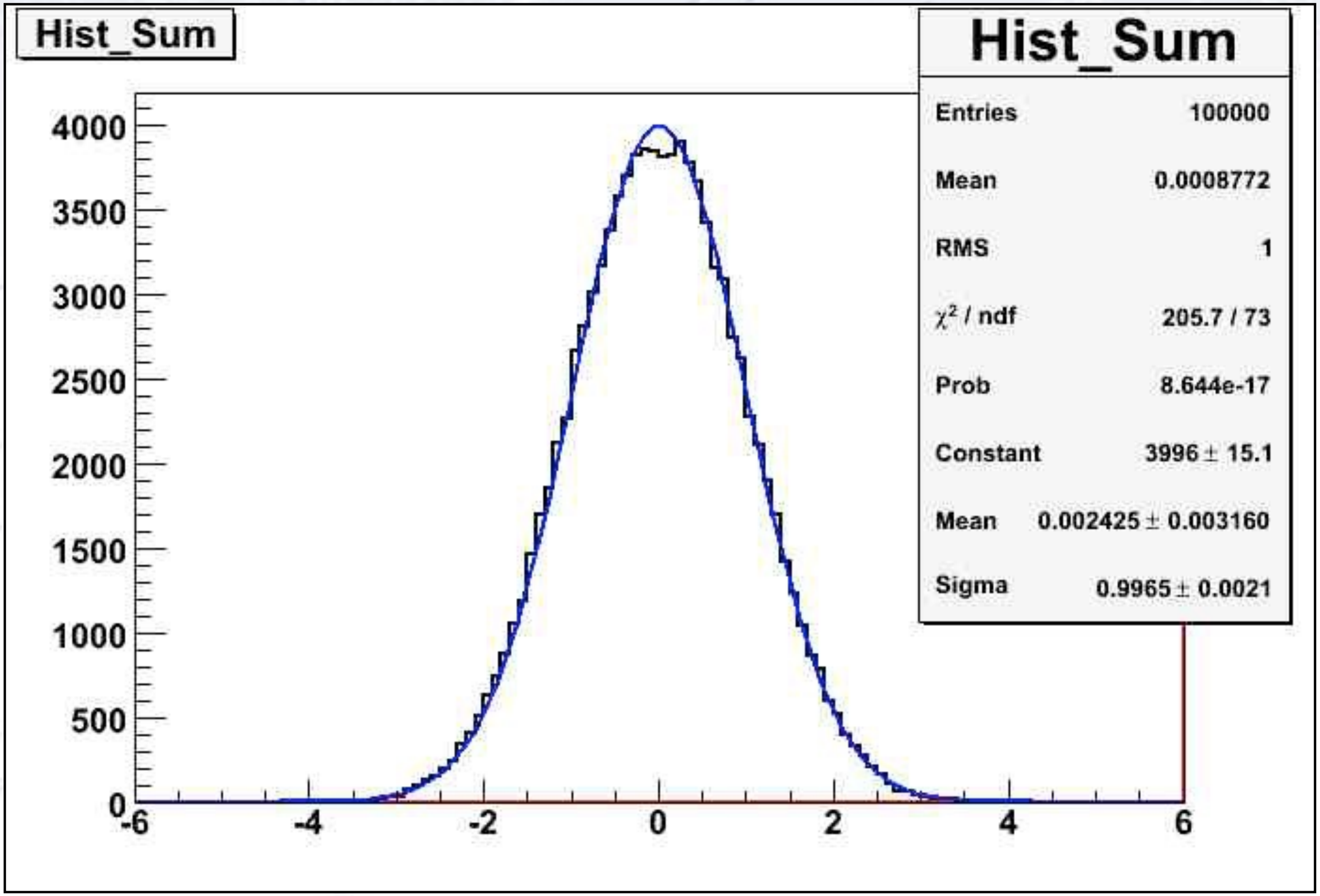


The result is a bell shaped curve, a so-called **normal** or **Gaussian** distribution.

It turns out, that this is very general!!!

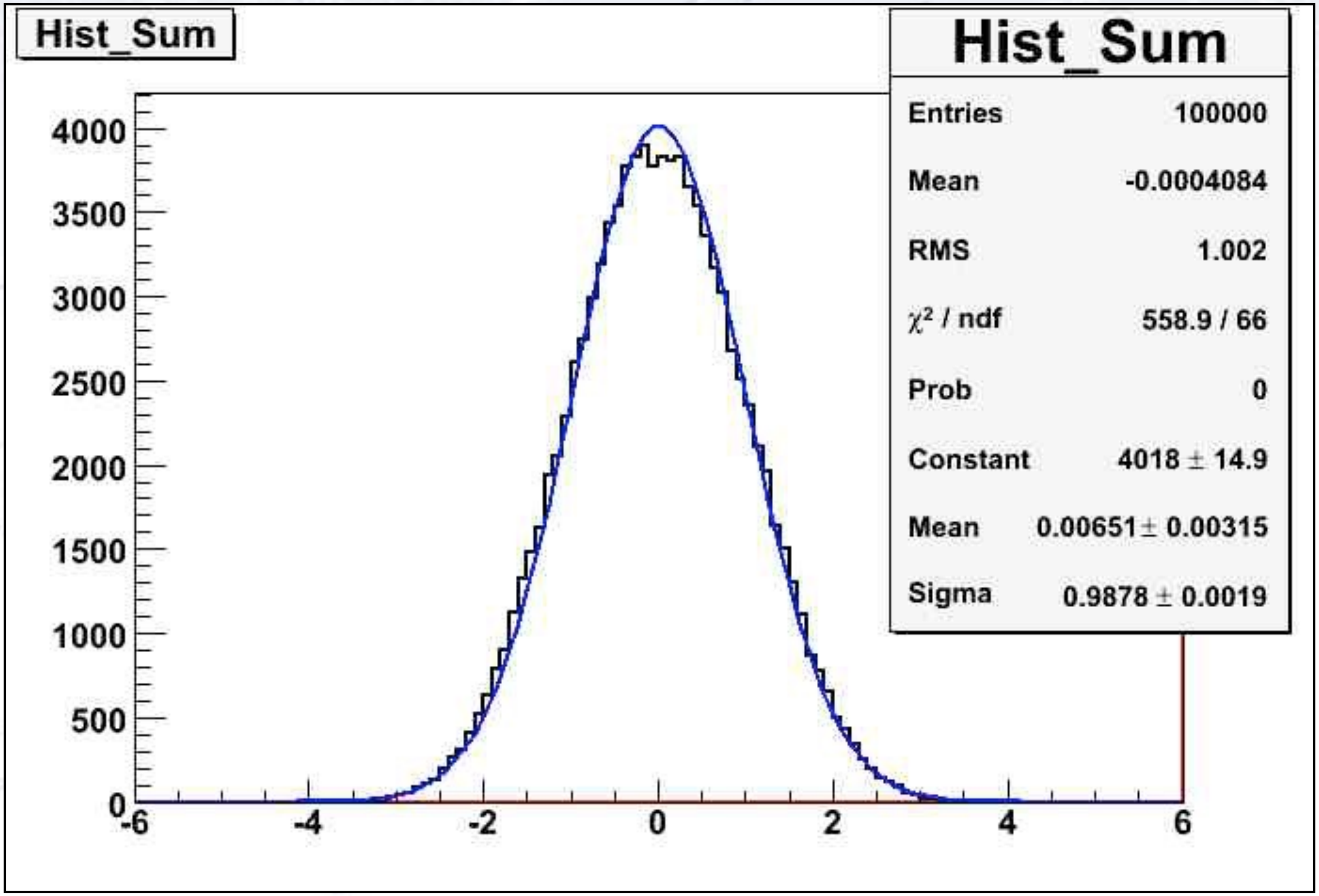
Example of Central Limit Theorem

Now take the sum of just **10** uniform numbers!



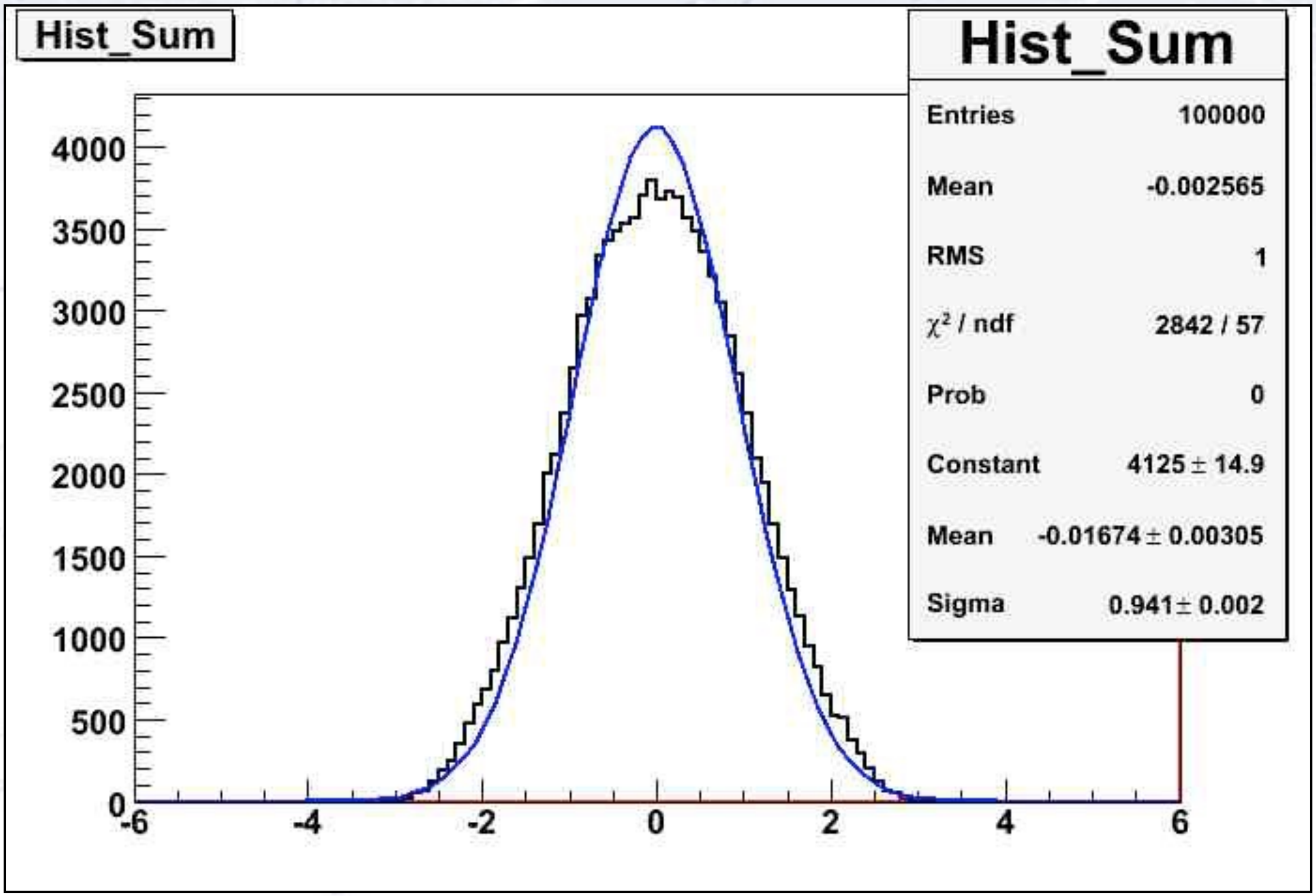
Example of Central Limit Theorem

Now take the sum of just **5** uniform numbers!



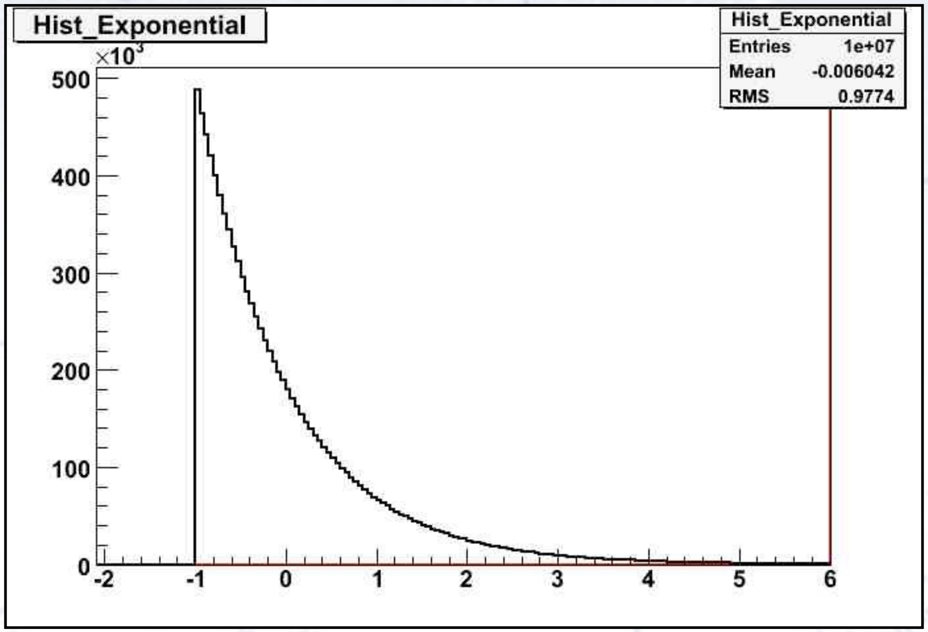
Example of Central Limit Theorem

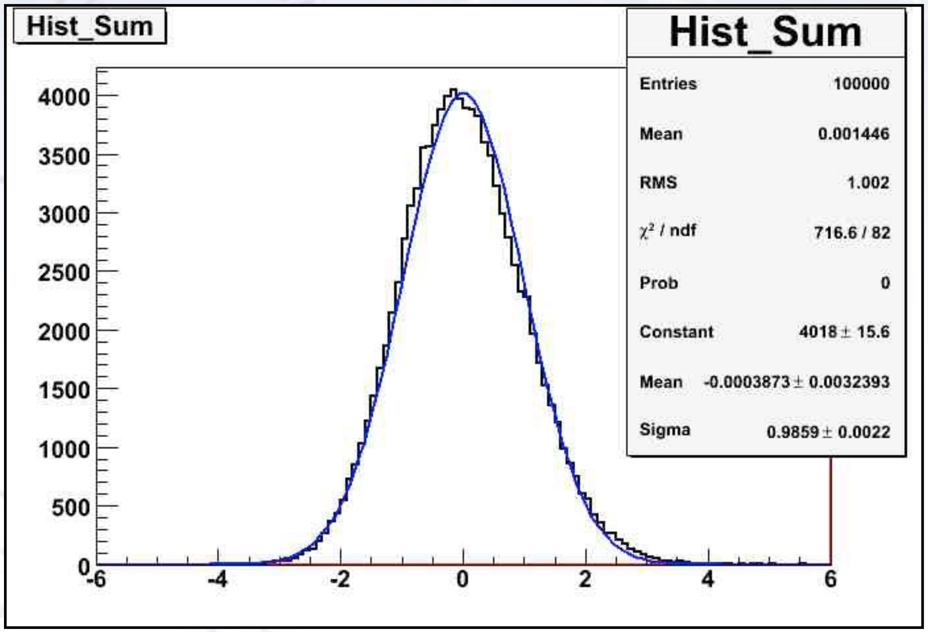
Now take the sum of just **3** uniform numbers!



Example of Central Limit Theorem

This time we will try with a much more "**nasty**" function. Take the sum of 100 *exponential* numbers! Repeat 100000 times to see the sum's distribution…





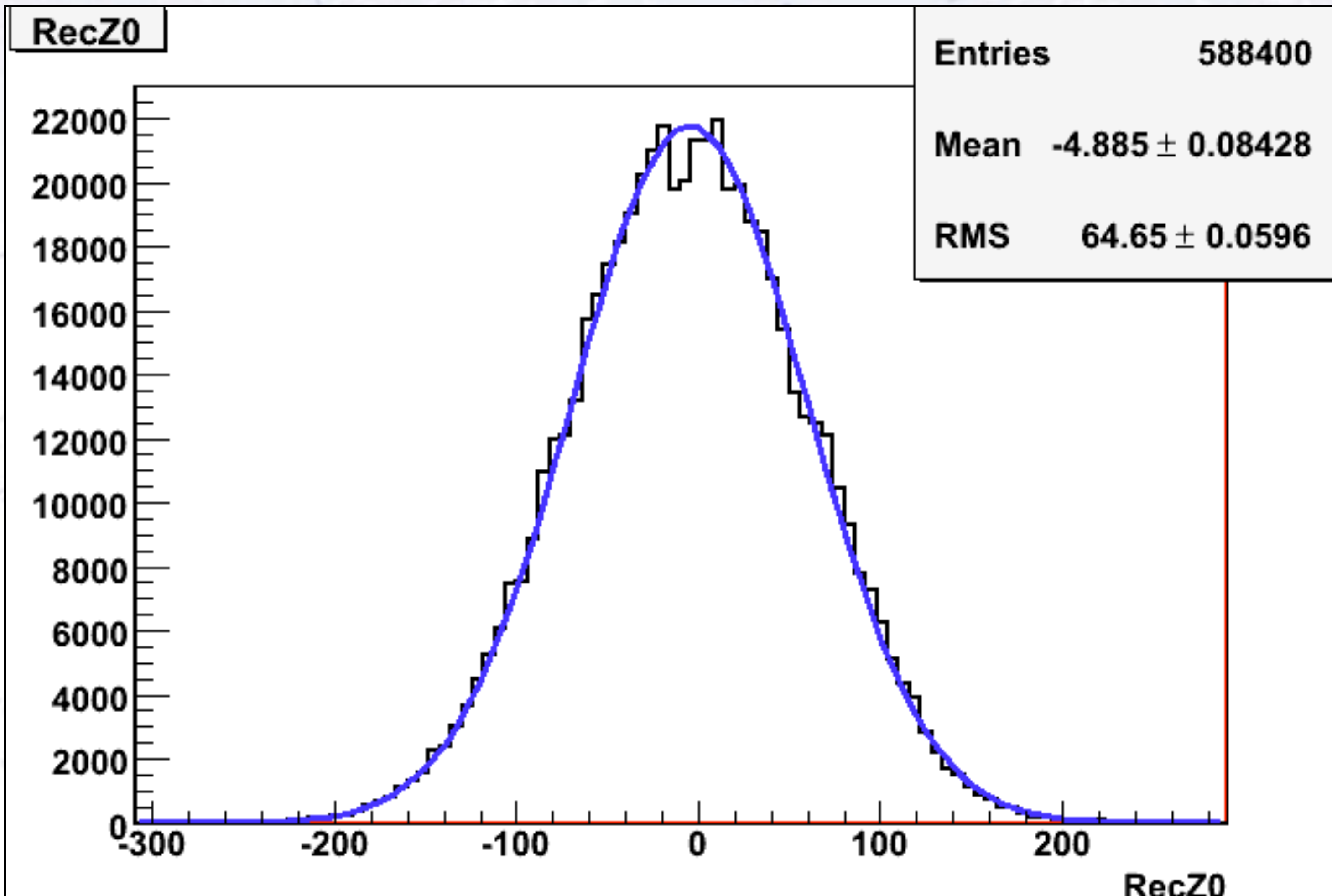
It doesn't matter what shape the input PDF has, as long as it has finite mean and width, which all numbers from the real world has! Sum quickly becomes:

Gaussian!!!

It turns out, that this fact saves us from much trouble: Makes statistics "easy"!

Example of Central Limit Theorem

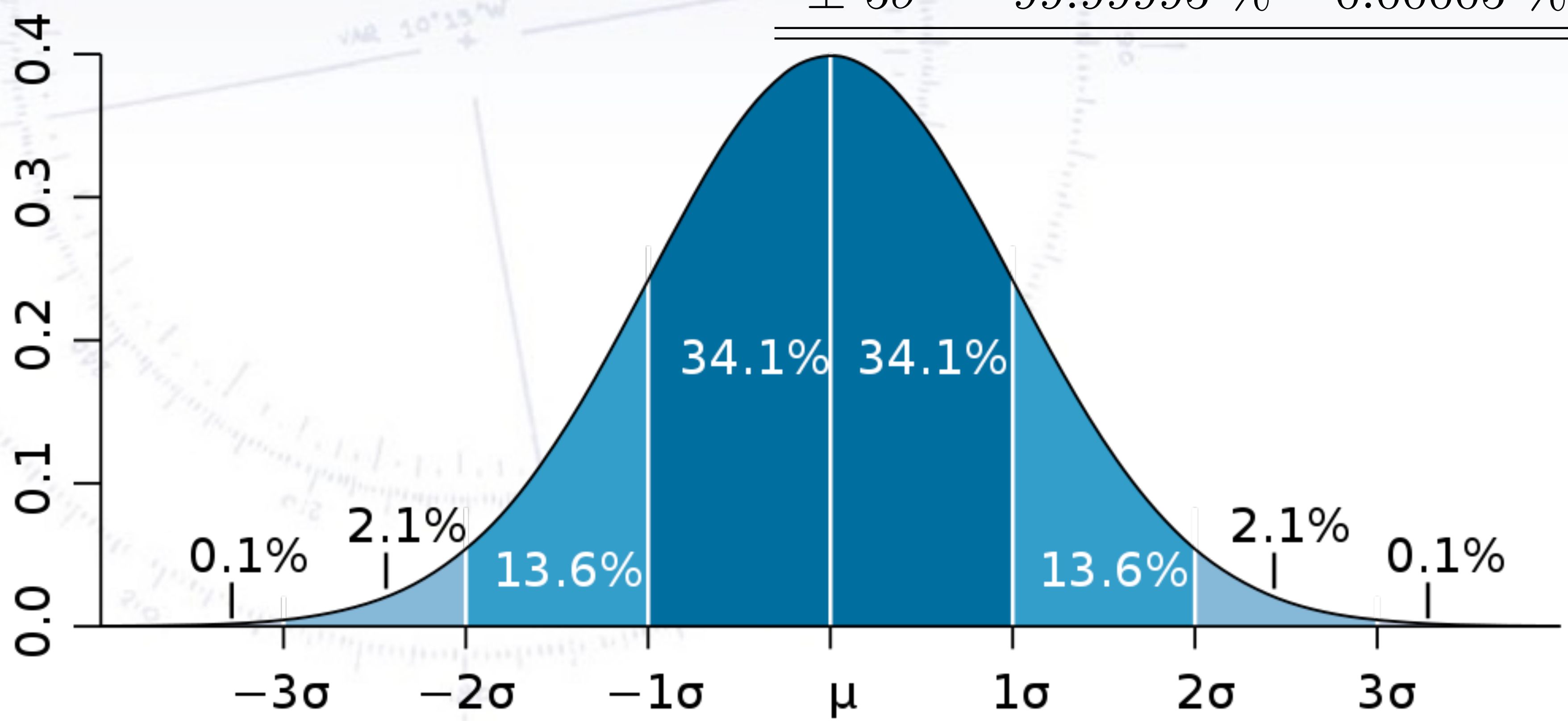
Looking at z-coordinate of tracks at vertex from proton collisions in CERNs LHC accelerator by the ATLAS detector, this is what you get:



The Gaussian distribution

It is useful to know just a few of the most common Gaussian integrals:

Range	Inside	Outside
$\pm 1\sigma$	**68** %	32 %
$\pm 2\sigma$	**95** %	5 %
$\pm 3\sigma$	**99.7** %	0.3 %
$\pm 5\sigma$	99.99995 %	0.00005 %



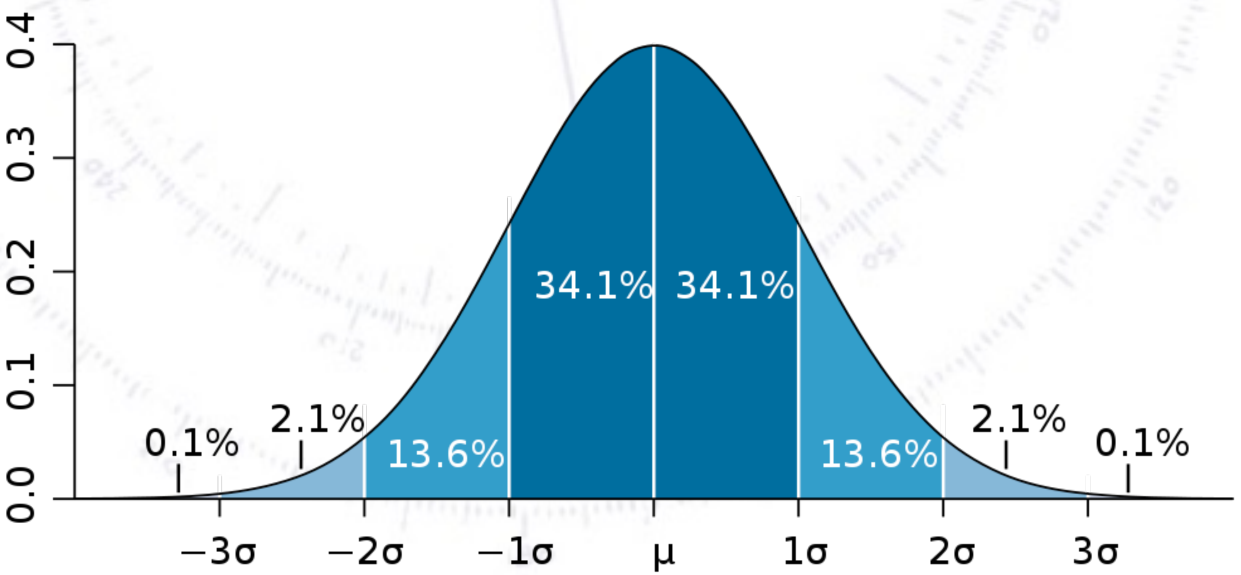
Summary

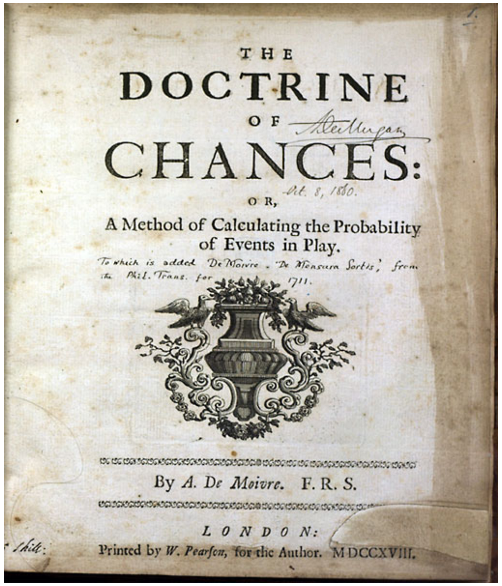
The Central Limit Theorem

...is your good friend because it...

ensures that uncertainties tend to be Gaussian

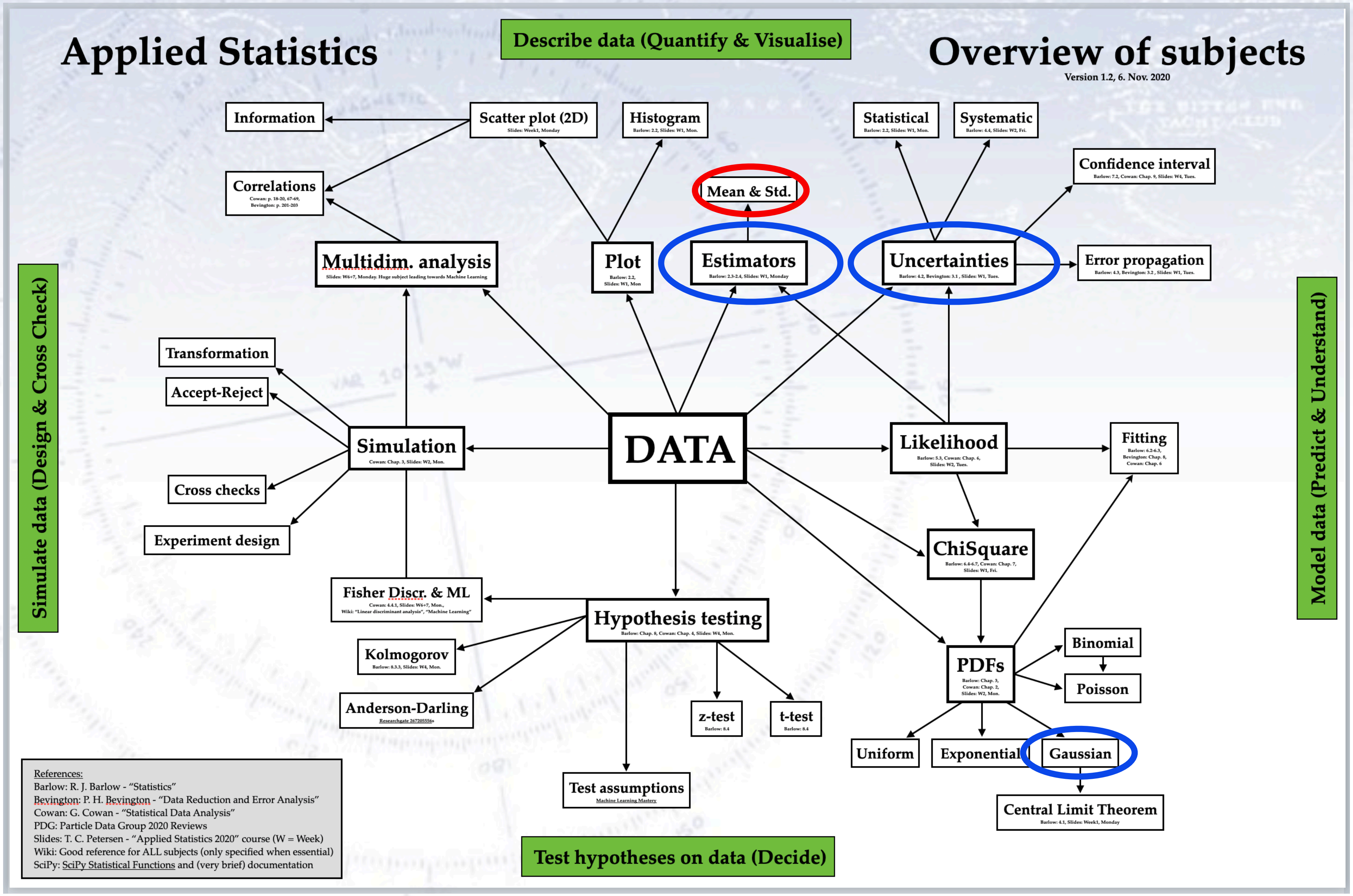
...which are the easiest to work with!





Mean & Width

Mean & Width



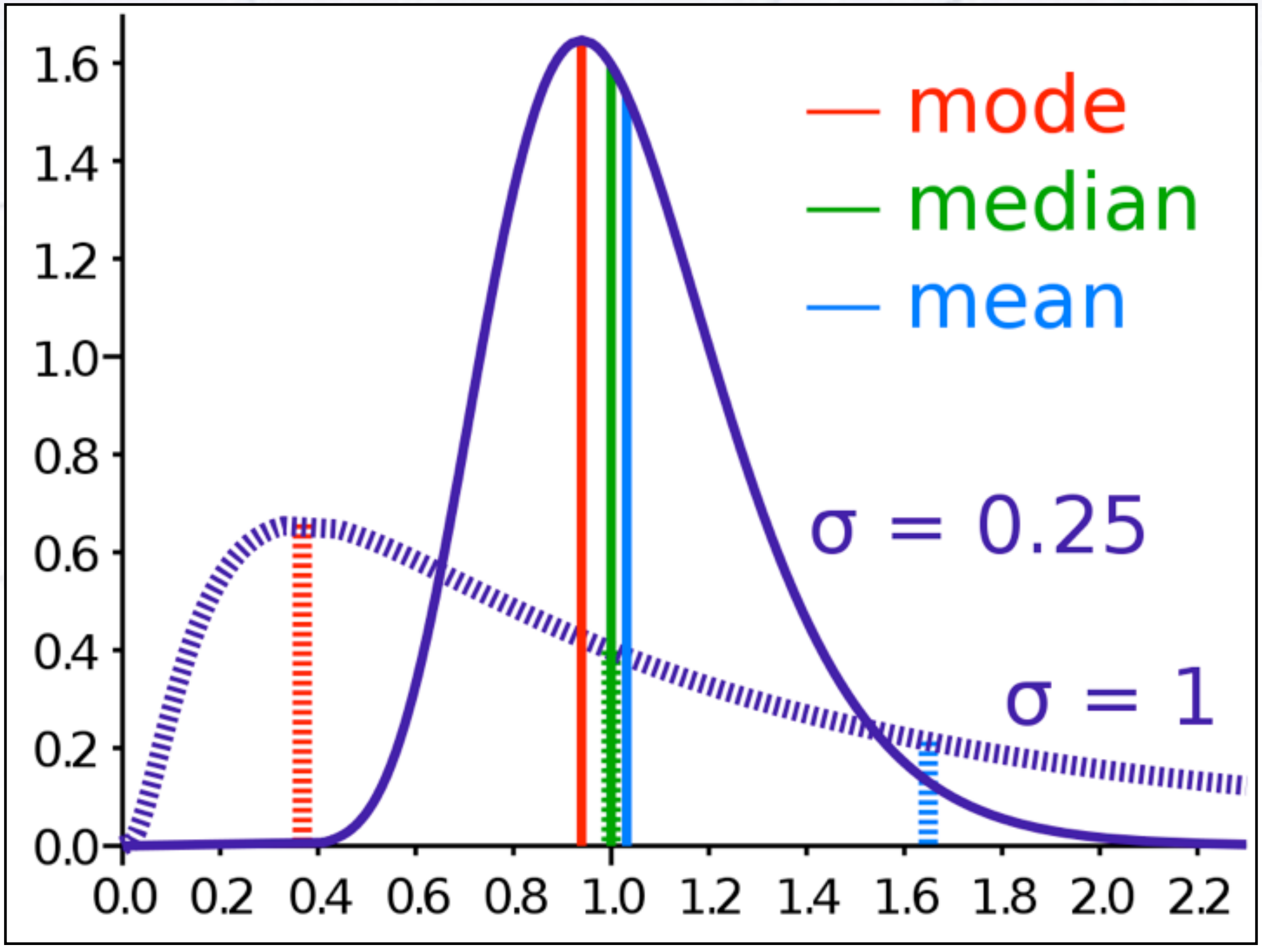
Applied Statistics
Describe data (Quantify & Visualise)
Overview of subjects
Version 1.2, 6. Nov. 2020
Information
Scatter plot (2D)
Slides: Week1, Monday
Histogram
Barlow: 2.2, Slides: W1, Mon.
Statistical
Barlow: 2.2, Slides: W1, Mon.
Systematic
Barlow: 4.4, Slides: W2, Fri.
Confidence interval
Barlow: 7.2, Cowan: Chap. 9, Slides: W4, Tues.
Correlations
Cowan: p. 18-20, 67-69,
Bevington: p. 201-203
Mean & Std.
Multidim. analysis
Slides: W6+7, Monday. Huge subject leading towards Machine Learning
Plot
Barlow: 2.2,
Slides: W1, Mon
Estimators
Barlow: 2.3-2.4, Slides: W1, Monday
Uncertainties
Barlow: 4.2, Bevington: 3.1 , Slides: W1, Tues.
Error propagation
Barlow: 4.3, Bevington: 3.2 , Slides: W1, Tues.
Simulate data (Design & Cross Check)
Model data (Predict & Understand)
Transformation
Accept-Reject
Simulation
DATA
Likelihood
Fitting
Cross checks
Experiment design
ChiSquare
Fisher Discr. & ML
Hypothesis testing
Kolmogorov
Anderson-Darling
PDFs
Binomial
Poisson
z-test
t-test
Uniform
Exponential
Gaussian
Test assumptions
Central Limit Theorem
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 (SD)**, 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):

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

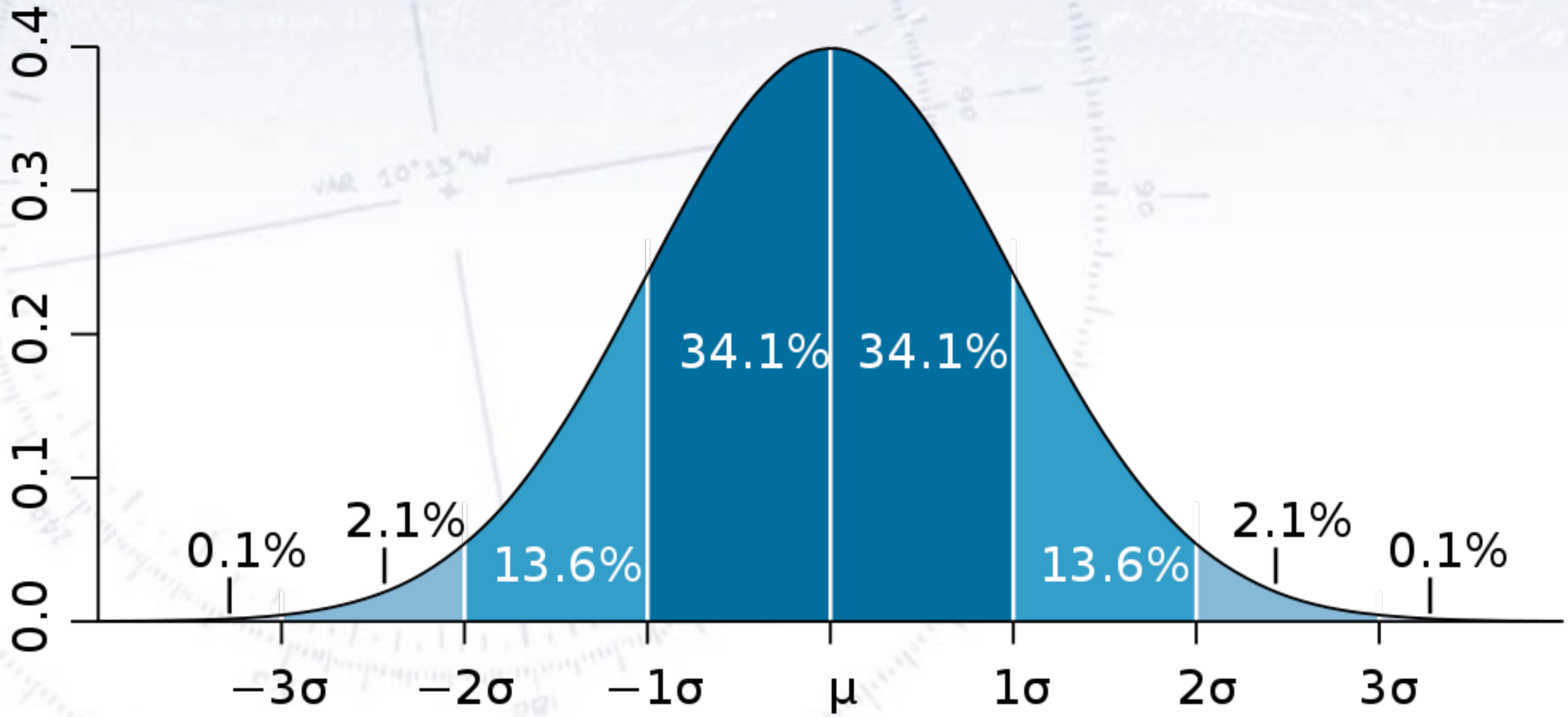
For the **standard deviation (SD)**, a.k.a. **width** or **RMSE**, it is:

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

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

SD and Gaussian σ relation

When a distribution is Gaussian, **the SD 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?

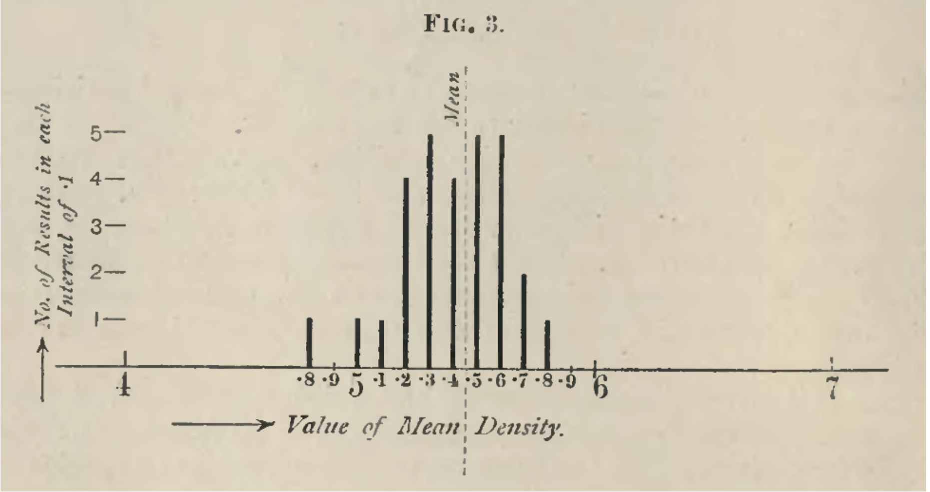
$$\hat{\sigma}_{\mu} = \hat{\sigma} / \sqrt{N}$$

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!

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

Example:

Cavendish Experiment

(measurement of Earth's density)

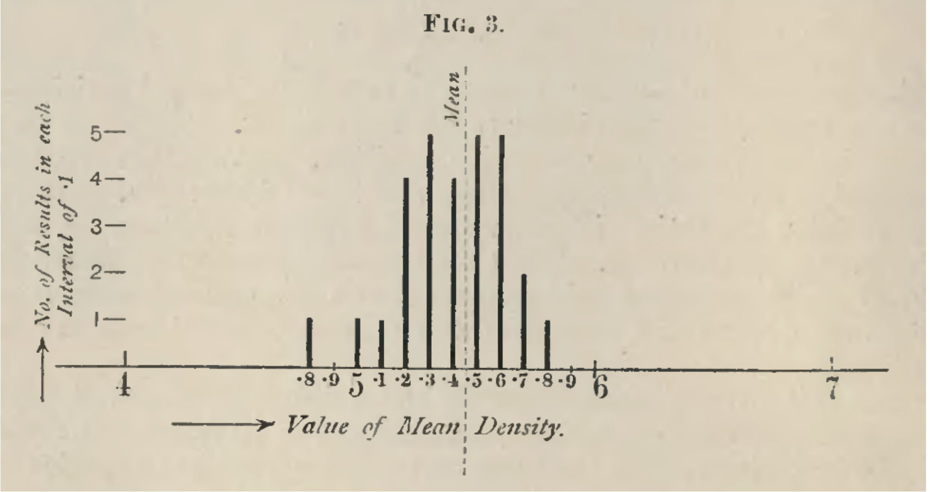
N = 29

mu = 5.42

sigma = 0.333

sigma(mu) = 0.06

Earth density = 5.42 ± 0.06



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 these, and what is the uncertainty on the average?

$$\sum x_i/\sigma_i^2$$

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.

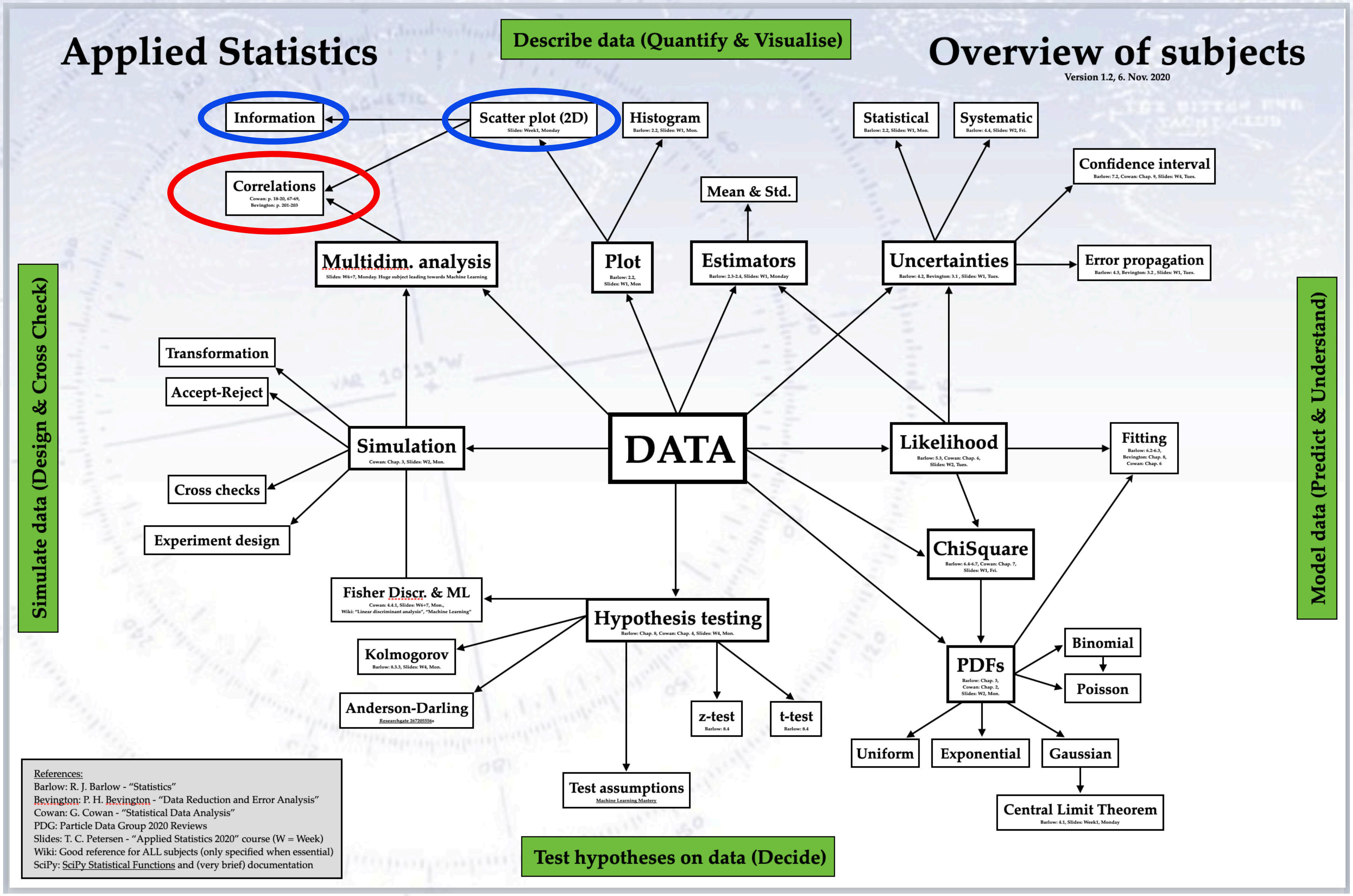
For measur … SD!
The uncerta …

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

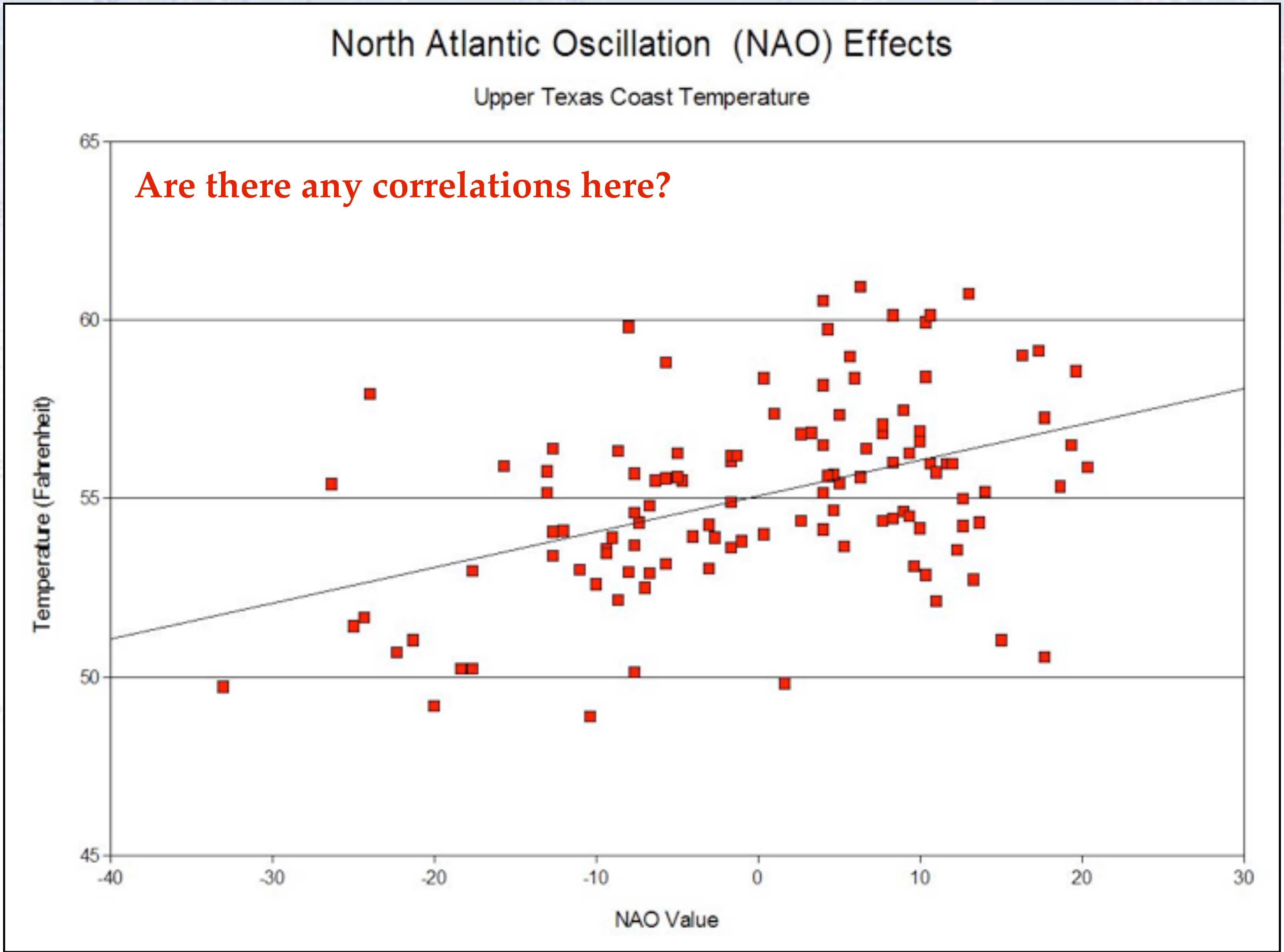
Can be understood intuitively, if two persons combine 1 vs. 4 measurements

Correlations

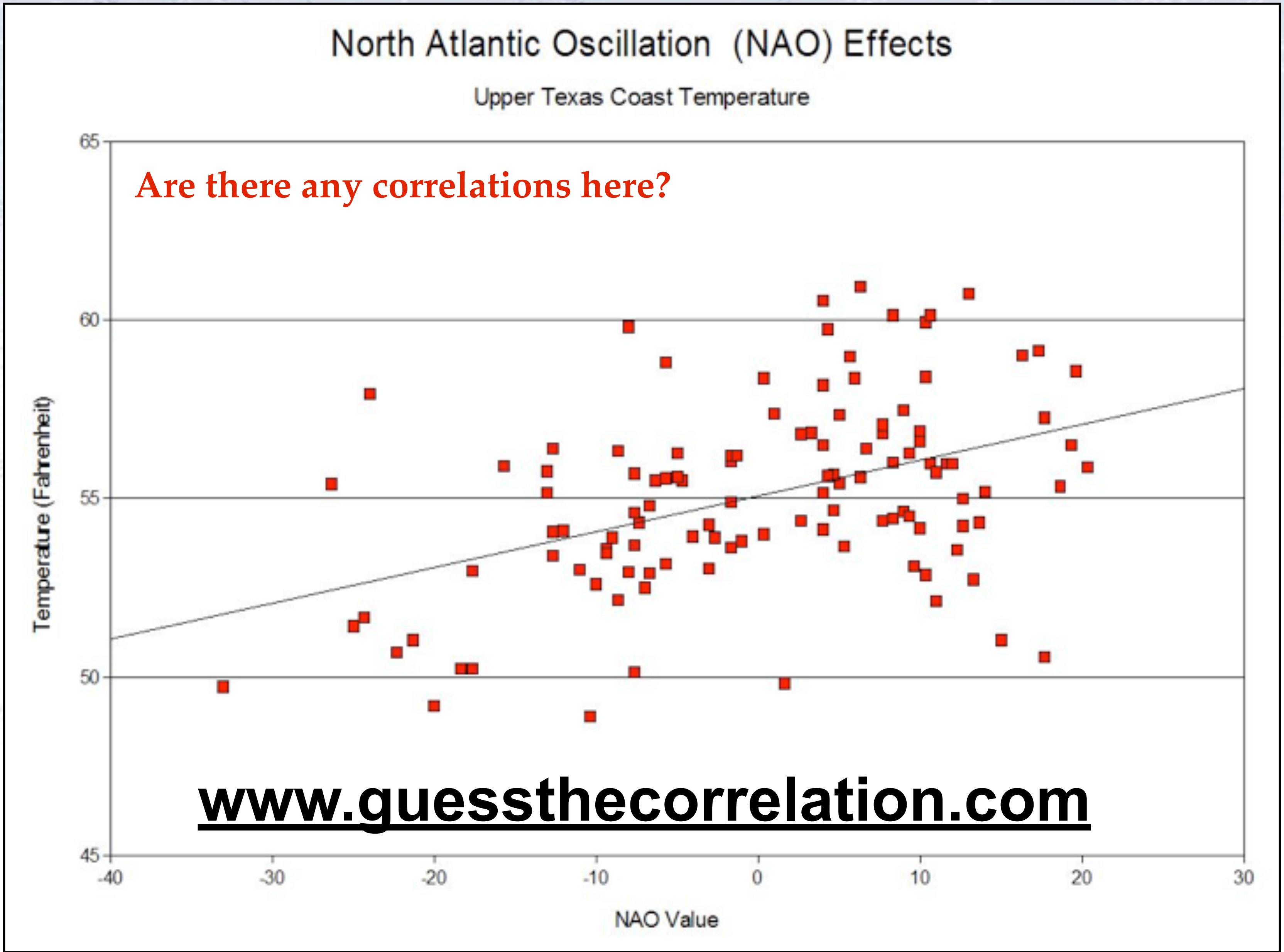
Correlations



Correlation



Correlation



www.guessthecorrelation.com

Correlation

Recall the definition of the Variance, V:

$$V = \sigma^2 = \frac{1}{N}\sum_i^n (x_i - \mu)^2 = E[(x-\mu)^2] = E[x^2] - \mu^2$$

Correlation

Recall the definition of the Variance, V:

$$V = \sigma^2 = \frac{1}{N}\sum_{i}^{n}(x_i - \mu)^2 = E[(x-\mu)^2] = E[x^2] - \mu^2$$

Likewise, one defines the **Covariance, $\mathbf{V_{xy}}$:**

$$V_{xy} = \frac{1}{N}\sum_{i}^{n}(x_i - \mu_x)(y_i - \mu_y) = E[(x_i - \mu_x)(y_i - \mu_y)]$$

Correlation

Recall the definition of the Variance, V:

$$V = \sigma^2 = \frac{1}{N}\sum_i^n (x_i - \mu)^2 = E[(x-\mu)^2] = E[x^2] - \mu^2$$

Likewise, one defines the **Covariance, $\mathbf{V_{xy}}$:**

$$V_{xy} = \frac{1}{N}\sum_i^n (x_i - \mu_x)(y_i - \mu_y) = E[(x_i - \mu_x)(y_i - \mu_y)]$$

"Normalising" by the widths, gives Pearson's (linear) correlation coefficient:

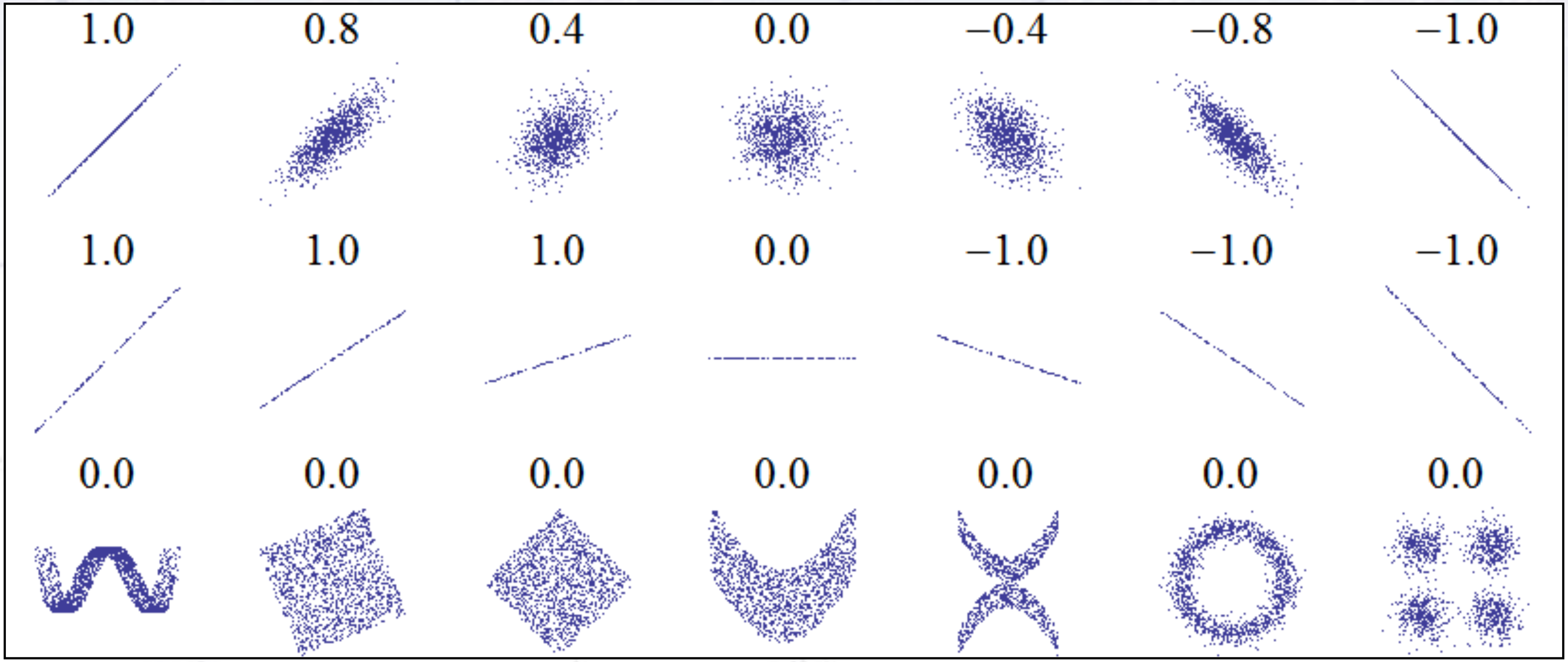
$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}$$

$$-1 < \rho_{xy} < 1$$

$$\sigma(\rho) \simeq \sqrt{\frac{1}{n}(1-\rho^2)^2 + O(n^{-2})}$$

Correlation

Correlations in 2D are in the Gaussian case the "degree of ovalness"!



Note how ALL of the bottom distributions have $\varrho = 0$, despite obvious correlations!

Significant Digits

Reporting results

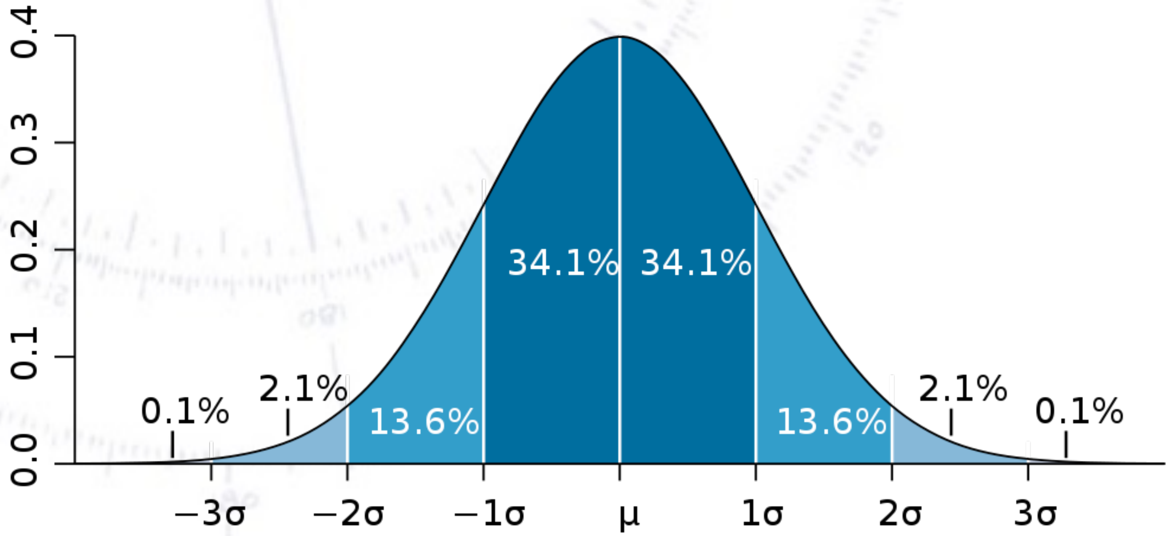
When reporting measurements, the notation is typically:

$$\mathbf{x = (0.24 \pm 0.05) \times 10^3\ m}$$

This should be interpreted as:

"with a mean of 0.24 km and a Gaussian uncertainty of 0.05 km".

This does **NOT** guaranty that x is within 0.19 km and 0.29 km!
Rather it says, that there is a 68% chance of being inside this range.



Reporting results

When reporting measurements, the notation is typically:

$$\mathbf{x = (0.24 \pm 0.05) \times 10^3\ m}$$

The reason for not writing 240 ± 50 m is that one might think, that the uncertainty has been determined with two significant digits, which is most often not the case.

Sometimes, one can find the following reporting:

$$\mathbf{x = (0.24 \pm 0.05_{stat} \pm 0.07_{syst}) \times 10^3\ m}$$

The tells the reader, that the statistical and systematic uncertainties have been kept apart, which allows for a better combination with other results (which might share some of the systematic uncertainty).

The good experimentalist gives an explained table of systematic uncertainties!

Reporting results

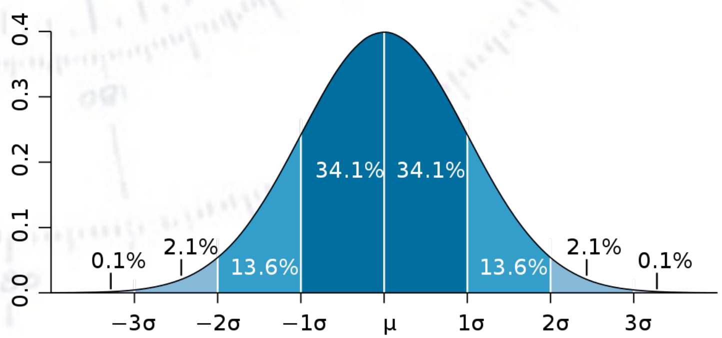
The "uncertainty on the uncertainty" follows the approximate rule:

$$\sigma_\sigma = \frac{1}{\sqrt{2N-2}}$$

Unless you have worked hard not only to reduce the uncertainty, but also to make it accurate, you should

only quote one significant digit errors, when giving results!

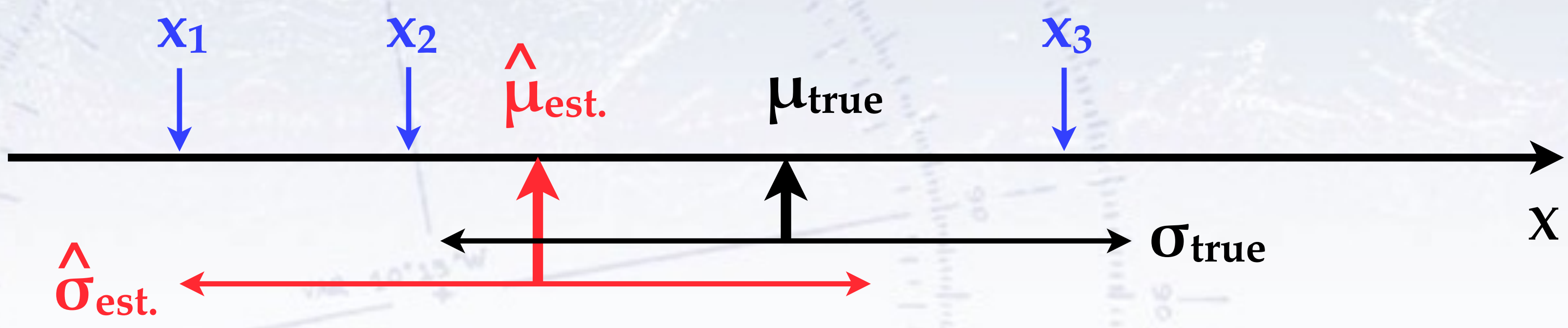
The (possible) exceptions are, if the first digit is a "1" (i.e. 0.51 ± 0.12), or internally while you are working to reduce your uncertainties.
Using two significant digits for the error is then acceptable (in this course).



Bonus Slides

Why not "just" the naive SD?

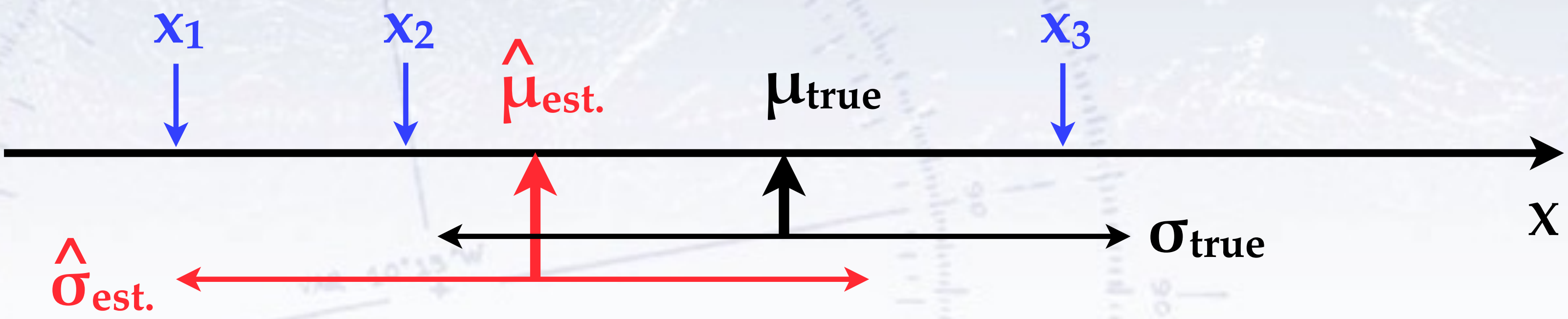
Imagine taking 3 independent measurements, and then the mean and SD:



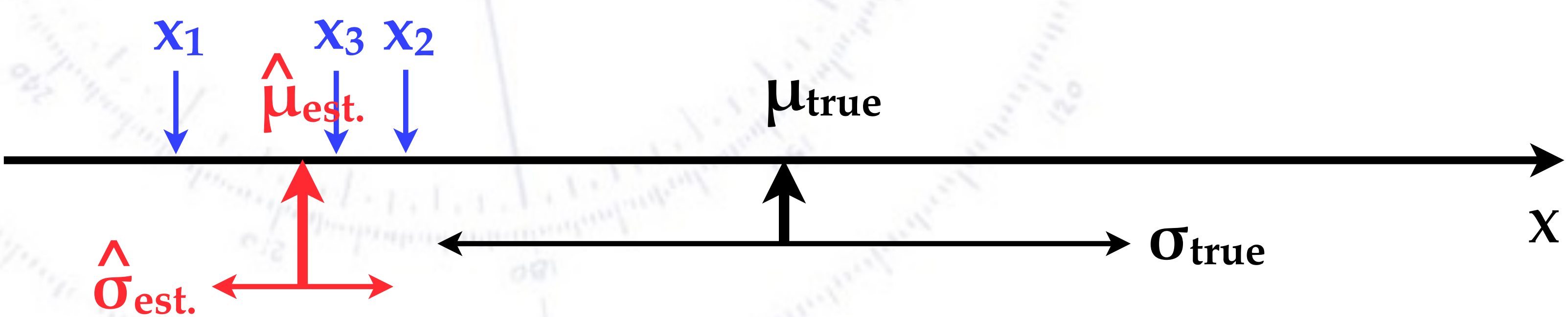
Above, all went well, because measurements were nicely distributed on both sides of the mean, and spread out according to SD.

Why not “just” the naive SD?

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



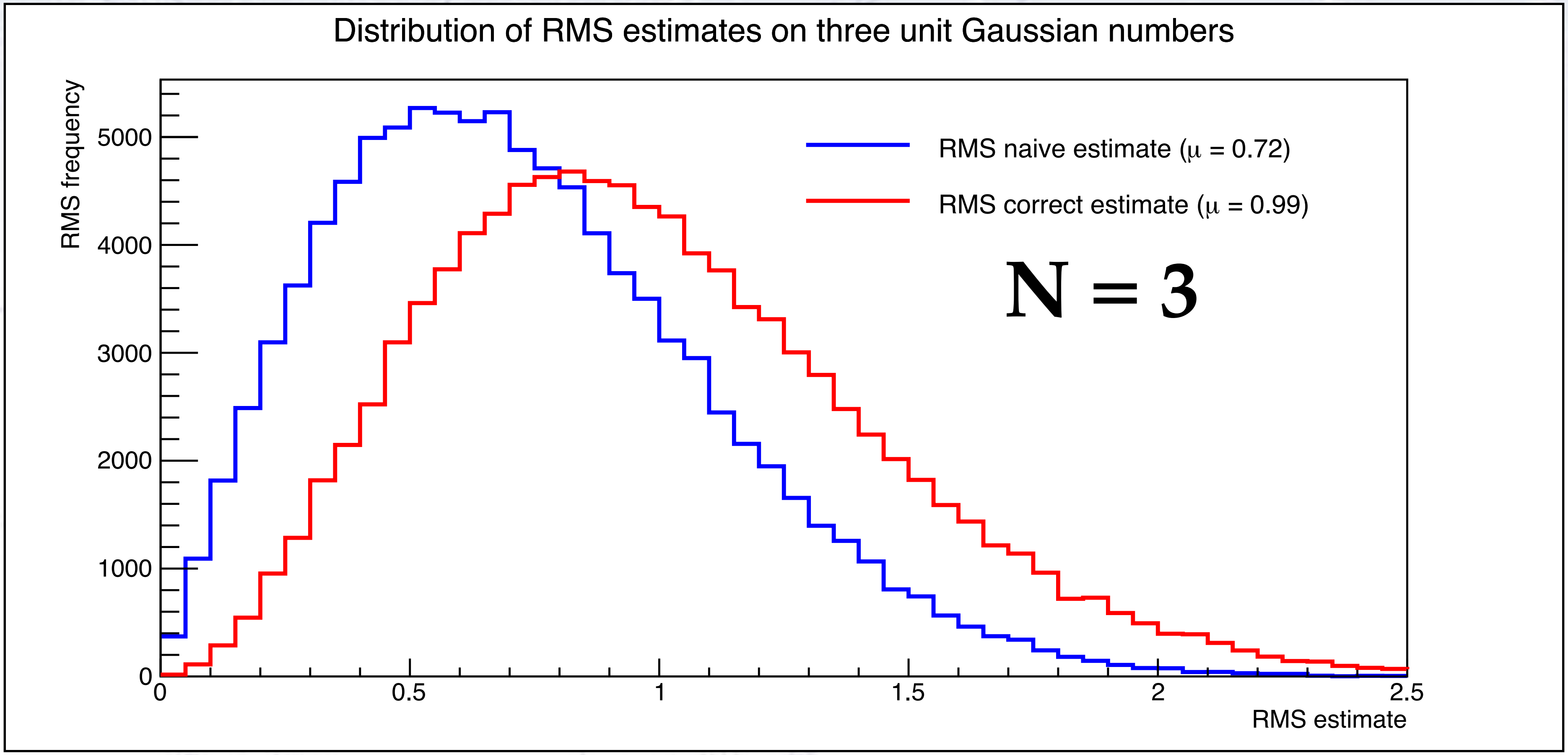
Above, all went well, because measurements were nicely distributed on both sides of the mean, and spread out according to SD.



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?

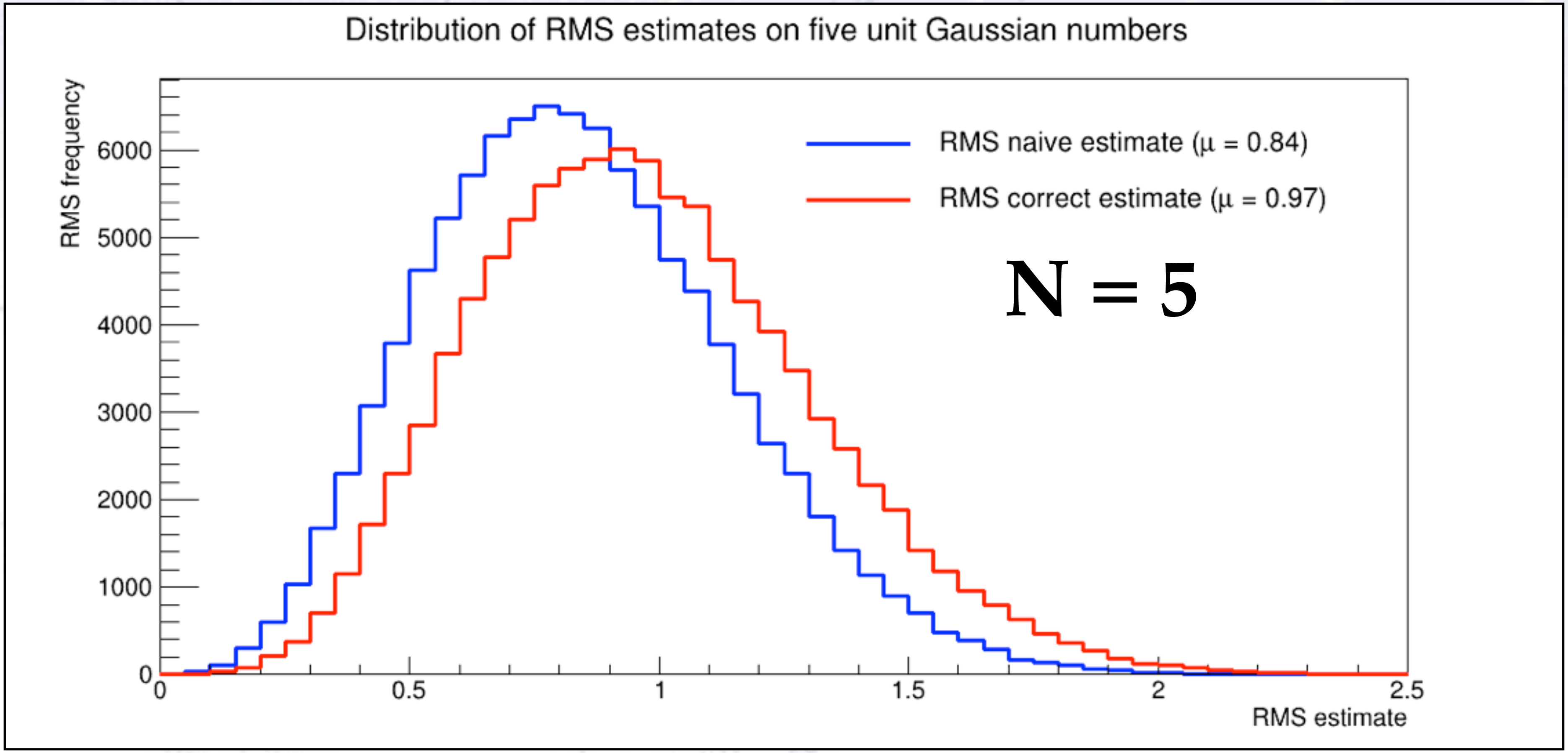
Such questions can most easily be answered by a small simulation…
Produce N=3 numbers from a unit Gaussian, and calculate the SD estimate:



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...