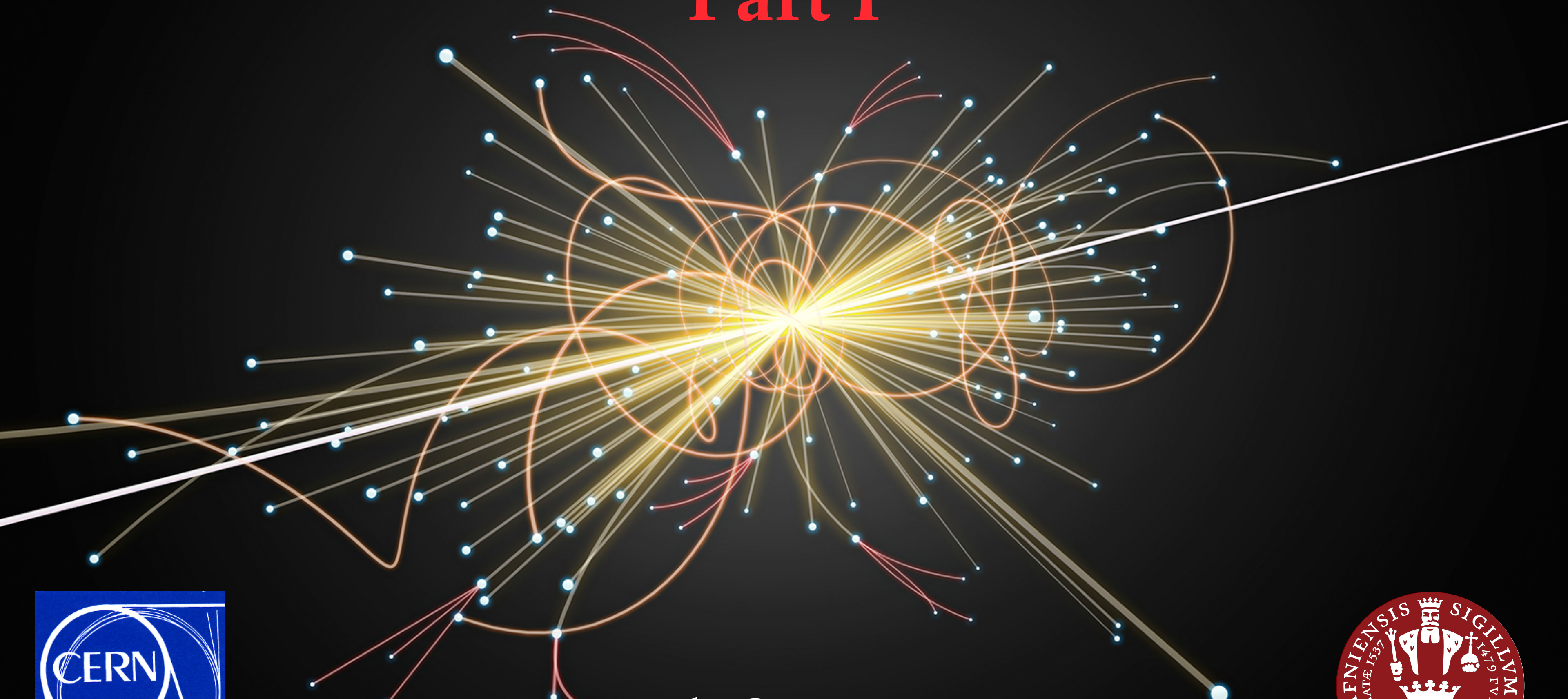
Practical Statistics

Part I



Troels C. Petersen

Niels Bohr Institute, Copenhagen

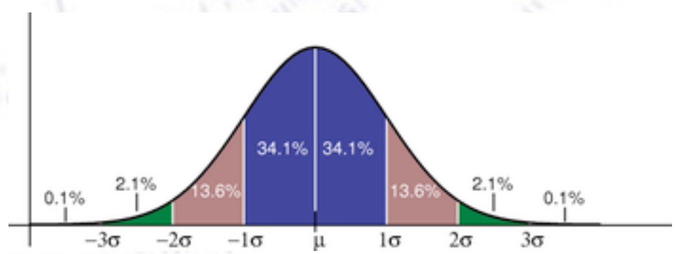
European School of High Energy Physics, 11th of September 2023

Practical Statistics

Part I - the basics

Estimators, Probability Density Functions, ChiSquare & p-value, Calibration and Simpson's Paradox

Troels C. Petersen (Niels Bohr Institute)



"Statistics is merely a quantisation of common sense"

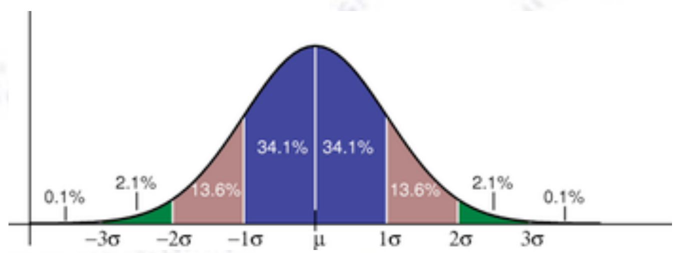
European School of High Energy Physics, 11th of September 2023

Practical Statistics

Part I - the basics

Estimators, Probability Density Functions, ChiSquare & p-value, Calibration and Simpson's Paradox

Troels C. Petersen (Niels Bohr Institute)



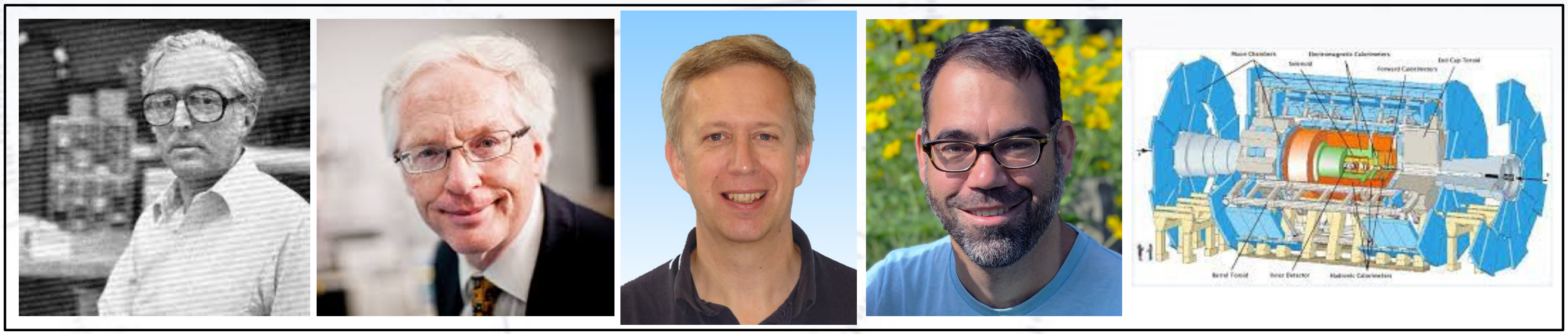
"Statistics is merely a quantisation of common sense"

European School of High Energy Physics, 11th of September 2023

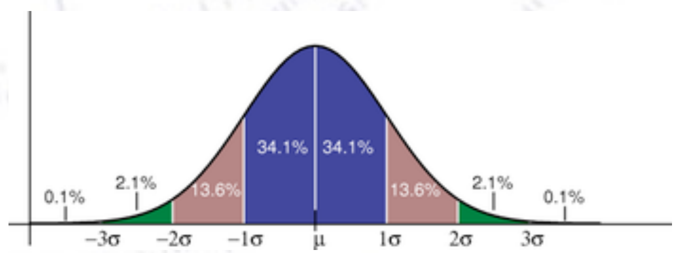
Practical Statistics

Part I - the basics

Estimators, Probability Density Functions, ChiSquare & p-value, Calibration and Simpson's Paradox



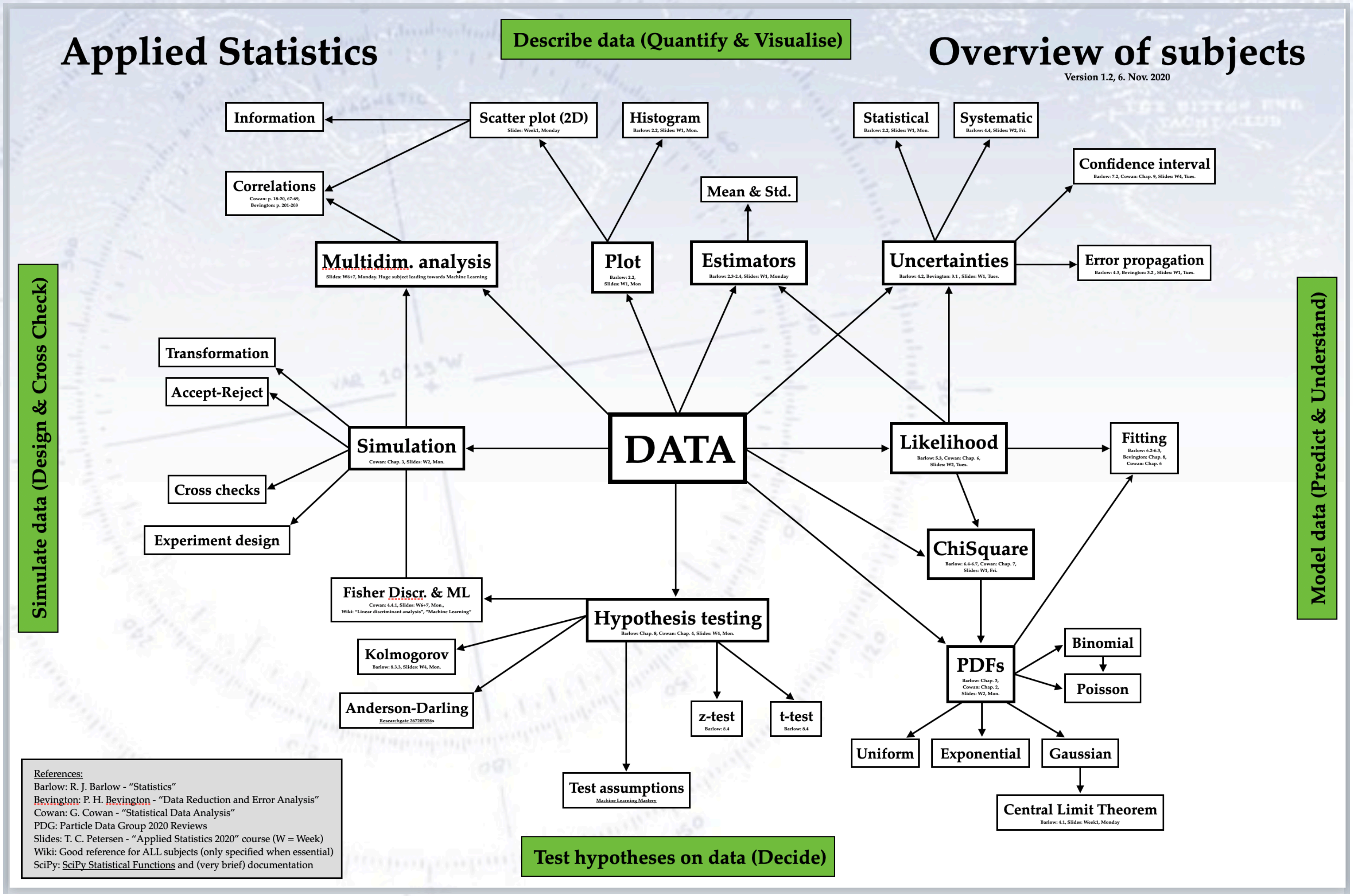
Troels C. Petersen (Niels Bohr Institute)



"Statistics is merely a quantisation of common sense"

European School of High Energy Physics, 11th of September 2023

Statistics - an overview



Outline of lectures

<u>Part I - the basics:</u>
- Estimators
- Probability Density Functions
- ChiSquare & p-values
- Calibration
- Simpson's Paradox

<u>Part II - the necessities:</u>
- Likelihood fitting
- Hypothesis testing
- Systematic uncertainties

<u>Part III - the cool:</u>
- Setting limits
- Look Elsewhere Effect
- The art of plotting
- The Fisher discriminant
- sPlots & sWeights

Outline of lectures

Part I - the basics:
- Estimators
- Probability Density Functions
- ChiSquare & p-values
- Calibration
- Simpson's Paradox

Part II - the necessities:
- Likelihood fitting
- Hypothesis testing
- Systematic uncertainties

Part III - the cool:
- Setting limits
- Look Elsewhere Effect
- The art of plotting
- The Fisher discriminant
- sPlots & sWeights

Part I - the missing:
- What is probability? Axioms!
- Bayes Theorem & Jeffrey Priors
- Proof of Central Limit Theorem
- Significant digits
- Uncertainty on uncertainties

Part II - the complicated:
- Proof of Minimum Variance Bound
- Fisher Information
- Systematic uncertainty types
- Nuisance parameters

Part III - the wierd:
- Details of Feldman-Cousins
- Time series
- …and surely lots more!

Why Statistics?

Why uncertainties?

In physics there are various elements of uncertainty:

- Theory is not deterministic
  Examples: Quantum effects & chaos
- Random measurement errors
  Fluctuations are present even without quantum effects!
- Things we could know in principle but don't…
  e.g. from limitations in cost, time, etc.

We can quantify the uncertainty using **PROBABILITY**

Armed with the realisation of limitations, we can make better calculations/experiments and informed conclusions.

Example: Speed of Gravity

Imagine that you measured the speed of gravity, and got the following result:

$$v_{\text{gravity}} = 2.89 \times 10^8 \text{ m/s}$$

That would tell you...

Example: Speed of Gravity

Imagine that you measured the speed of gravity, and got the following result:

$$v_{\mathrm{gravity}} = 2.89 \times 10^{8}\ \mathrm{m/s}$$

That would tell you...

Nothing!!!

Because you have no idea of the uncertainty.

Example: Speed of Gravity

Imagine that you measured the speed of gravity, and got the following result:

$$v_{\text{gravity}} = 2.89 \times 10^8 \text{ m/s}$$

Depending on the uncertainty, you might foresee three very different conclusions:

$$v_{\text{gravity}} = (2.89 \pm 9.21) \times 10^8 \text{ m/s}$$

Could be anything,
even negative!

$$v_{\text{gravity}} = (2.89 \pm 0.09) \times 10^8 \text{ m/s}$$

Consistent with c,
and not much else!

$$v_{\text{gravity}} = (2.89 \pm 0.01) \times 10^8 \text{ m/s}$$

Inconsistent with c:
New Discovery!!!

(extreme) Conclusion:

Numbers without stated uncertainties are meaningless!

Why precision?

How well do we know Newton's Law of Gravity?

$$F = G\frac{mM}{r^2}$$

Newton's Law of Gravity

How well do we know Newton's Law of Gravity? Well, reasonably well, but...

Force central?

Valid for all masses?

$$F = G\frac{mM}{r^2}$$

Range of validity?

Square Law?

No other dependencies?

Newton's Law of Gravity

How well do we know Newton's Law of Gravity? Well, reasonably well, but...

Force central?

Seemingly...

Valid for all masses?

NO - not large ones!

Why is G so small?

$$F = G\frac{mM}{r^2}$$

Being tested: Related to search for more dimensions

Maybe not short ranges

Range of validity?

Square Law?

Yes, from generel relativity

No other dependencies?

Why statistics in physics?

Experimental measurements are only **SAMPLES** of the reality, they can never represent the entire set of possibilities, so
→ they are affected by uncertainties
→ results can be expressed as probabilities

Theoretical calculations are mostly **APPROXIMATIONS** limited by finite resources to do the calculations or by imprecise input parameters, so
→ they are also affected by uncertainties
→ predictions can also be expressed in terms of probability

Statistics gives the understanding of uncertainty and probability in relating data and theory!!!

Why statistics in physics?

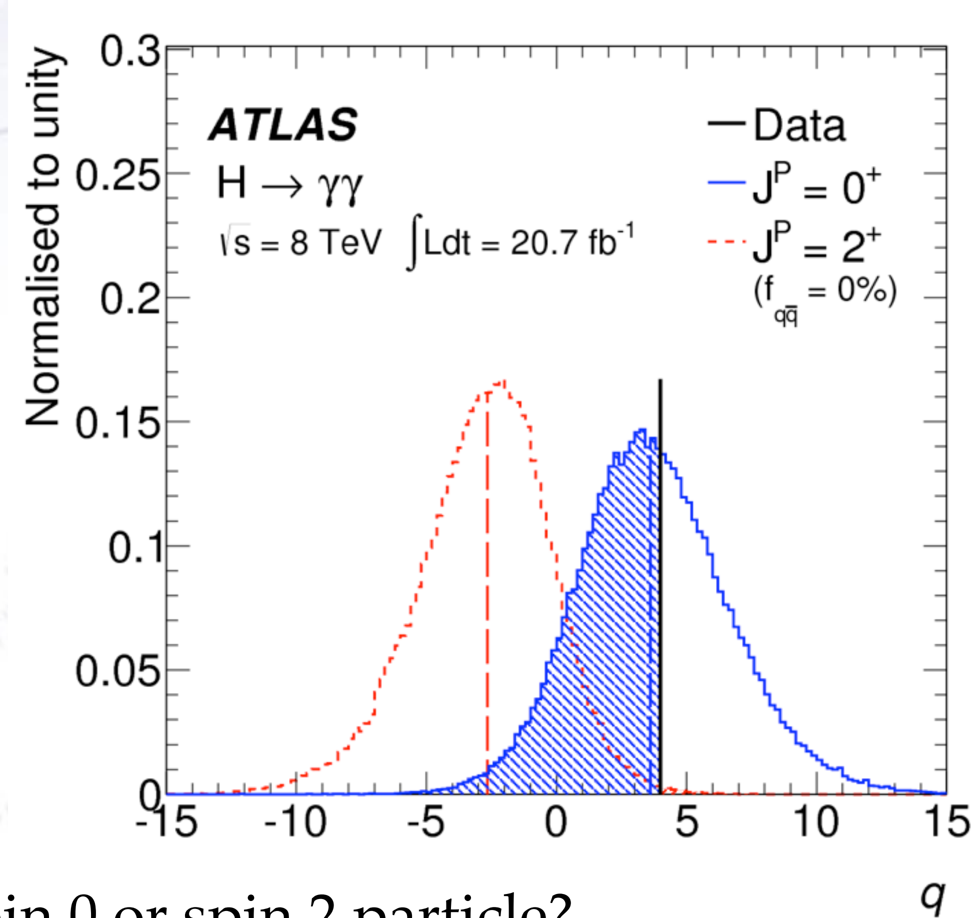
Statistics is about **hypothesis testing**, quantifying the answer to the question

"which theory matches the data best?"

Statistics is about collecting data and logically analysing it, not being fooled by coincidences and chance observations.

Statistics is about fitting trends in data, allowing for projections and predictions.

Statistics is about understanding data, and extracting the essential information from it in the most powerful way.



Is the Higgs a spin 0 or spin 2 particle?

Biases in statistics...

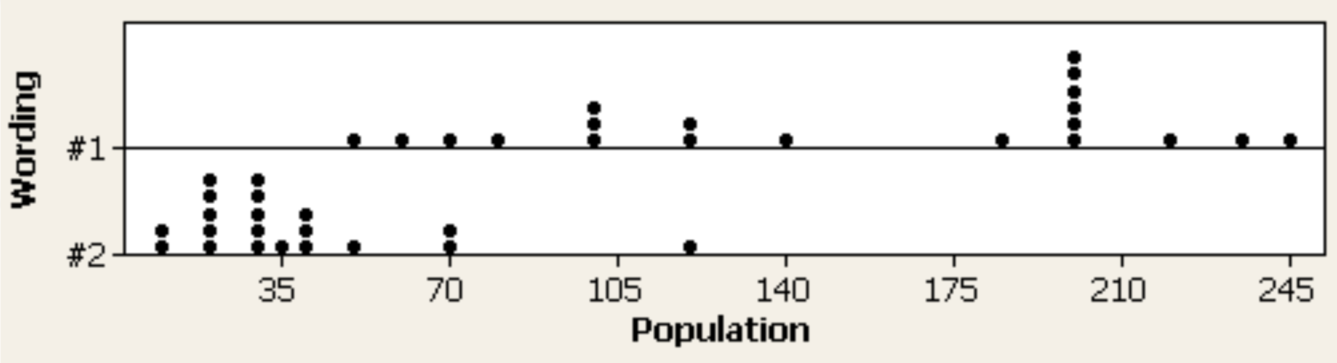
When ASKING people, one may introduce (deliberate?) biases:

- ***Wording 1*:** Pick a color: red or blue?
- ***Wording 2*:** Pick a color: blue or red?

Color Choice	Red	Blue
Wording 1	59 %	41 %
Wording 2	45 %	55 %

One may also bias answers by giving (ir-)relevant information:

- ***Wording 1*:** Knowing that the population of the U.S. is 270 million, what is the population of Canada?
- ***Wording 2*:** Knowing that the population of Australia is 15 million, what is the population of Canada?



Biases in statistics...

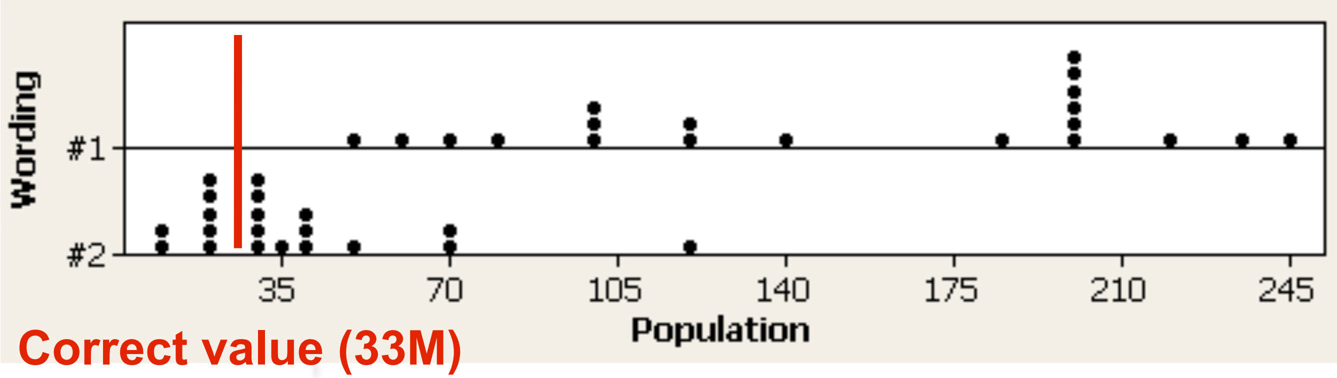
When ASKING people, one may introduce (deliberate?) biases:

- ***Wording 1*:** Pick a color: red or blue?
- ***Wording 2*:** Pick a color: blue or red?

Color Choice	Red	Blue
Wording 1	59 %	41 %
Wording 2	45 %	55 %

One may also bias answers by giving (ir-)relevant information:

- ***Wording 1*:** Knowing that the population of the U.S. is 270 million, what is the population of Canada?
- ***Wording 2*:** Knowing that the population of Australia is 15 million, what is the population of Canada?



Mark Twain:

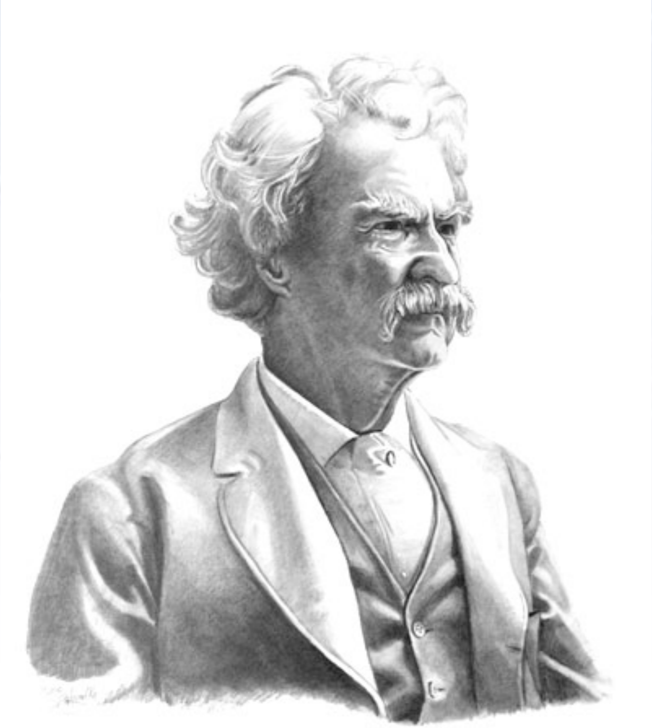
"There are three kinds of lies:

lies, damned lies, and statistics."

My opinion:

"The only way to convey accurate

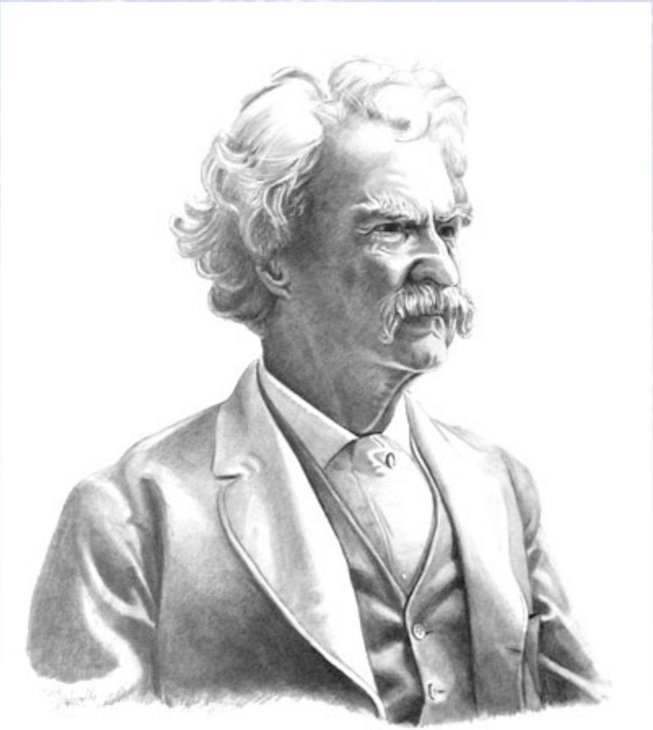
information is by statistics."



Mark Twain:

"There are three kinds of lies:

lies, damned lies, and statistics."



My opinion:

"The only way to convey accurate

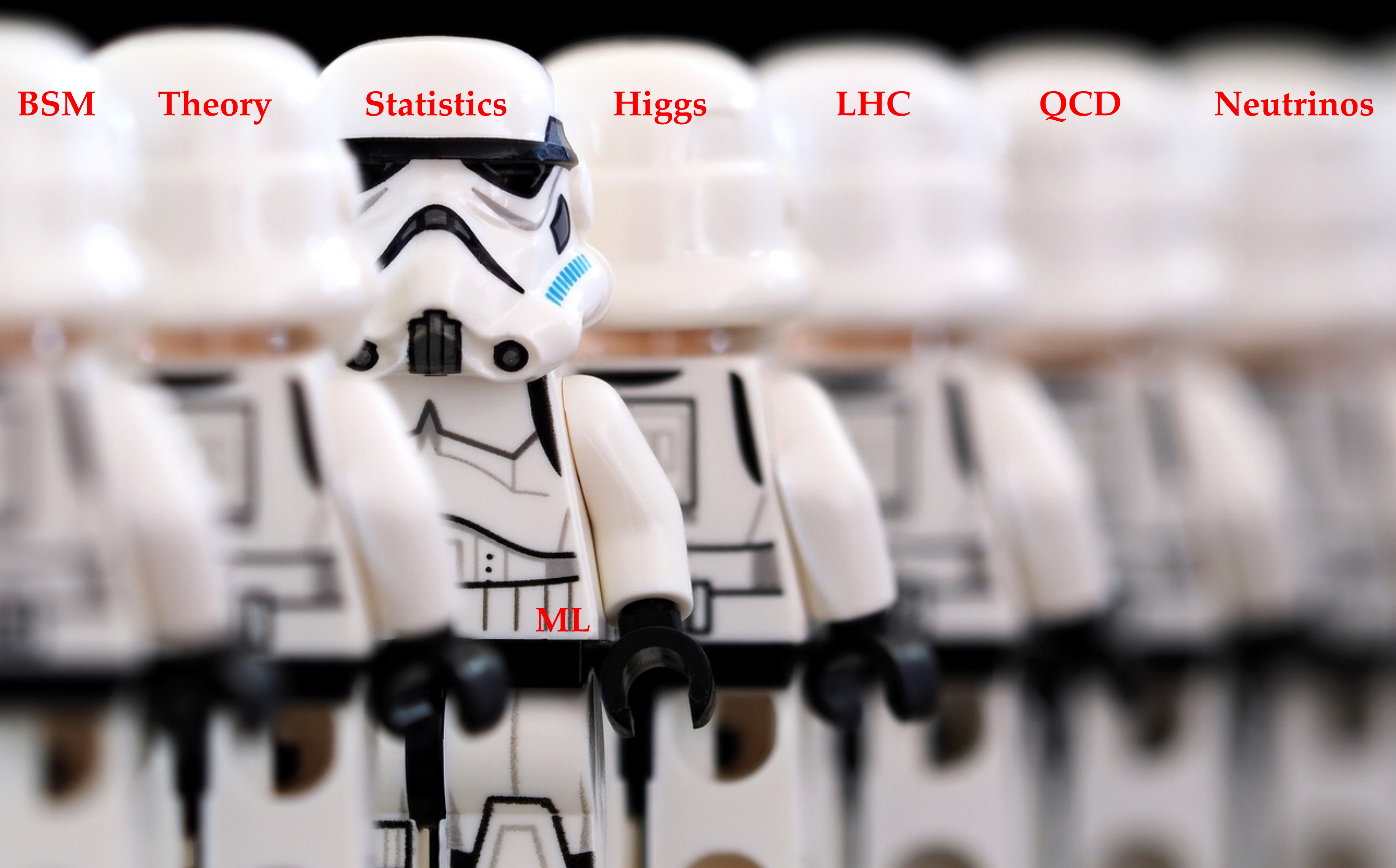
information is by statistics."

Hal Varian [Chief economist of Google]:

"I keep saying the sexy job in the next ten

years will be statisticians."

Why statistics?



Why statistics?

BSM Theory Statistics Higgs LHC QCD Neutrinos

Because you will need it!

(and maybe even like it)

ML

Central Limit Theorem

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 RMSEs 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 RMSEs 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!!!

by the 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.

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

Central Limit Theorem

<u>Central Limit Theorem:</u>

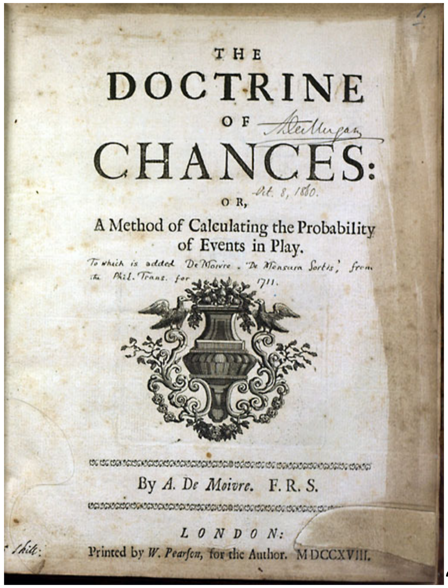
The sum of N *independent* continuous random variables x_i with means μ_i and variances σ_i^2 becomes a Gaussian random variable with mean $\mu = \Sigma_i \mu_i$ and variance $\sigma^2 = \Sigma_i \sigma_i^2$ in the limit that N approaches infinity.

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..**



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.

"The epistemological value of probability theory is based on the fact that chance phenomena, considered collectively and on a grand scale, create non-random regularity."

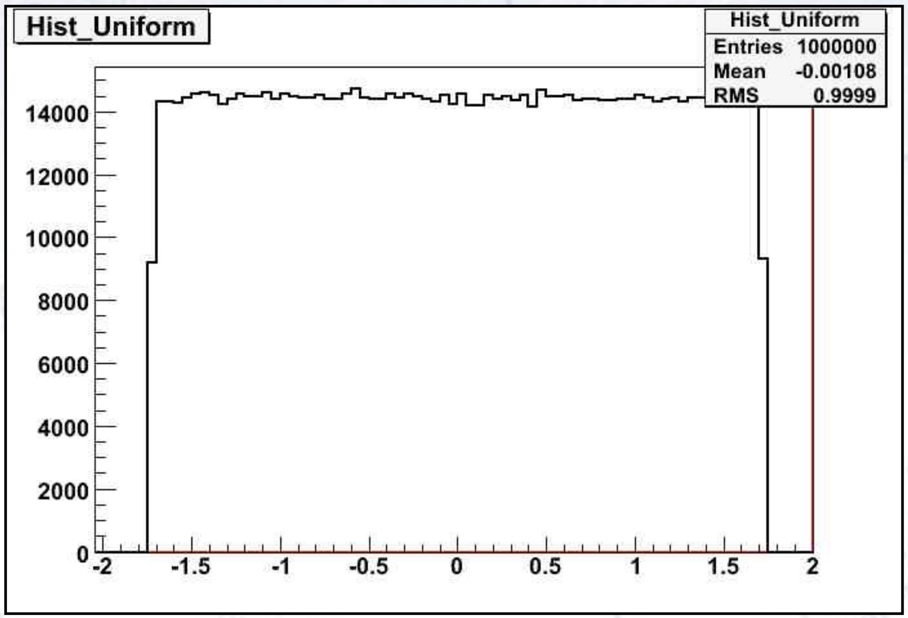
[Andrey Kolmogorov, Soviet mathematician, 1954]

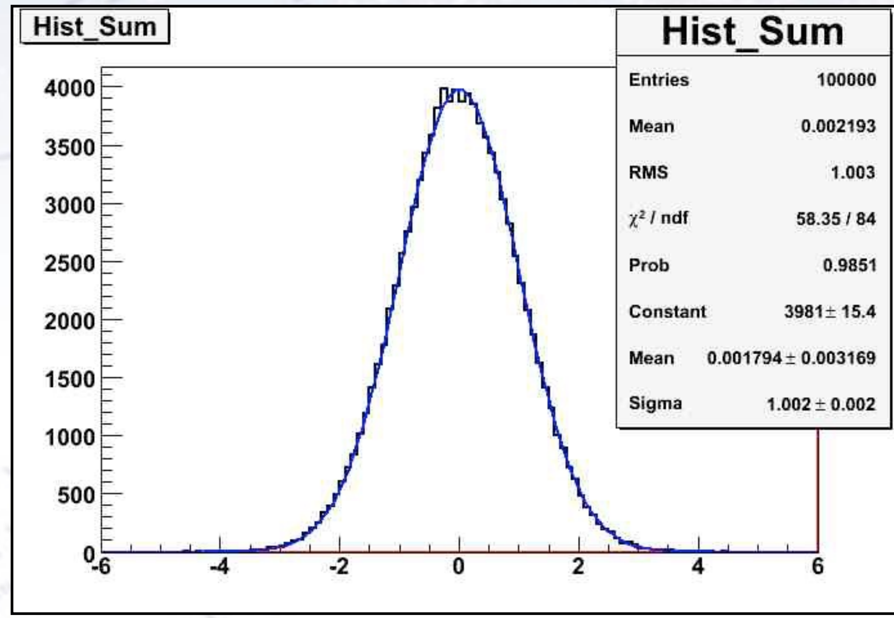
"Nowadays, the central limit theorem is considered to be the unofficial sovereign of probability theory."

[Henk Tijms, Dutch mathematician 2004]

Example of Central Limit Theorem

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



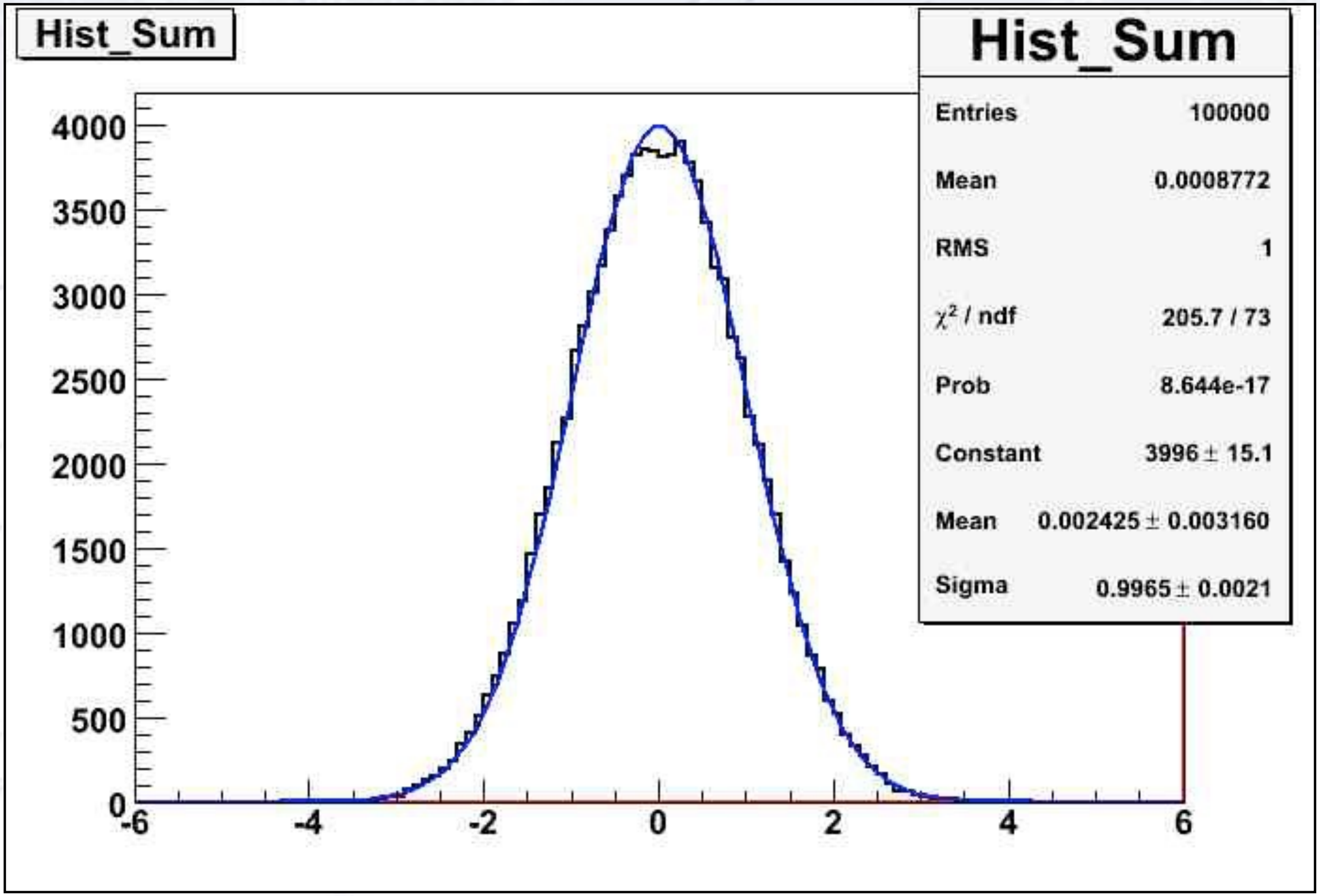


The result is a bell shaped curve, a so-called **normal** or **Gaussian** distribution.

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

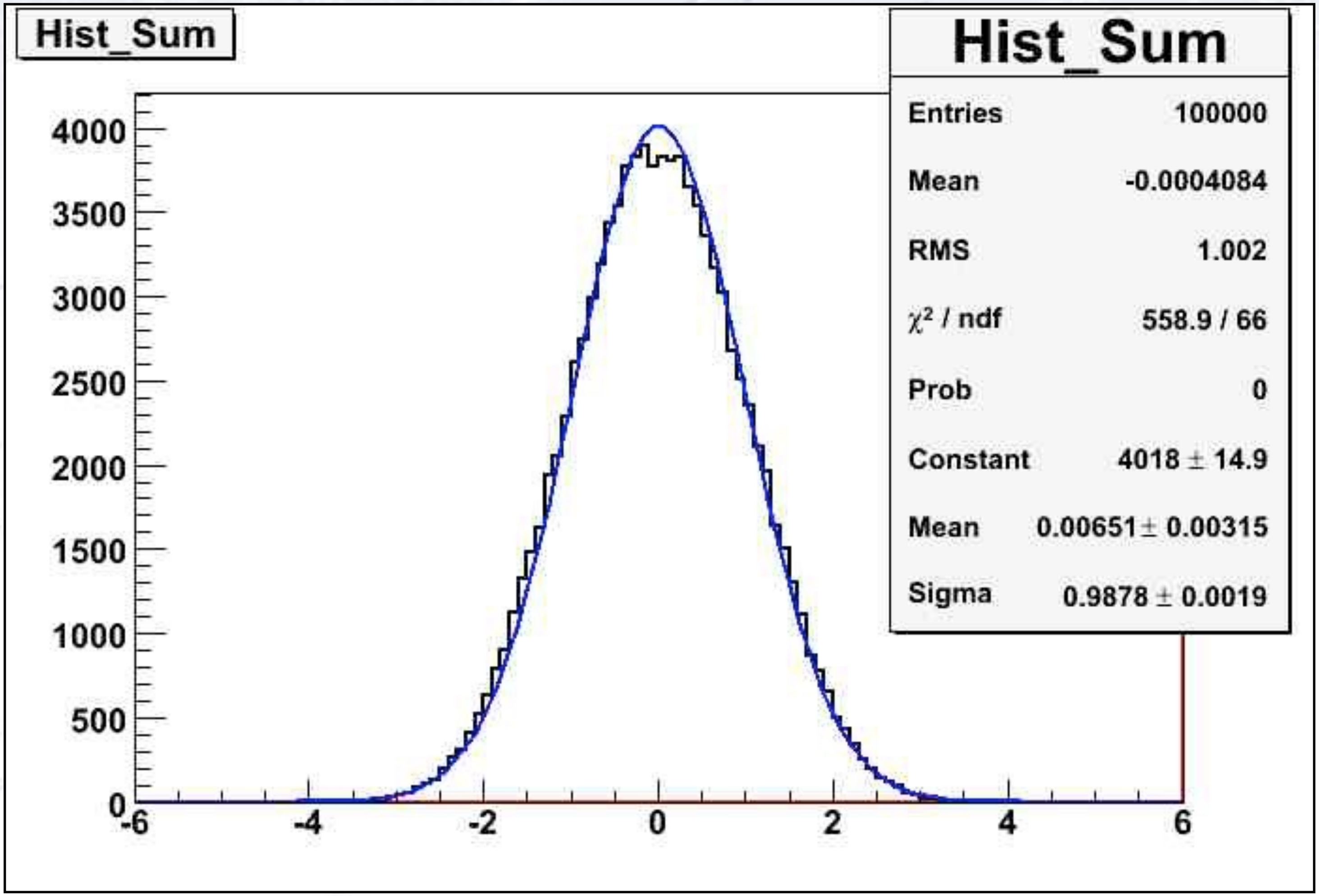
Example of Central Limit Theorem

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



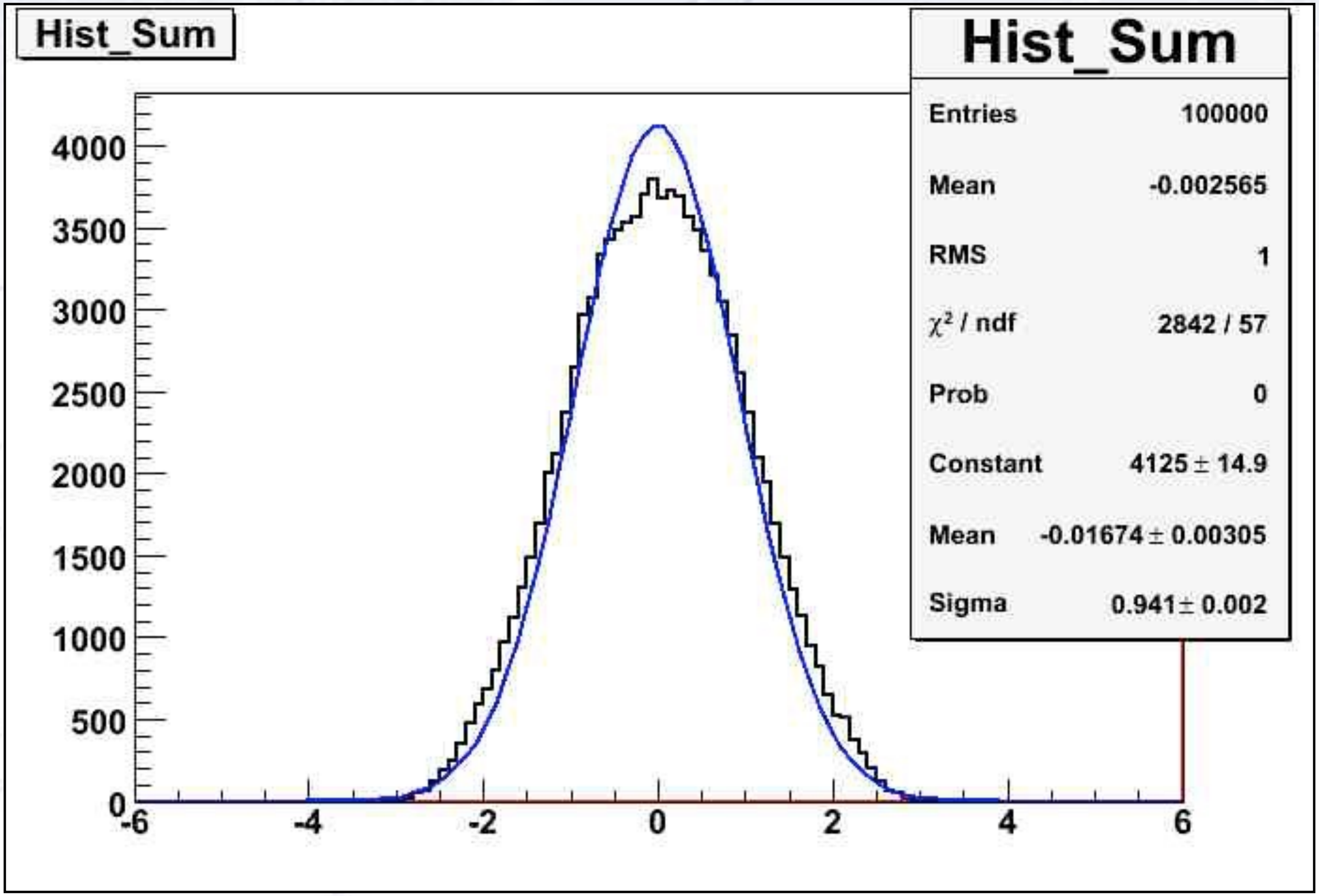
Example of Central Limit Theorem

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



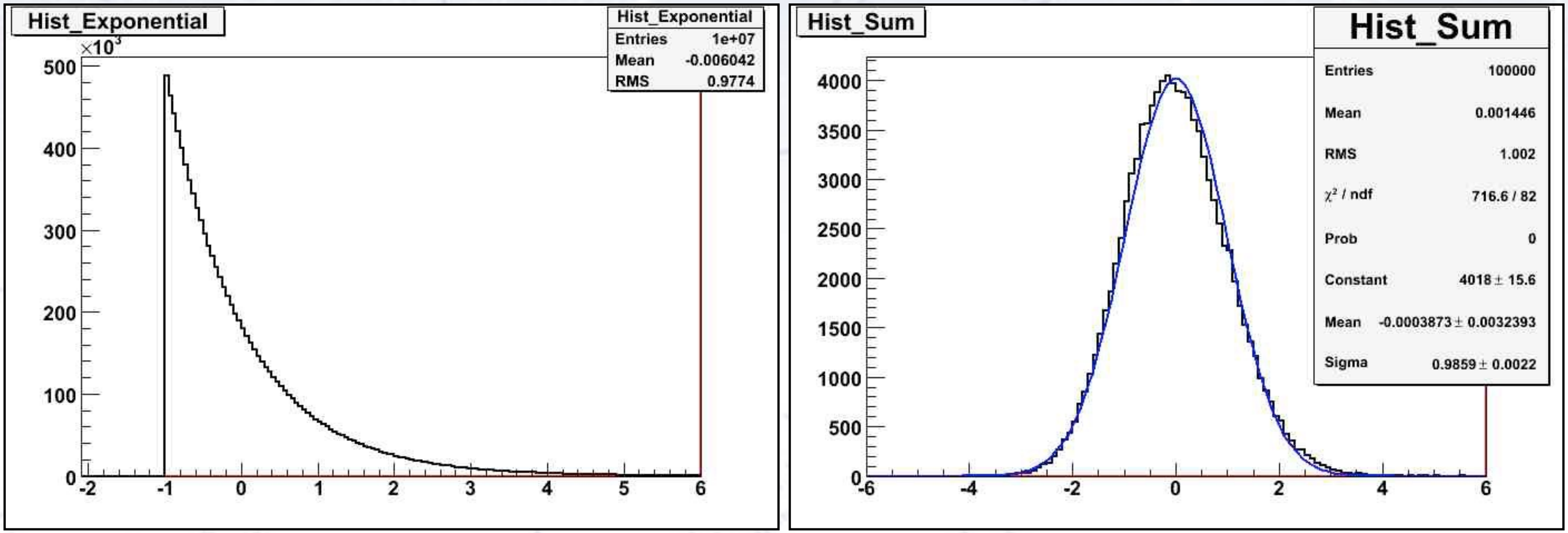
Example of Central Limit Theorem

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…



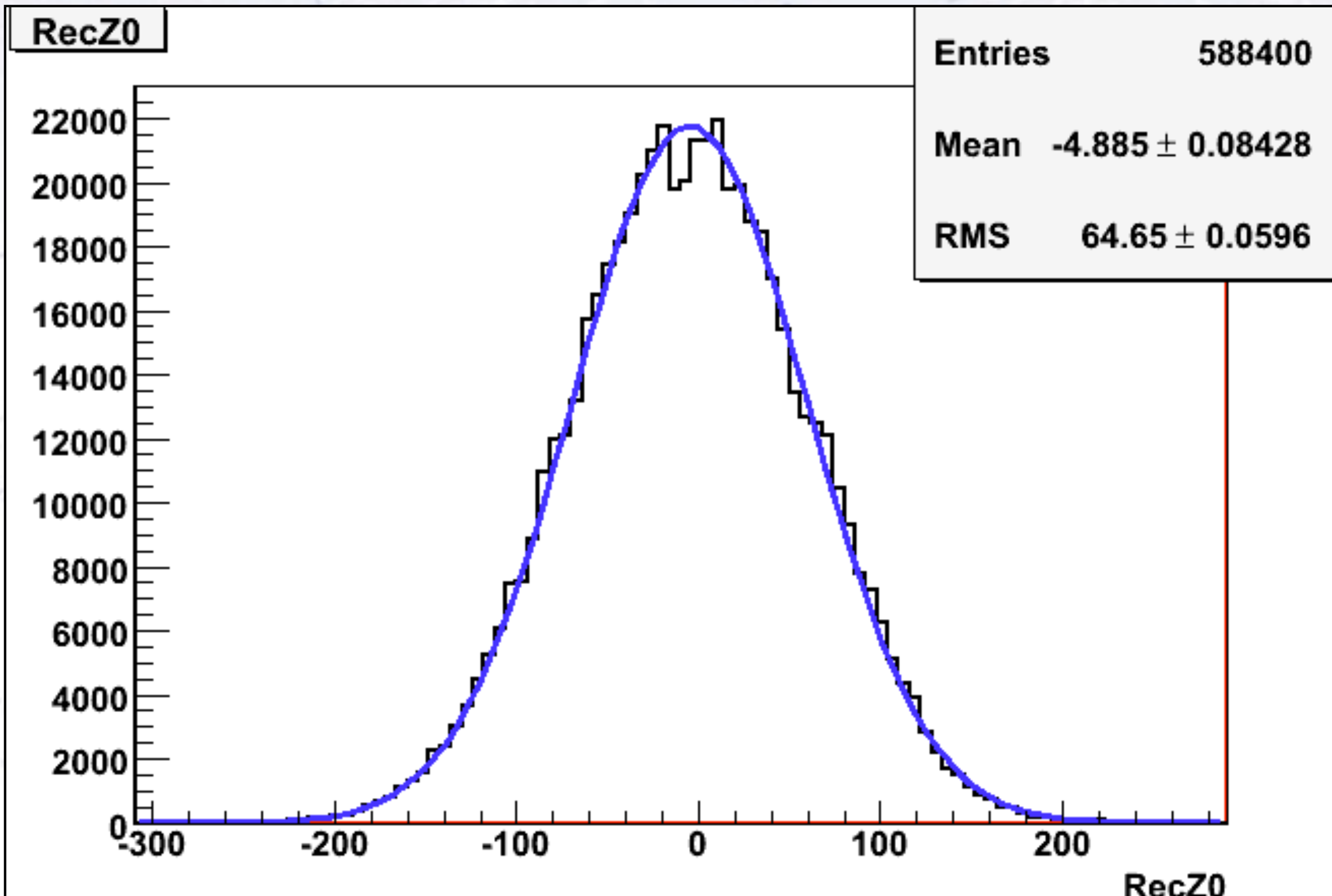
It doesn't matter what shape the input PDF has, as long as it has finite mean and width, which all numbers from the real world has! Sum quickly becomes:

Gaussian!!!

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

Example of Central Limit Theorem

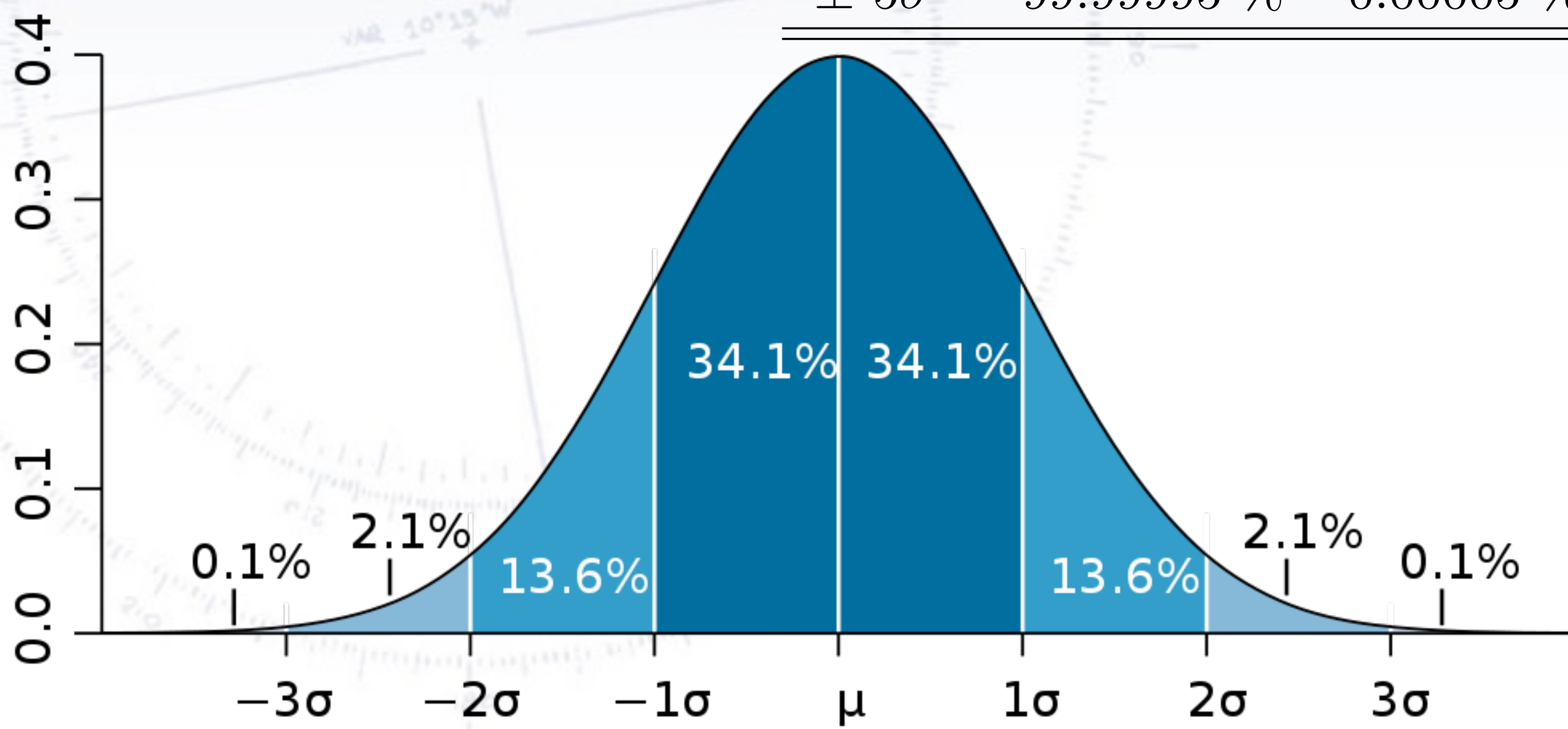
Looking at z-coordinate of tracks at vertex from proton collisions in CERNs LHC accelerator by the ATLAS detector, this is what you get:



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 %



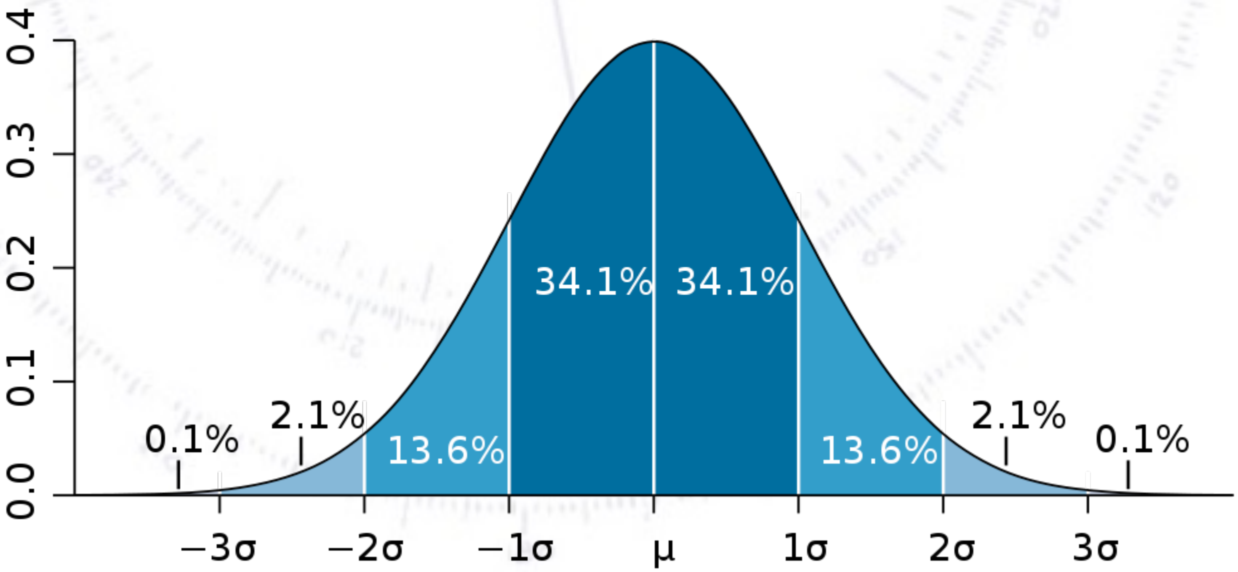
Summary

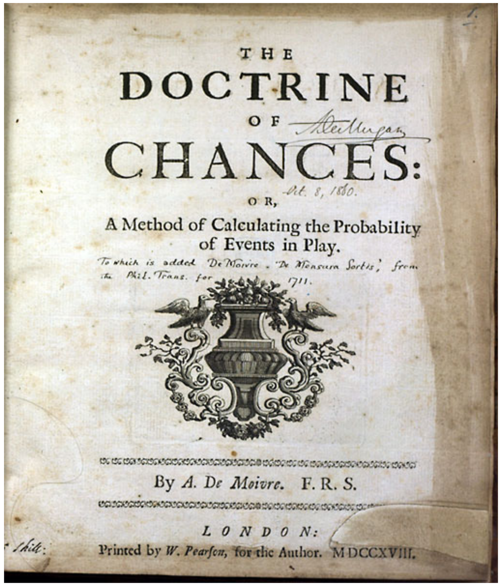
The Central Limit Theorem

...is your good friend because it…

ensures that uncertainties tend to be Gaussian

…which are the easiest to work with!

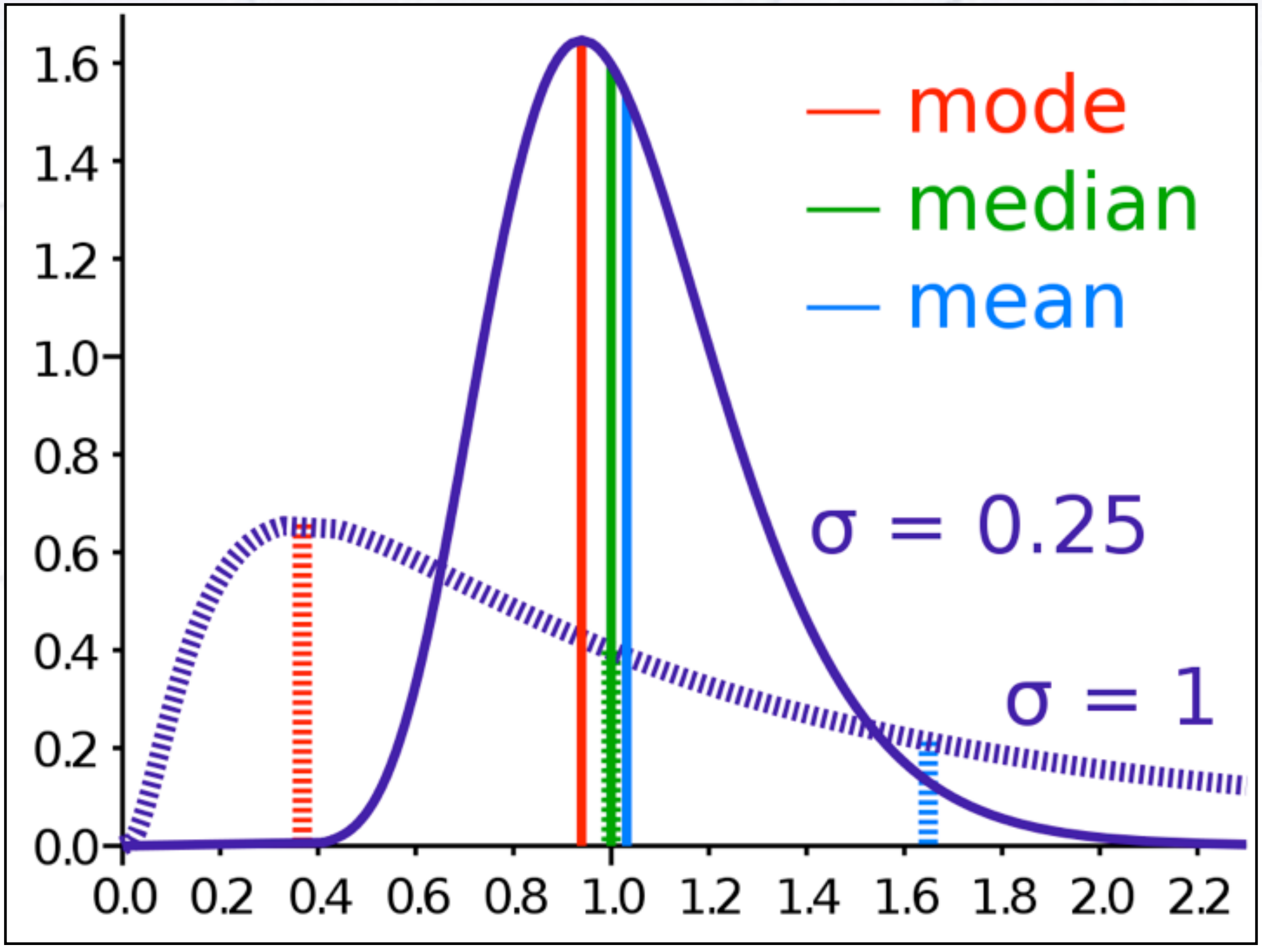




Estimators

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):

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

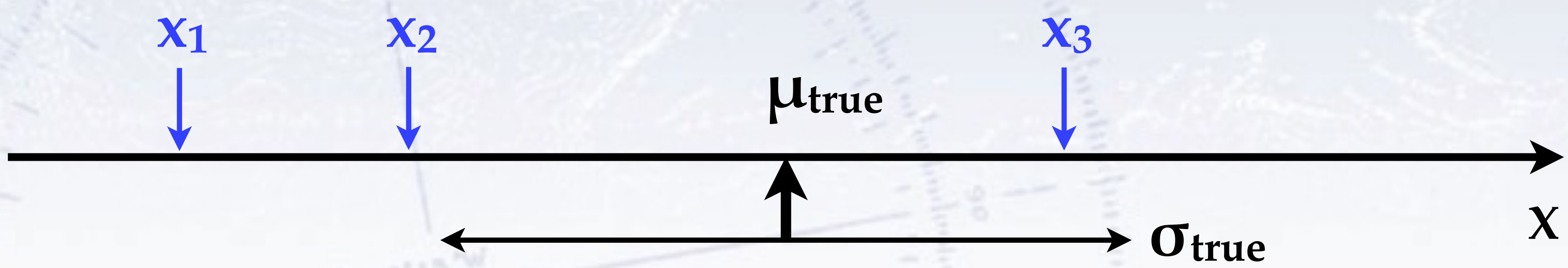
For the **standard deviation (Std)**, a.k.a. **width** or **RMSE**, it is:

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

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

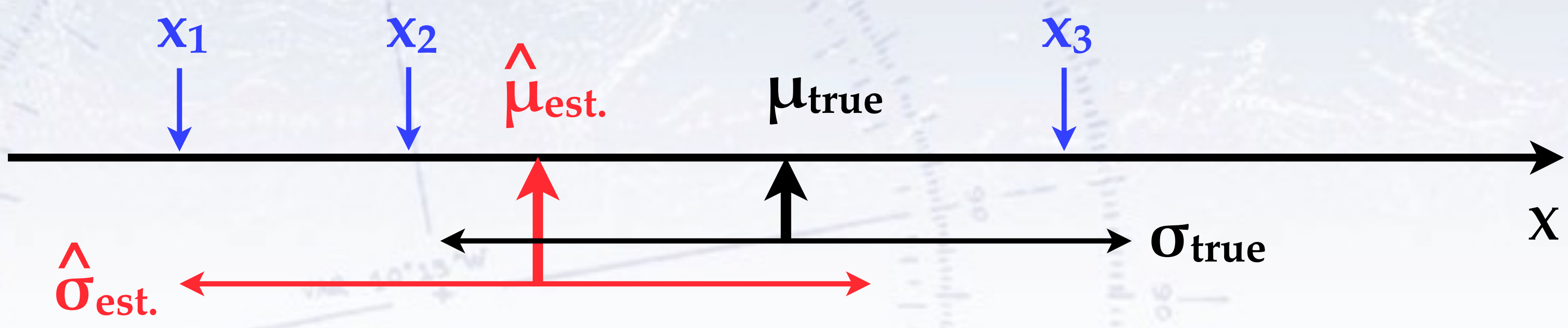
Why not "just" the naive SD?

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



Why not "just" the naive SD?

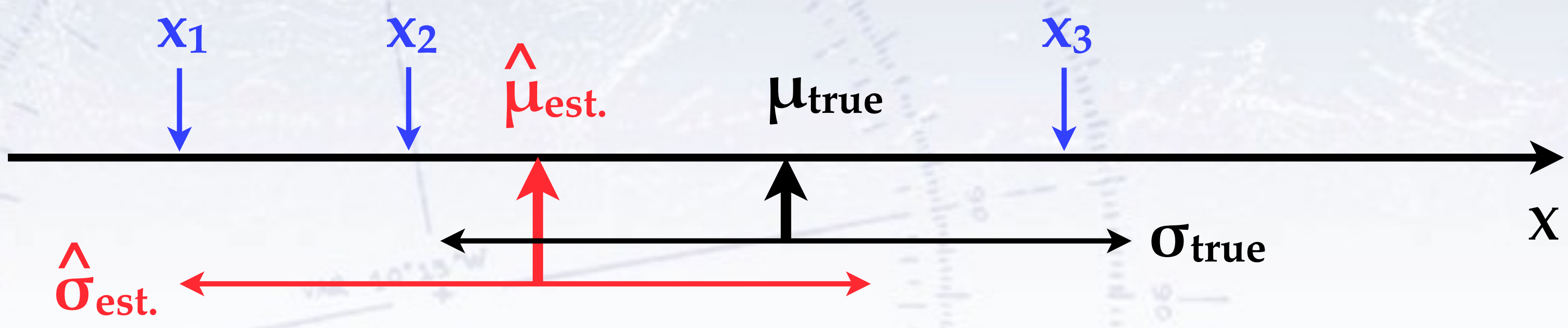
Imagine taking 3 independent measurements, then estimating mean and SD:



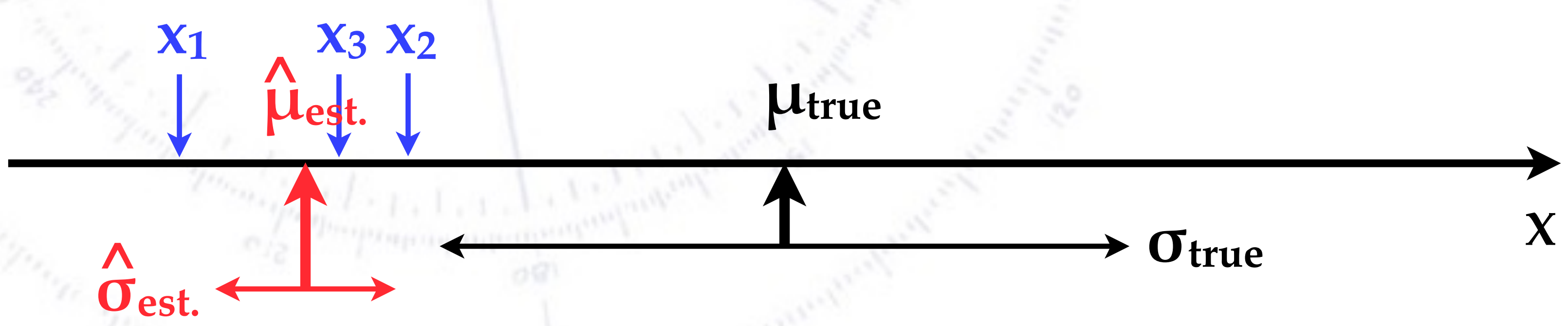
Above, all went well, because measurements were nicely distributed on both sides of the mean, and spread out according to SD.

Why not "just" the naive SD?

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



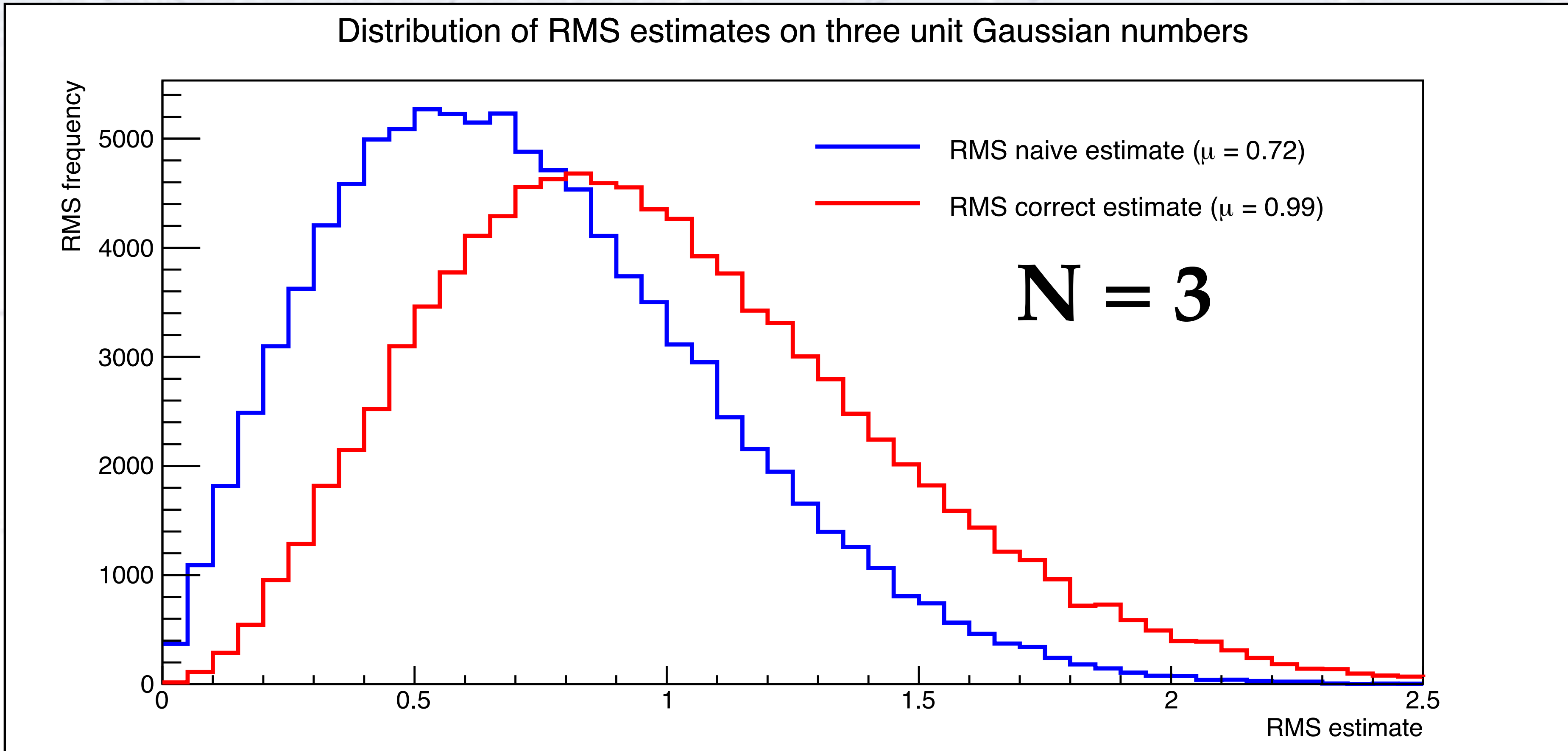
Above, all went well, because measurements were nicely distributed on both sides of the mean, and spread out according to SD.



However, now the mean is off and the Std way off (terribly so!).
If we had used the true mean in the formula, it would have been less of a problem.

How incorrect is the naive SD?

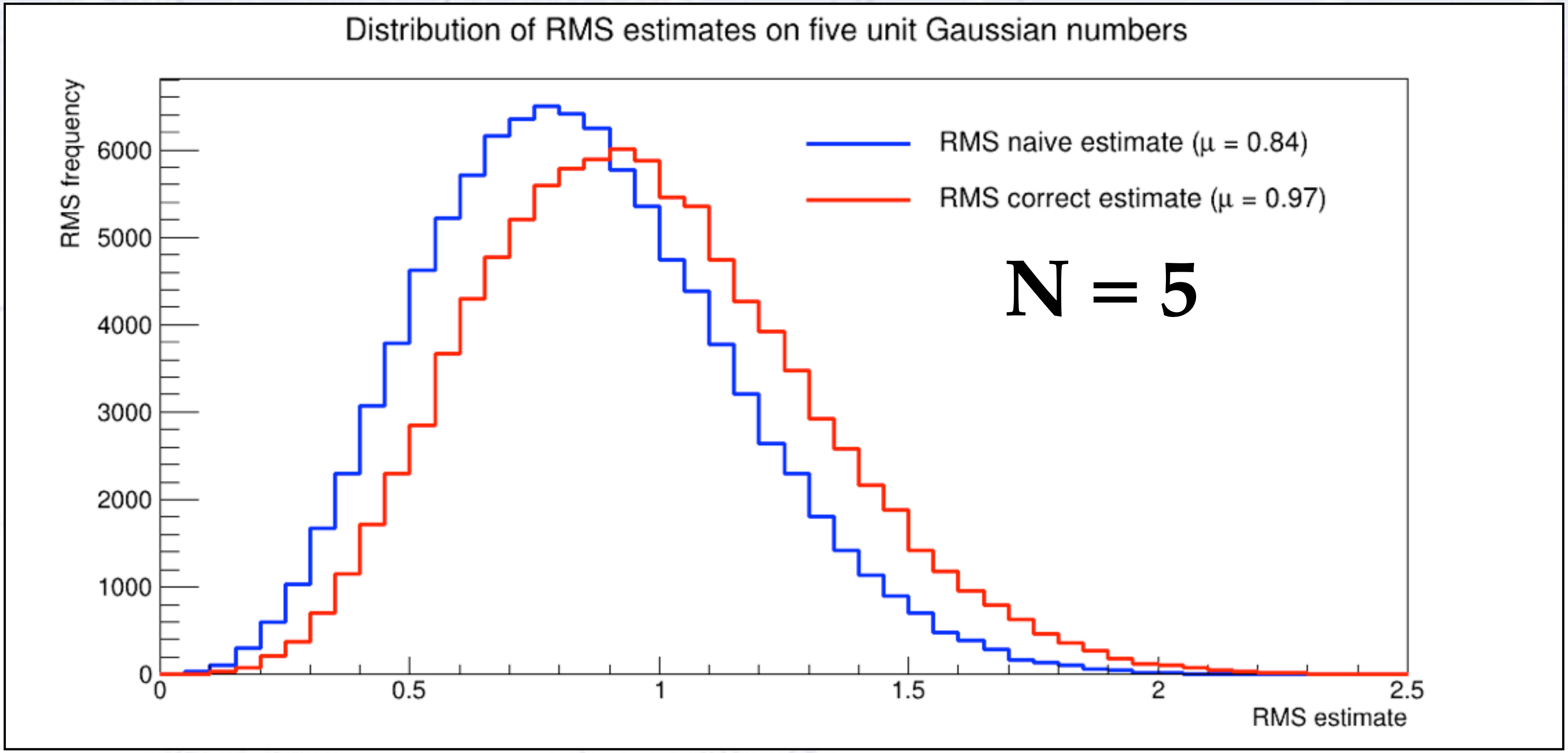
Such questions can most easily be answered by a small simulation…
Produce N=3 numbers from a unit Gaussian, and calculate the SD estimate:



So, the "naive" SD underestimates the uncertainty significantly…

How incorrect is the naive SD?

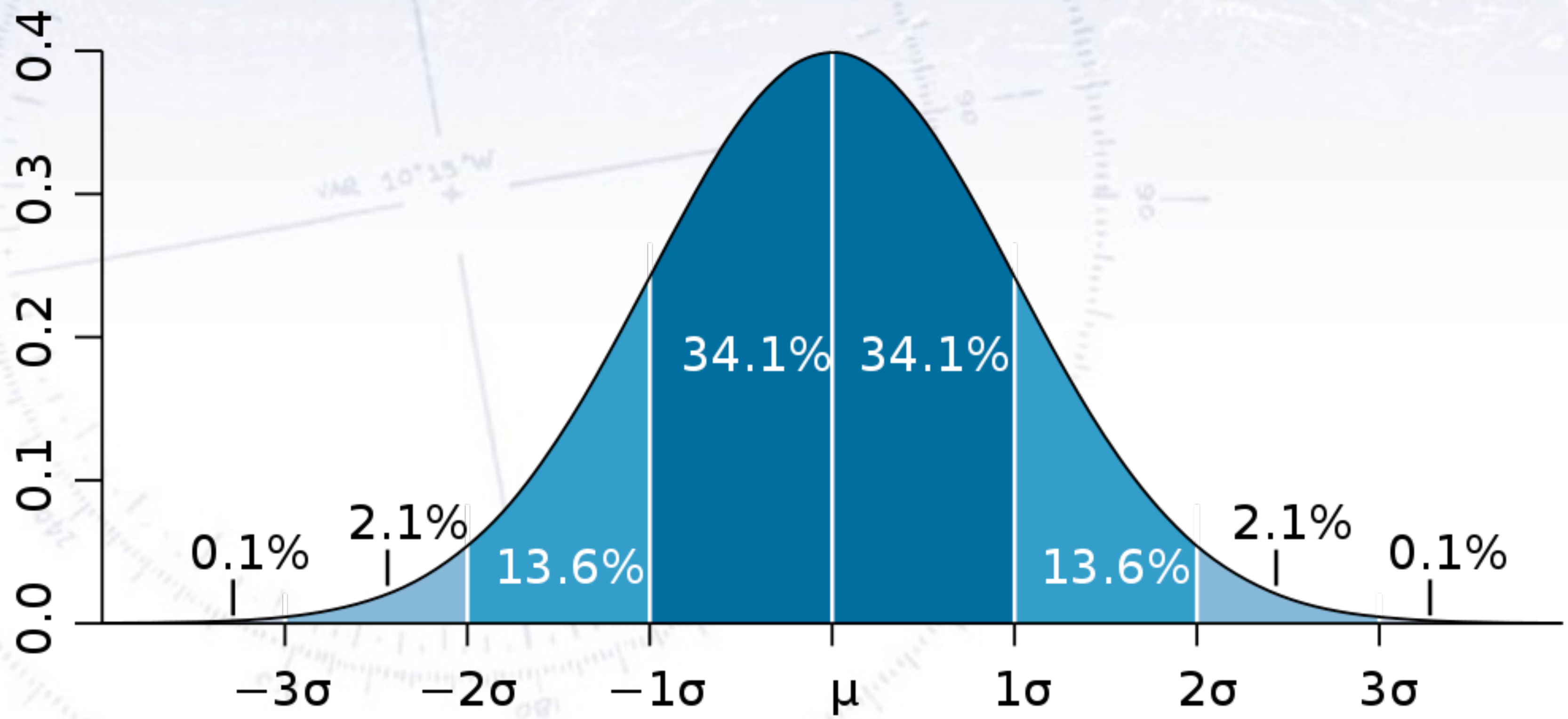
Such questions can most easily be answered by a small simulation…
Produce N=5 numbers from a unit Gaussian, and calculate the SD estimate:



Here, the "naive" SD underestimates the uncertainty a bit...

SD and Gaussian σ relation

When a distribution is Gaussian, **the Std. 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?

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

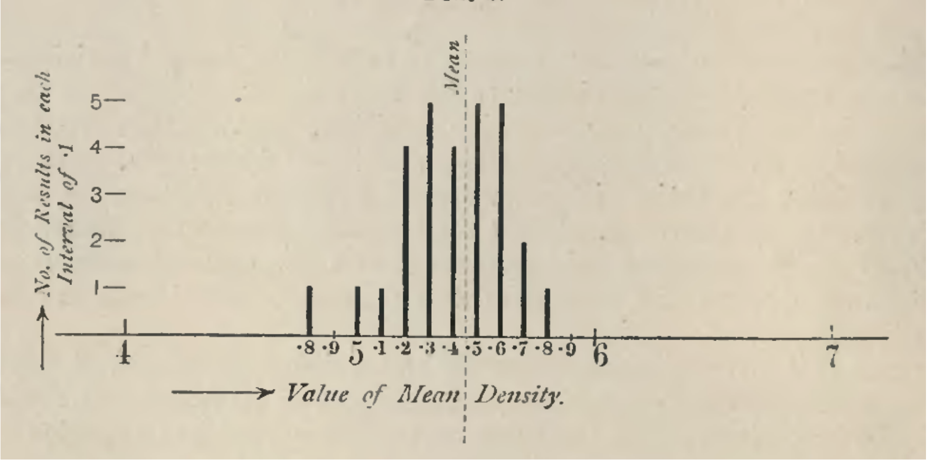
Mean and Width

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

FIG. 3.



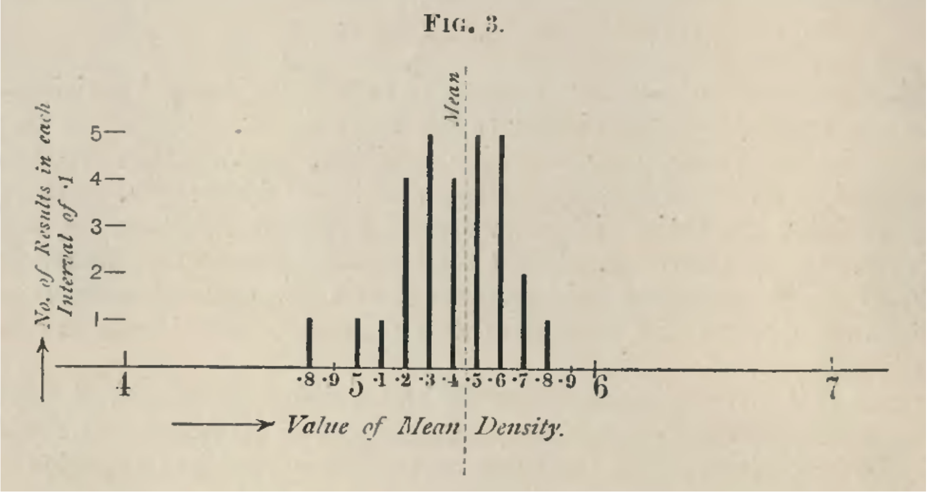
Mean and Width

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

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

Please, commit to memory now!

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

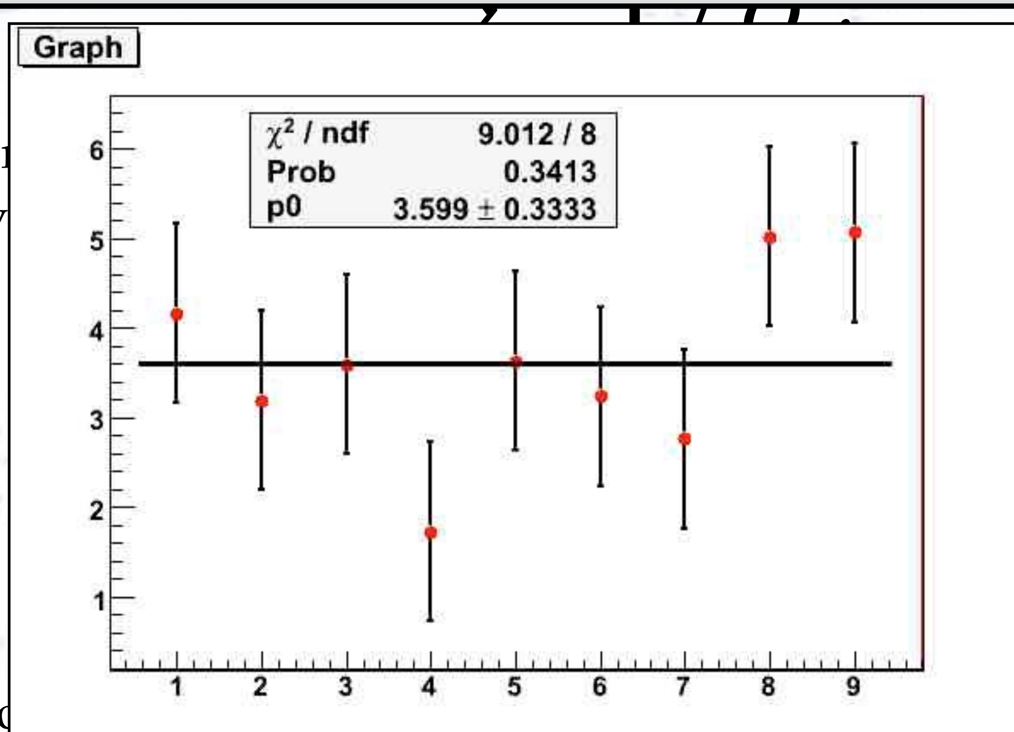
What if we are given data, which has different uncertainties?

How to aver

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 measureme … gful SD!

The uncertainty



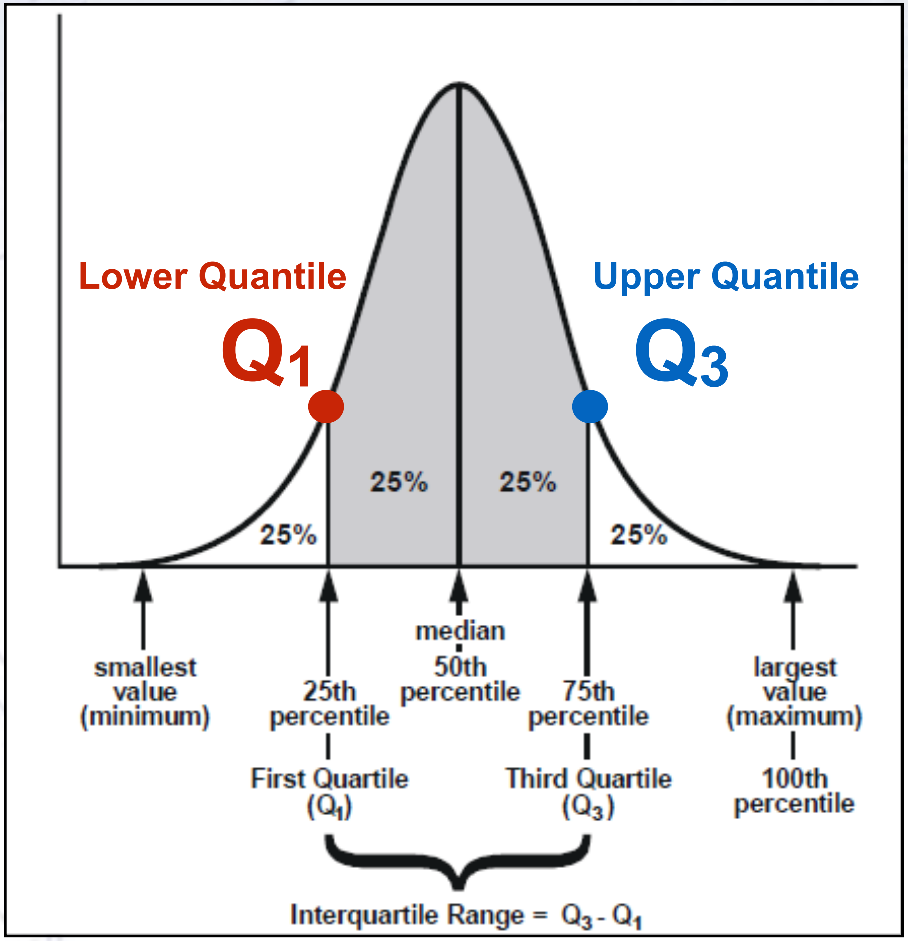
Can be underst … easurements

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

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



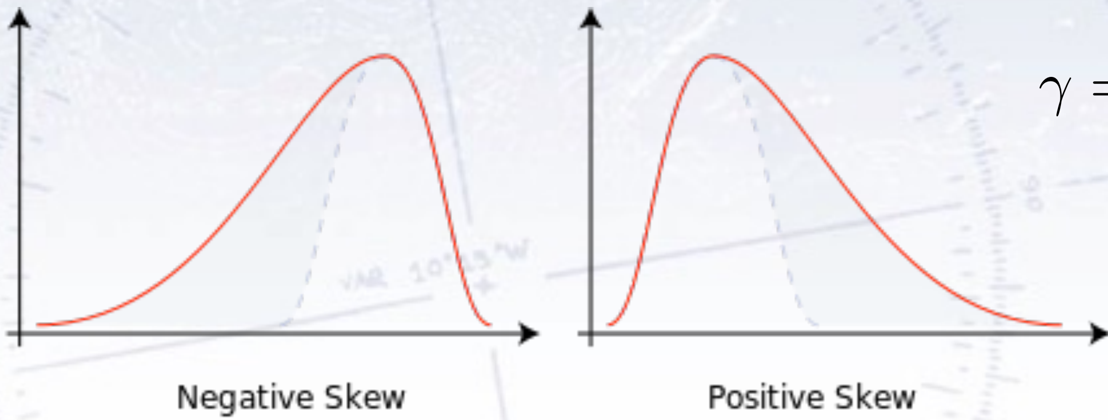
The InterQuantile Efficiency (IQE) is defined as:

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

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

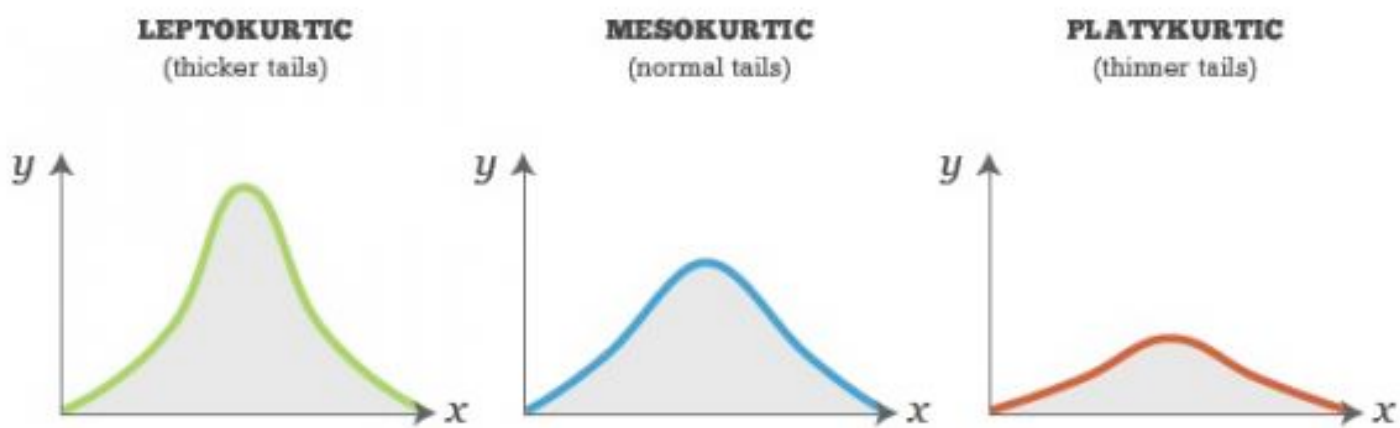
Skewness and Kurtosis

Higher moments reveal something about a distributions asymmetry and tails:

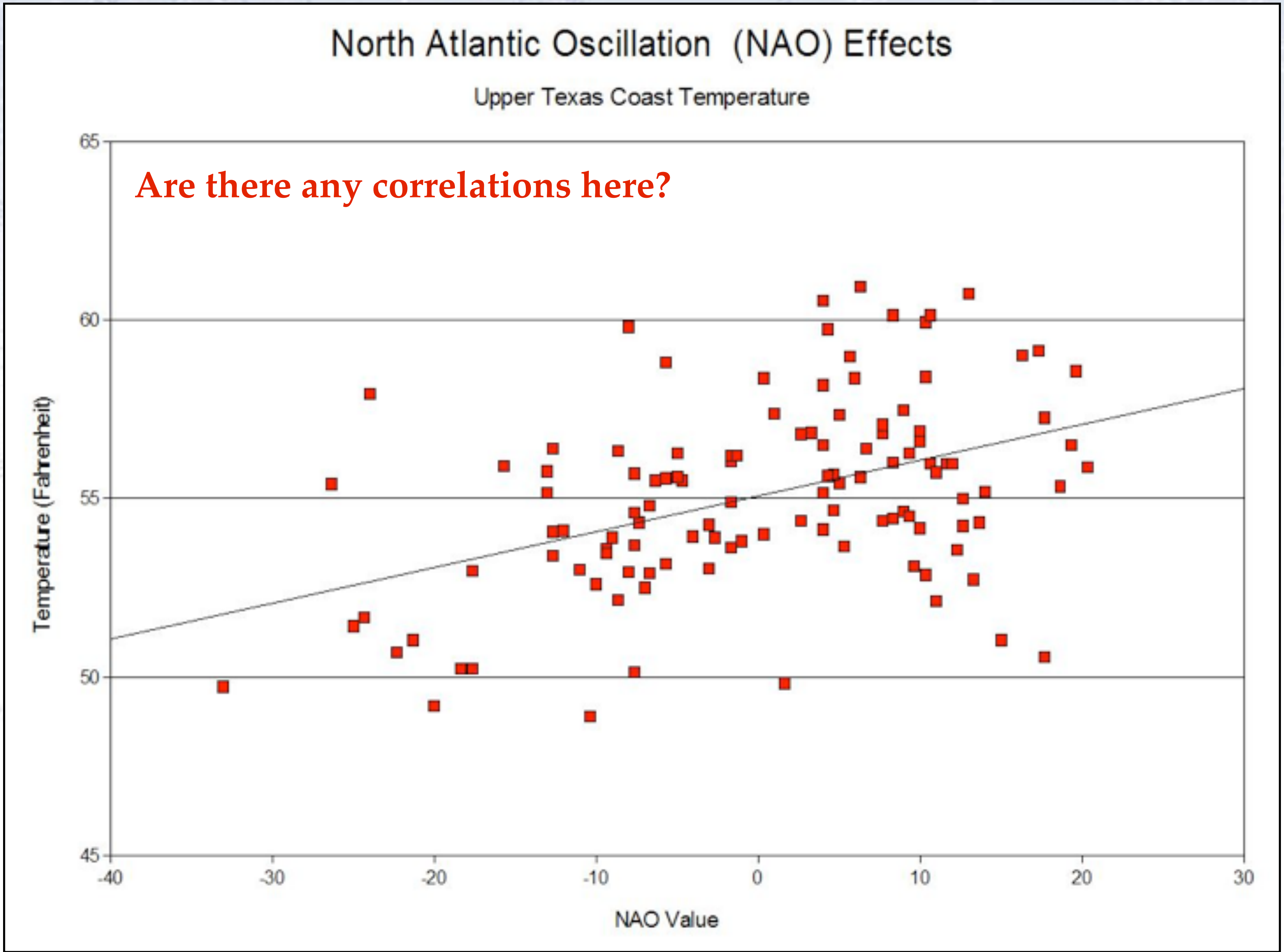


$$\gamma = \frac{\frac{1}{N}\sum_i (x_i - \bar{x})^3}{(\frac{1}{N}\sum_i (x_i - \bar{x})^2)^3/2}$$

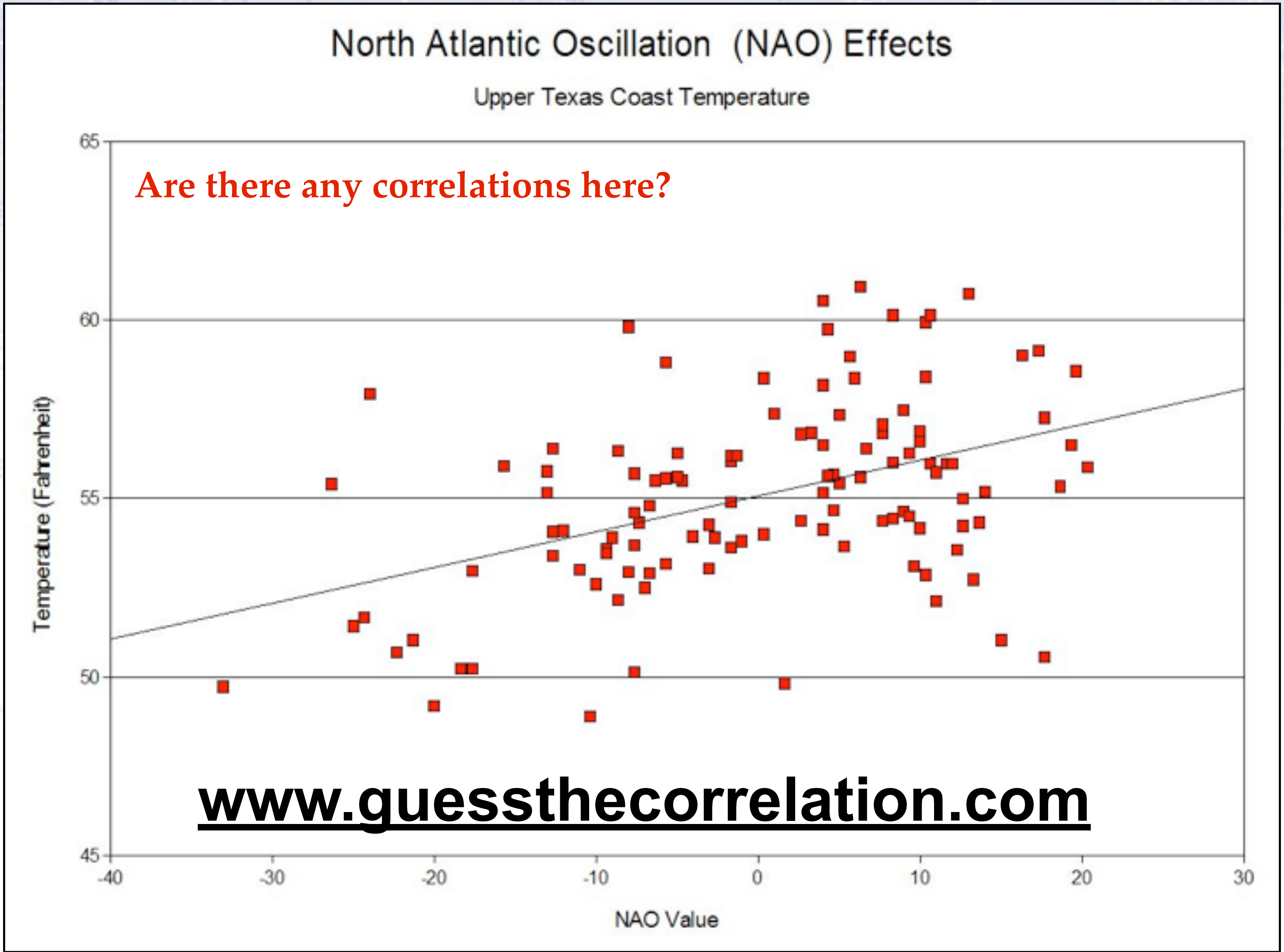
$$\kappa = \frac{\frac{1}{N}\sum_i (x_i - \bar{x})^4}{(\frac{1}{N}\sum_i (x_i - \bar{x})^2)^2} - 3$$



Correlation

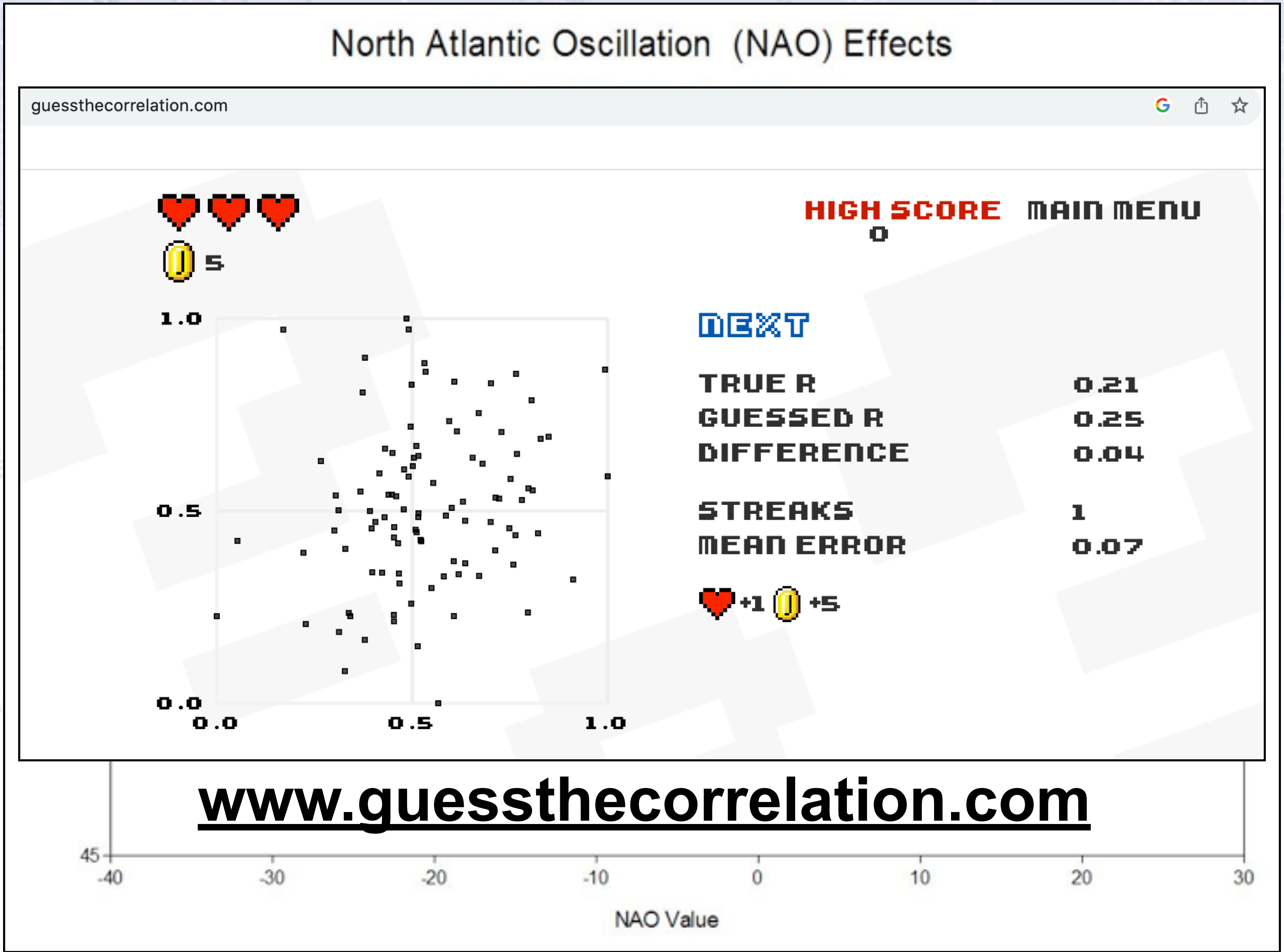


Correlation



www.guessthecorrelation.com

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, 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, 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:

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

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

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

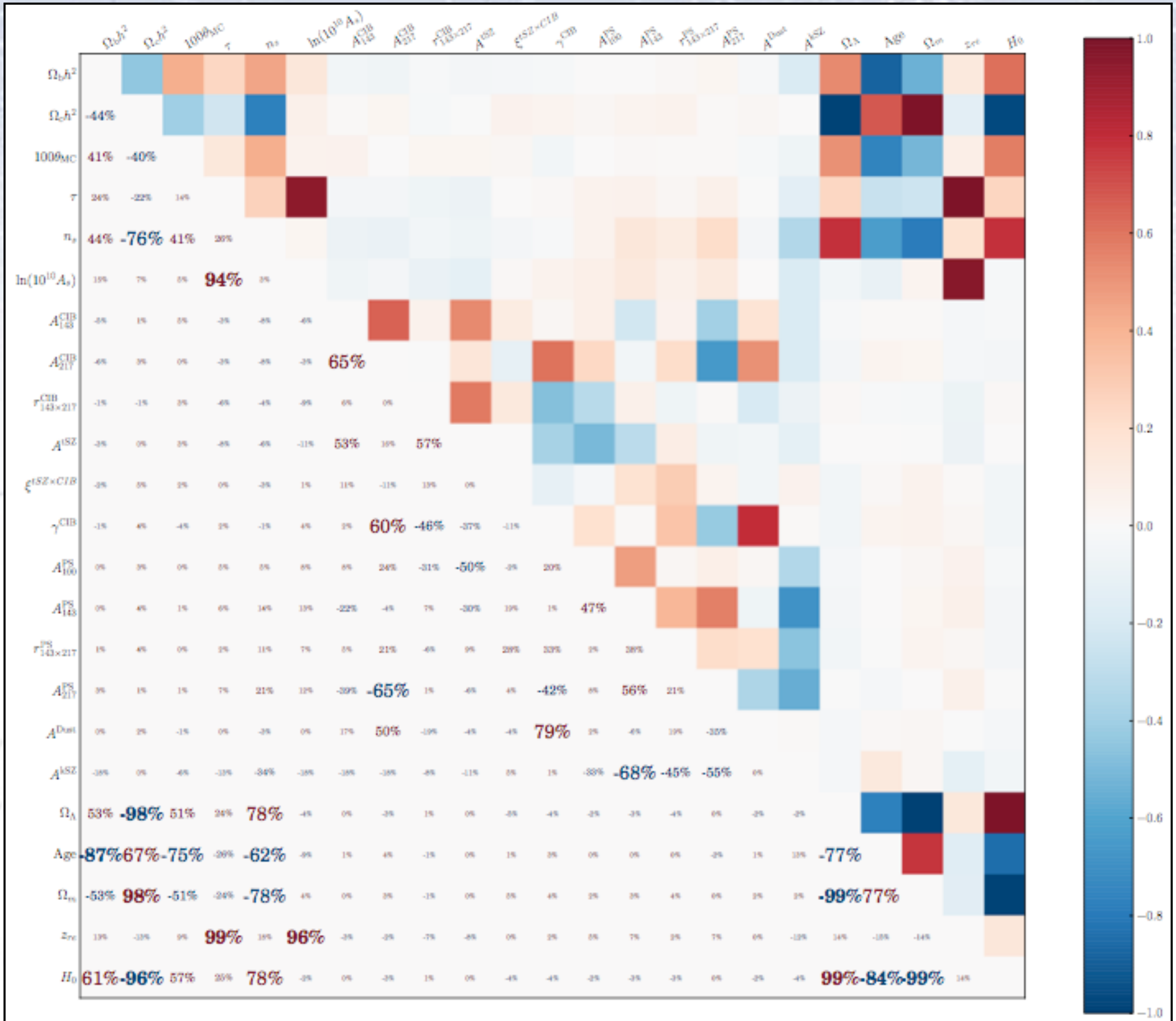
Correlation Matrix

The correlation matrix V_{xy} explicitly looks as:

$$V_{xy} = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \cdots & \sigma_{1N}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2N}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_N^2 & \sigma_{N2}^2 & \cdots & \sigma_{NN}^2 \end{bmatrix}$$

The variance of variables can be found along the diagonal, while the (symmetric) off-diagonal terms show the co-variances.

Planck example



Correlation and Information

Correlations influence results in complex ways!

They need to be taken into account, for example in **Error Propagation!**

Correlations may contain a significant amount of information.

We will consider this more when we play with multivariate analysis.

Rank correlations

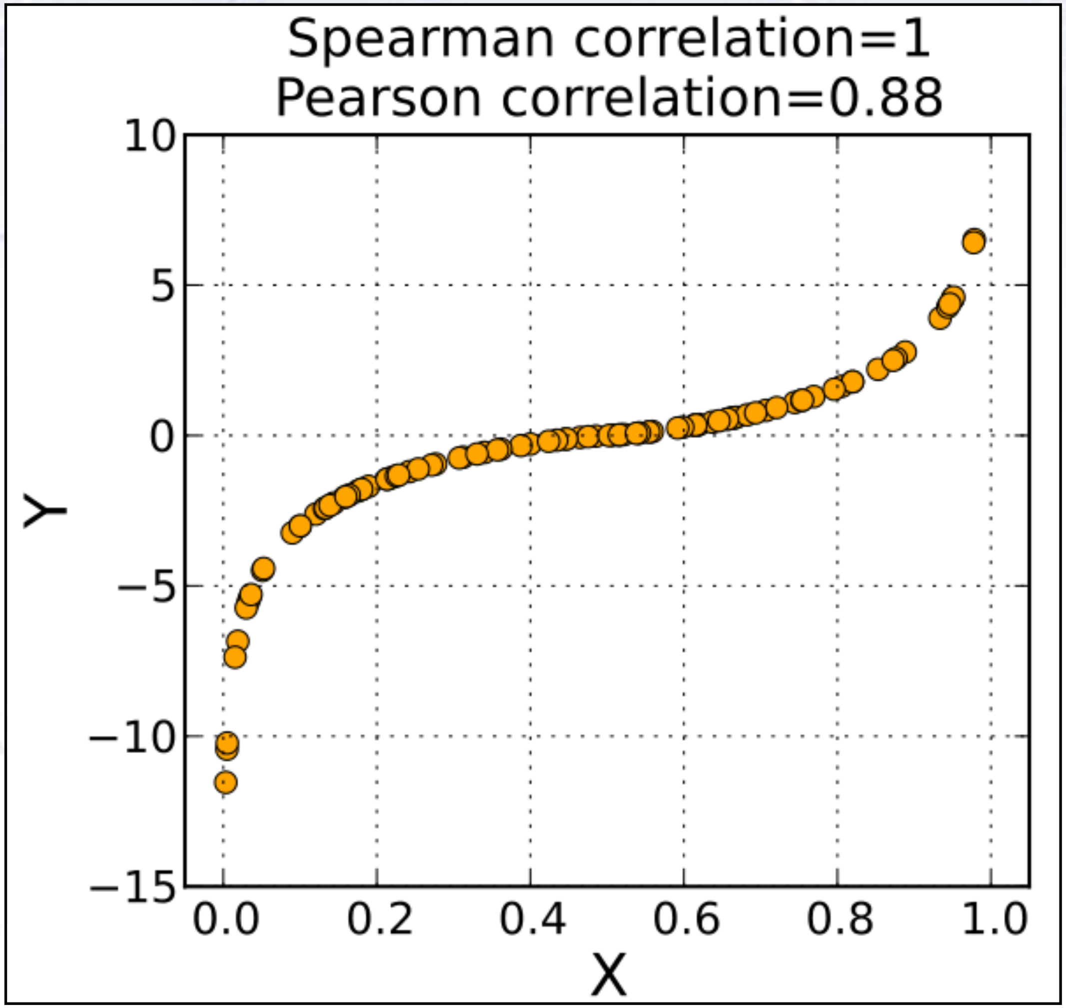
Sometimes, variables are perfectly correlated, just not linearly:

In this case the Pearson correlation is not the best measure.

Rank correlation compares the ranking between the two sets, and therefore gets a good measure of the correlation (see figure).

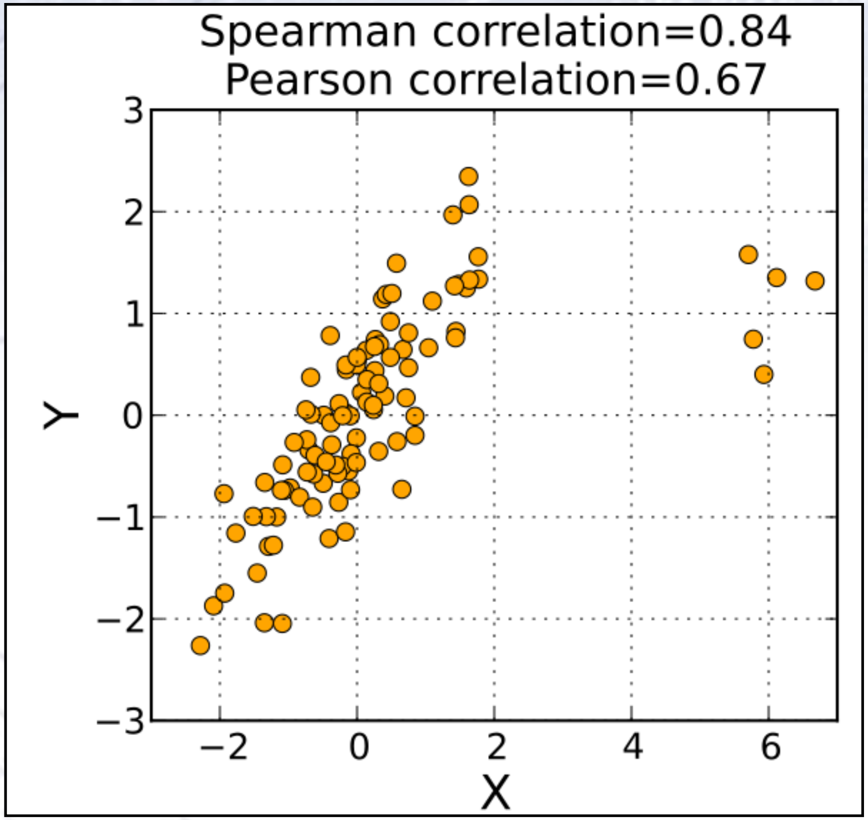
The two main cases of rank correlations are:

- Spearman's rho
- Kendall's tau



Rank correlations

An additional advantage is, that the rank correlation is less sensitive to outliers:



The two rank correlations are special cases of a more general rank correlation.

Typically, Spearman's rank correlation is used.

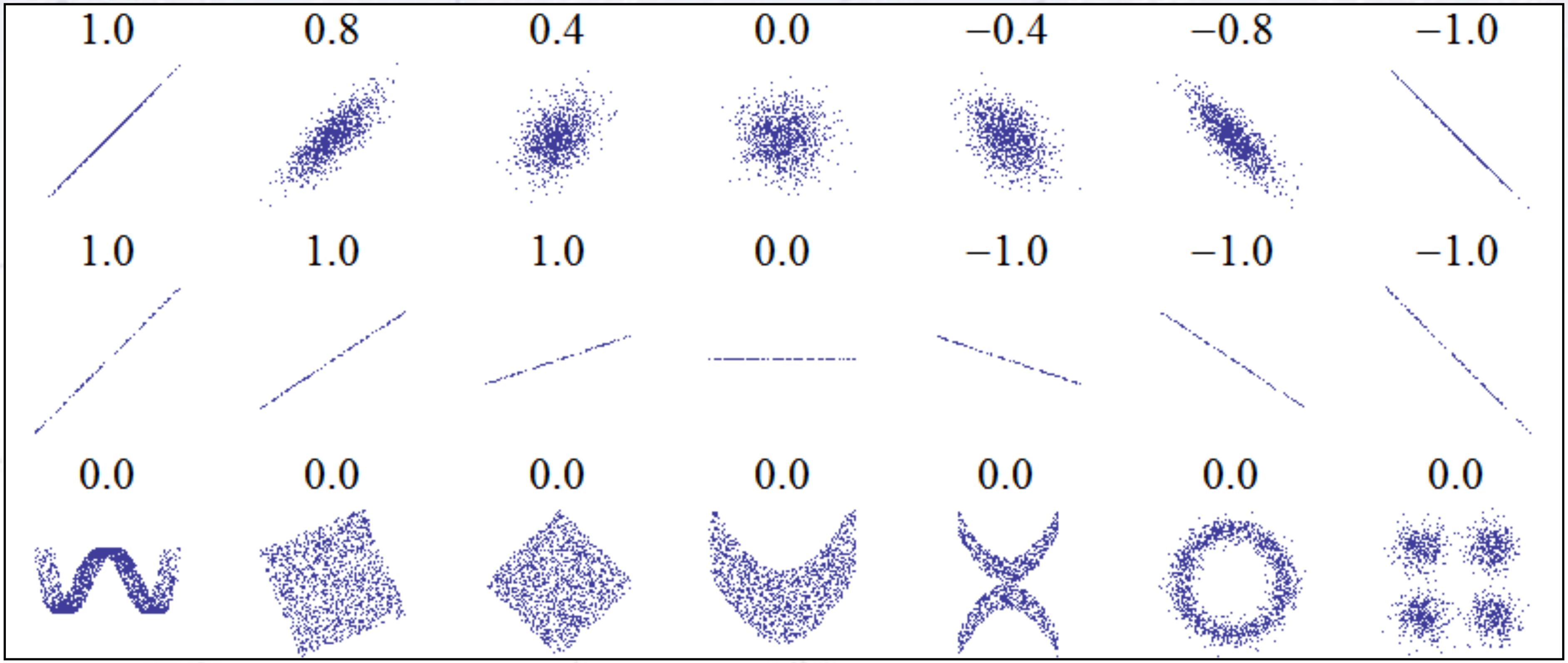
The definition is:

$$\rho = 1 - 6\sum_i (r_i - s_i)^2 / (n^3 - n)$$

where r_i and s_i is the rank of the i'th element.

Correlation

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

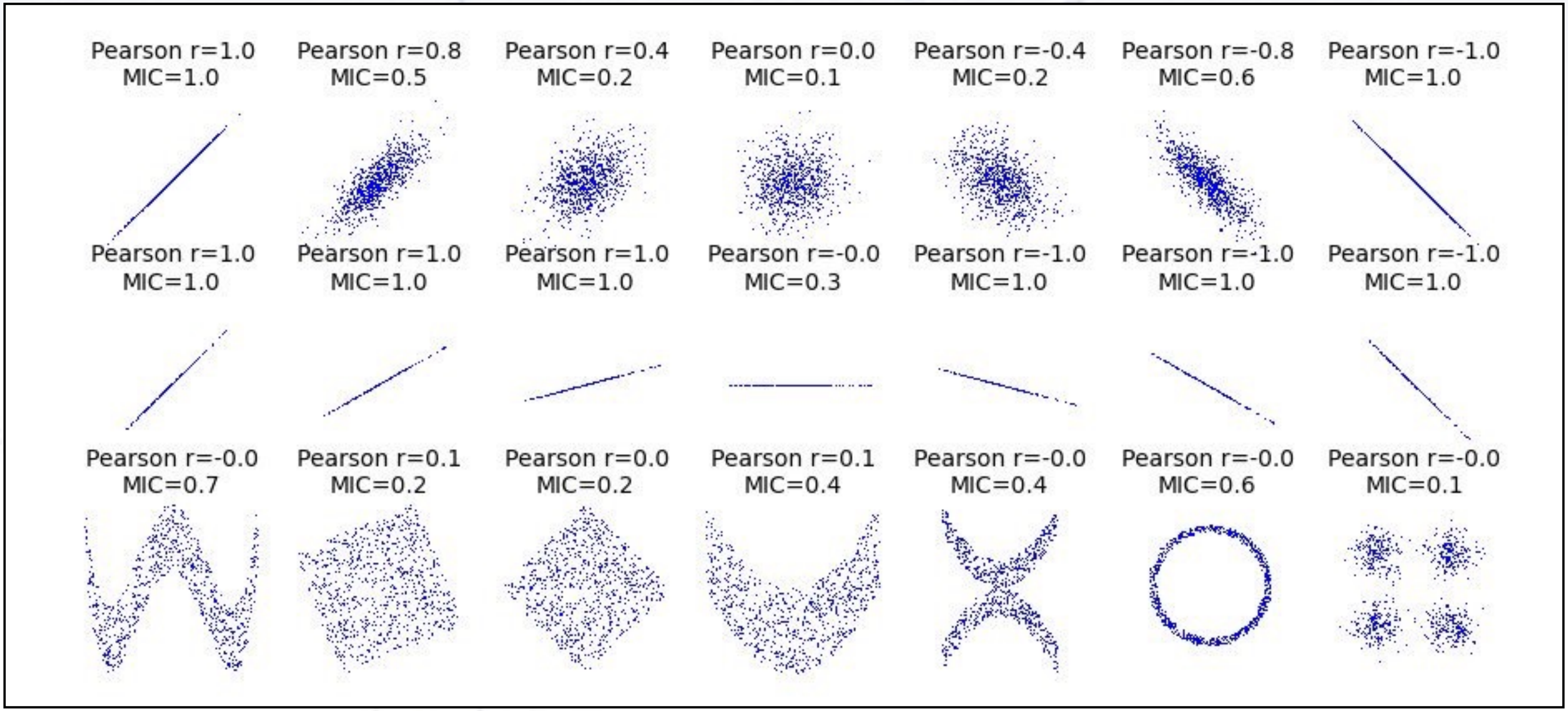


Note how ALL of the bottom distributions have $\varrho = 0$, despite obvious correlations!

Non-linear correlations

Non-linear correlations (associations) are harder to measure, but possible:

- Maximal Information Coefficient (MIC), see reference and Wikipedia on MIC.
- Mutual Information (MI), linked to entropy, see Wikipedia on MI and SKLearn.
- Distance Correlation (DC) between paired vectors, see Wikipedia on DC.



Original paper: "Detecting Novel Associations in Large Data Sets" (2011). Science 334 (6062): 1518–1524.

Correlation Vs. Causation

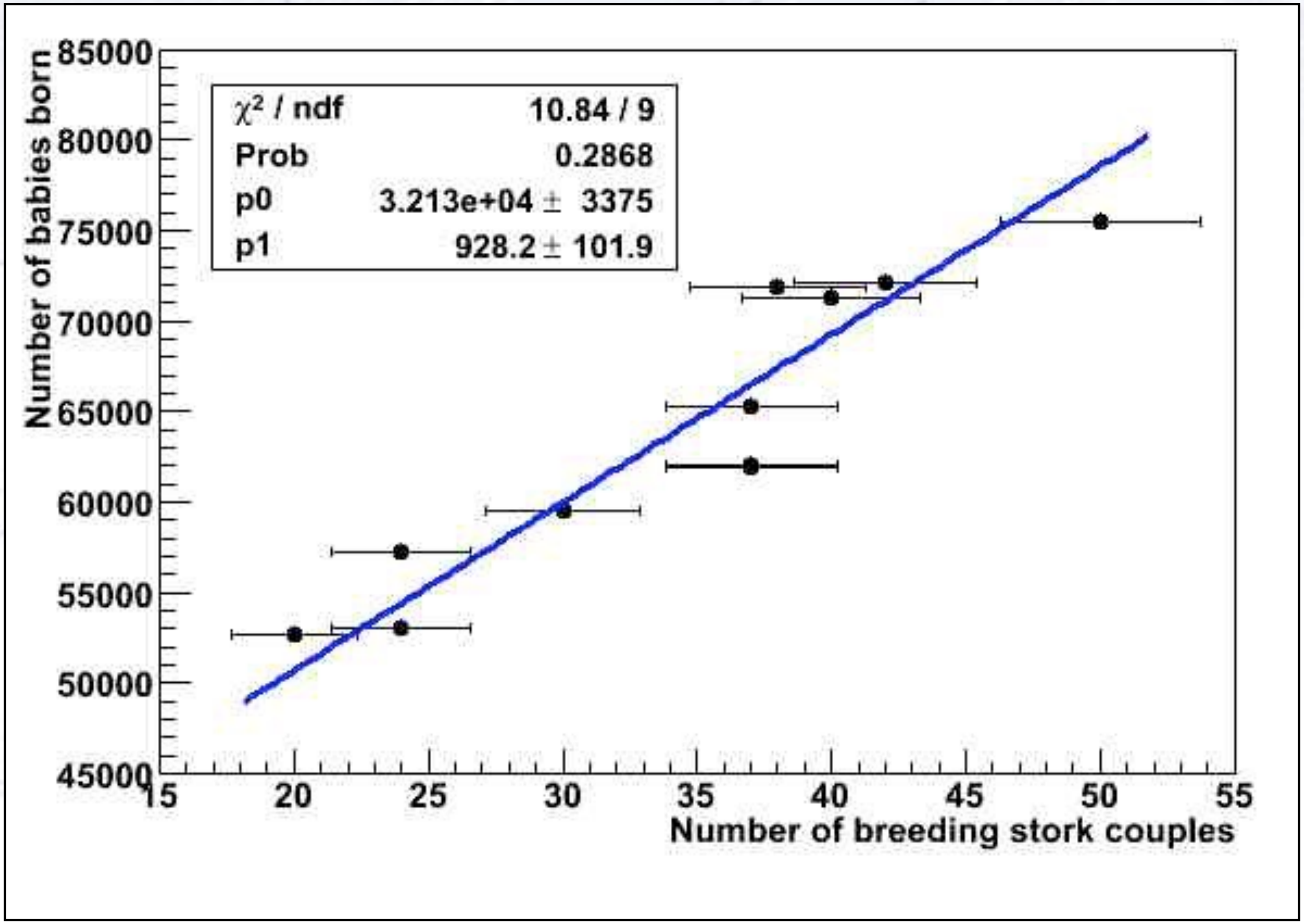
"Com hoc ergo propter hoc"

(with this, therefore because of this)

Correlation Vs. Causation

"Com hoc ergo propter hoc"

(with this, therefore because of this)



Digression on correlations

Why do correlations play a fundamental role?

1. It is the fundamental relation between variables.
2. Possible independent variables give you handles (see below).
3. The degree of simplicity/linearity tells you what methods to use.
4. Correlation with variable of interest is often key.

Imagine, that you find two sets of PID variables, which are **uncorrelated**. In this case, you can produce two **independent** ways to identify signal, giving you a method for measuring performance, cross checking results, and producing enriched samples of each type.
The two methods can of course be combined (with Likelihood or ML).

PDFs

Probability Density Functions

Probability Density Functions

A Probability Density Function (PDF) f(x) describes the probability of an outcome x:

probability to observe x in the interval [x, x+dx] = f(x) dx

PDFs are required to be normalised:

$$\int_S f(x)dx = 1$$

The expectation value (aka. mean) and the variance (i.e. standard deviation squared) are then defined as follows:

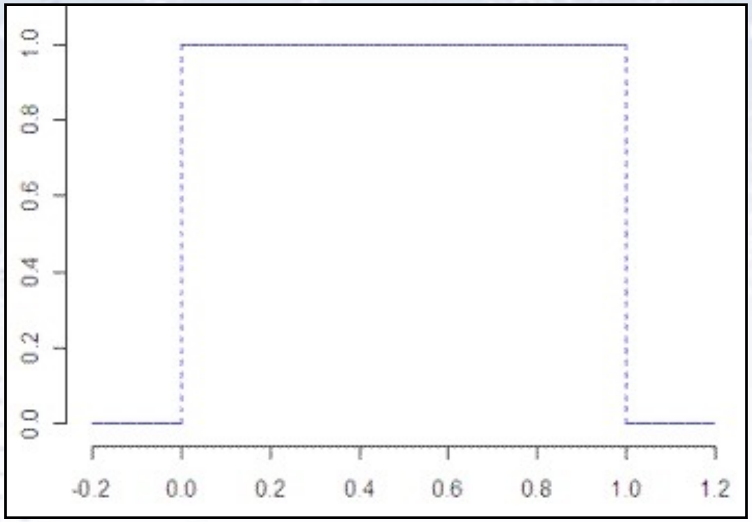
$$\mu = \int_{-\infty}^{\infty} x f(x) dx$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

Probability Density Functions

<u>Example:</u>
Consider a uniform distribution:

$$f(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & else \end{cases}$$



Calculating the mean and variance:

$$\mu = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x dx = [\frac{1}{2}x^2]_0^1 = \frac{1}{2}$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_0^1 (x - \frac{1}{2})^2 dx =$$

$$[\frac{1}{3}x^3 - \frac{1}{2}x^2 + \frac{1}{4}x]_0^1 = \frac{1}{3} - \frac{1}{2} + \frac{1}{4} = \frac{1}{12}$$

Cumulative distributions functions

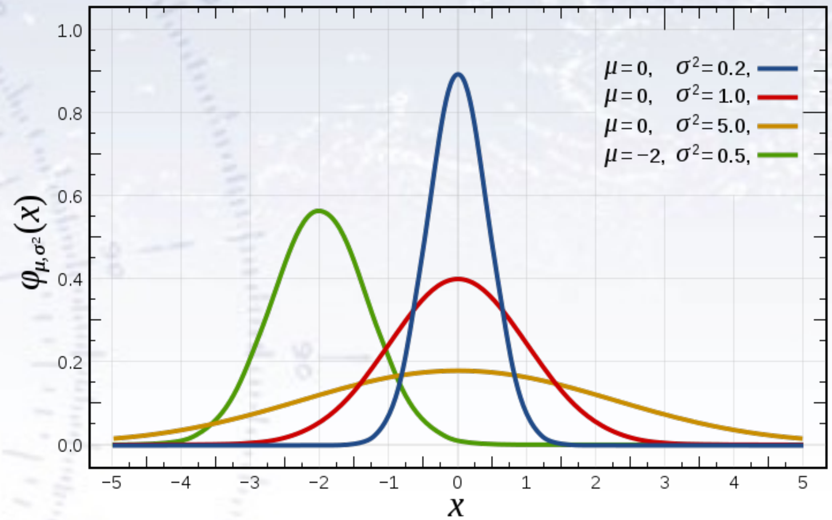
Completely basic to every PDF is the **cumulative distribution function,** CDF, defined as:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

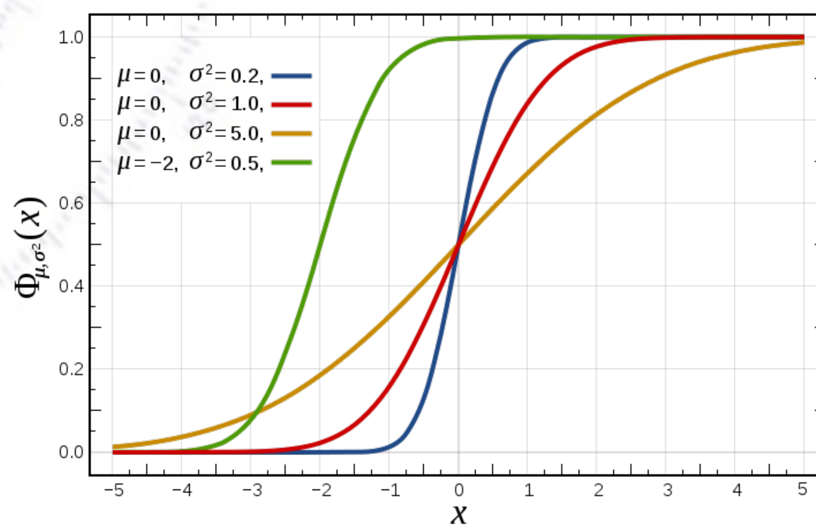
In words, this means that it is the probability of getting x, or something below that value.

The CDF is used in many ways, and we will meet it again soon, when we discuss hypothesis testing.

Gaussian PDF



Gaussian CDF



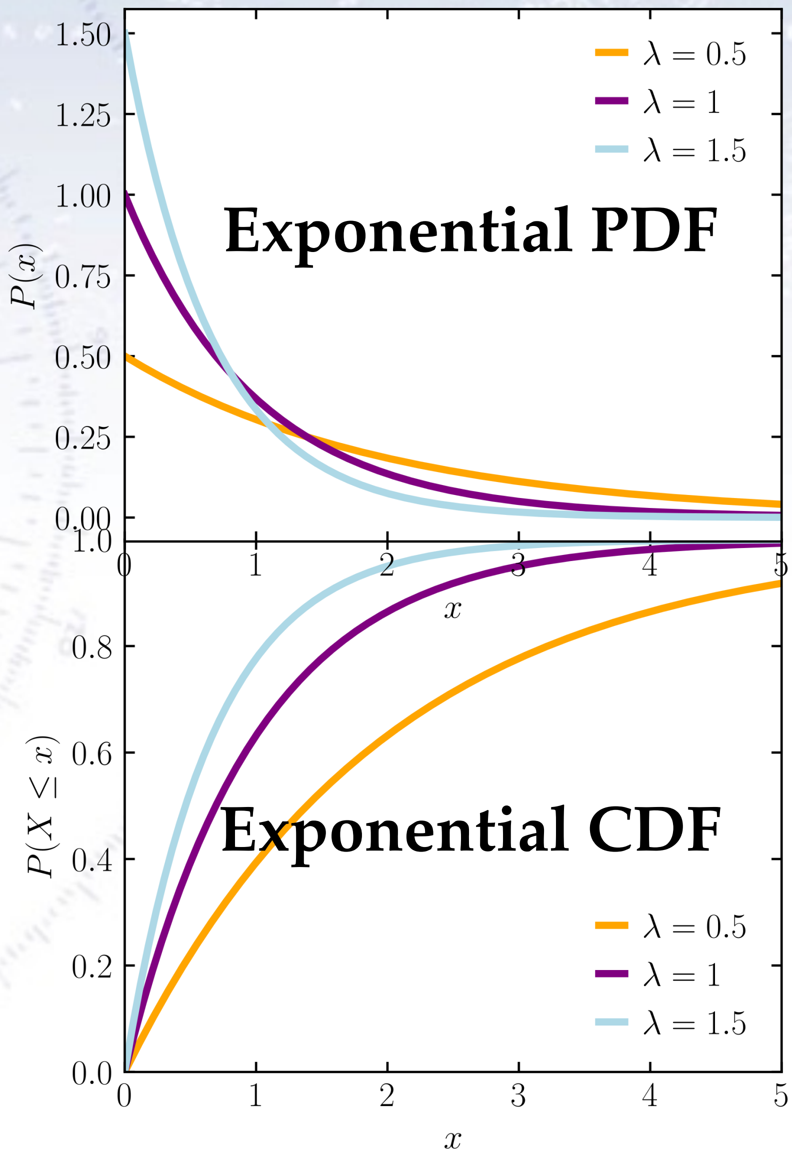
Cumulative distributions functions

Completely basic to every PDF is the **cumulative distribution function,** CDF, defined as:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

In words, this means that it is the probability of getting x, or something below that value.

The CDF is used in many ways, and we will meet it again soon, when we discuss hypothesis testing.



Probability Density Functions

The number of PDFs is infinite, and nearly so is the list of known ones:

Discrete distributions [edit source | edit beta]

With finite support [edit source | edit beta]

- The Bernoulli distribution, which takes value 1 with
- The Rademacher distribution, which takes value 1
- The binomial distribution, which describes the numb
- The beta-binomial distribution, which describes the
- The degenerate distribution at x_0, where X is certa random variables in the same formalism.
- The discrete uniform distribution, where all element shuffled deck.
- The hypergeometric distribution, which describes th there is no replacement.
- The Poisson binomial distribution, which describes
- Fisher's noncentral hypergeometric distribution
- Wallenius' noncentral hypergeometric distribution
- Benford's law, which describes the frequency of th

With infinite support [edit source | edit beta]

- The beta negative binomial distribution
- The Boltzmann distribution, a discrete distribution i analogue. Special cases include:
  - The Gibbs distribution
  - The Maxwell–Boltzmann distribution
- The Borel distribution
- The extended negative binomial distribution
- The extended hypergeometric distribution
- The generalized log-series distribution
- The generalized normal distribution
- The geometric distribution, a discrete distribution w
- The hypergeometric distribution
- The logarithmic (series) distribution
- The negative binomial distribution or Pascal distribu
- The parabolic fractal distribution
- The Poisson distribution, which describes a very la Poisson, the hyper-Poisson, the general Poisson b
  - The Conway–Maxwell–Poisson distribution, a tw
- The Polya-Eggenberger distribution
- The Skellam distribution, the distribution of the diff
- The skew elliptical distribution
- The skew normal distribution
- The Yule–Simon distribution
- The zeta distribution has uses in applied statistics
- Zipf's law or the Zipf distribution. A discrete power-
- The Zipf–Mandelbrot law is a discrete power law dis

Continuous distributions [edit source | edit beta]

Supported on a bounded interval [edit source | edit

- The Arcsine distribution on [a,b], which is a speci
- The Beta distribution on [0,1], of which the uniforr
- The Logitnormal distribution on (0,1).
- The Dirac delta function although not strictly a fur but the notation treats it as if it were a continuous
- The continuous uniform distribution on $[a,b]$, wher
  - The rectangular distribution is a uniform distrib
- The Irwin-Hall distribution is the distribution of the
- The Kent distribution on the three-dimensional sph
- The Kumaraswamy distribution is as versatile as t
- The logarithmic distribution (continuous)
- The PERT distribution is a special case of the bet
- The raised cosine distribution on $[\mu - s, \mu + s]$
- The reciprocal distribution
- The triangular distribution on $[a, b]$, a special case
- The truncated normal distribution on $[a, b]$.
- The U-quadratic distribution on $[a, b]$.
- The von Mises distribution on the circle.
- The von Mises-Fisher distribution on the N-dimens
- The Wigner semicircle distribution is important in t

Supported on semi-infinite intervals, usually $[0,\infty)$

- The Beta prime distribution
- The Birnbaum–Saunders distribution, also known a
- The chi distribution
  - The noncentral chi distribution
- The chi-squared distribution, which is the sum of t
  - The inverse-chi-squared distribution
  - The noncentral chi-squared distribution
  - The Scaled-inverse-chi-squared distribution
- The Dagum distribution
- The exponential distribution, which describes the t
- The F-distribution, which is the distribution of the ratio of two chi-squared variates which are not no
  - The noncentral F-distribution
- Fisher's z-distribution
- The folded normal distribution
- The Fréchet distribution
- The Gamma distribution, which describes the time
  - The Erlang distribution, which is a special case
  - The inverse-gamma distribution
- The generalized Pareto distribution
- The Gamma/Gompertz distribution
- The Gompertz distribution
- The half-normal distribution

- Hotelling's T-squared distribution
- The inverse Gaussian distribution, also kn
- The Lévy distribution
- The log-Cauchy distribution
- The log-gamma distribution
- The log-Laplace distribution
- The log-logistic distribution
- The log-normal distribution, describing vari
- The Mittag–Leffler distribution
- The Nakagami distribution
- The Pareto distribution, or "power law" dist
- The Pearson Type III distribution
- The phased bi-exponential distribution is c
- The phased bi-Weibull distribution
- The Rayleigh distribution
- The Rayleigh mixture distribution
- The Rice distribution
- The shifted Gompertz distribution
- The type-2 Gumbel distribution
- The Weibull distribution or Rosin Rammler grinding, milling and crushing operations.

Supported on the whole real line [edit sour

- The Behrens–Fisher distribution, which aris
- The Cauchy distribution, an example of a resonance energy distribution, impact and
- Chernoff's distribution
- The Exponentially modified Gaussian distri
- The Fisher–Tippett, extreme value, or log-
  - The Gumbel distribution, a special case
- Fisher's z-distribution
- The generalized logistic distribution
- The generalized normal distribution
- The geometric stable distribution
- The Holtsmark distribution, an example of
- The hyperbolic distribution
- The hyperbolic secant distribution
- The Johnson SU distribution
- The Landau distribution
- The Laplace distribution
- The Lévy skew alpha-stable distribution or distribution, Lévy distribution and normal d
- The Linnik distribution
- The logistic distribution
- The map-Airy distribution
- The normal distribution, also called the Ga independent, identically distributed variabl
- The Normal-exponential-gamma distributior
- The Pearson Type IV distribution (see Pea
- The skew normal distribution

- Student's t-distribution, useful for estimating u
  - The noncentral t-distribution
- The type-1 Gumbel distribution
- The Voigt distribution, or Voigt profile, is the c
- The Gaussian minus exponential distribution is

With variable support [edit source | edit beta]

- The generalized extreme value distribution has parameter
- The generalized Pareto distribution has a sup
- The Tukey lambda distribution is either suppo
- The Wakeby distribution

Mixed discrete/continuous distributions [edi

- The rectified Gaussian distribution replaces ne

Joint distributions [edit source | edit beta]

For any set of independent random variables the

Two or more random variables on the same sar

- The Dirichlet distribution, a generalization of th
- The Ewens's sampling formula is a probability
- The Balding–Nichols model
- The multinomial distribution, a generalization o
- The multivariate normal distribution, a generali
- The negative multinomial distribution, a genera
- The generalized multivariate log-gamma distrib

Matrix-valued distributions [edit source | edit b

- The Wishart distribution
- The inverse-Wishart distribution
- The matrix normal distribution
- The matrix t-distribution

Non-numeric distributions [edit source | edit

- The categorical distribution

newton distribution

Miscellaneous distributions [edit source | edit

- The Cantor distribution
- The generalized logistic distribution family
- The Pearson distribution family
- The phase-type distribution

And surely more!

Probability Density Functions

The number of PDFs is infinite, and nearly so is the list of known ones:

Discrete distributions [edit source | edit beta]

With finite support [edit source | edit beta]

- The Bernoulli distribution, which takes value 1 with
- The Rademacher distribution, which takes value 1
- The binomial distribution, which describes the numb
- The beta-binomial distribution, which describes the
- The degenerate distribution at x_0, where X is certa random variables in the same formalism.
- The discrete uniform distribution, where all element shuffled deck.
- The hypergeometric distribution, which describes th there is no replacement.
- The Poisson binomial distribution, which describes
- Fisher's noncentral hypergeometric distribution
- Wallenius' noncentral hypergeometric distribution

Continuous distributions [edit source | edit beta]

Supported on a bounded interval [edit source | edit

- The Arcsine distribution on [a,b], which is a speci
- The Beta distribution on [0,1], of which the unifor
- The Logitnormal distribution on (0,1).
- The Dirac delta function although not strictly a fur but the notation treats it as if it were a continuous
- The continuous uniform distribution on $[a,b]$, wher
  - The rectangular distribution is a uniform distrib
- The Irwin-Hall distribution is the distribution of the
- The Kent distribution on the three-dimensional sph
- The Kumaraswamy distribution is as versatile as t
- The logarithmic distribution (continuous)
- The PERT distribution is a special case of the bet
- The raised cosine distribution on $[\mu - s, \mu + s]$

- Hotelling's T-squared distribution
- The inverse Gaussian distribution, also kn
- The Lévy distribution
- The log-Cauchy distribution
- The log-gamma distribution
- The log-Laplace distribution
- The log-logistic distribution
- The log-normal distribution, describing vari
- The Mittag–Leffler distribution
- The Nakagami distribution
- The Pareto distribution, or "power law" dist
- The Pearson Type III distribution
- The phased bi-exponential distribution is c
- The phased bi-Weibull distribution
- The Rayleigh distribution
- The Rayleigh mixture distribution
- The Rice distribution

- Student's t-distribution, useful for estimating u
  - The noncentral t-distribution
- The type-1 Gumbel distribution
- The Voigt distribution, or Voigt profile, is the c
- The Gaussian minus exponential distribution is

With variable support [edit source | edit beta]

- The generalized extreme value distribution has parameter
- The generalized Pareto distribution has a sup
- The Tukey lambda distribution is either suppor
- The Wakeby distribution

Mixed discrete/continuous distributions [edi

- The rectified Gaussian distribution replaces n

Joint distributions [edit source | edit beta]

For any set of independent random variables the

https://docs.scipy.org/doc/scipy/reference/stats.html

- The Gibbs distribution
- The Maxwell–Boltzmann distribution
- The Borel distribution
- The extended negative binomial distribution
- The extended hypergeometric distribution
- The generalized log-series distribution
- The generalized normal distribution
- The geometric distribution, a discrete distribution w
- The hypergeometric distribution
- The logarithmic (series) distribution
- The negative binomial distribution or Pascal distrib
- The parabolic fractal distribution
- The Poisson distribution, which describes a very la Poisson, the hyper-Poisson, the general Poisson b
  - The Conway–Maxwell–Poisson distribution, a tw
- The Polya-Eggenberger distribution
- The Skellam distribution, the distribution of the diff
- The skew elliptical distribution
- The skew normal distribution
- The Yule–Simon distribution
- The zeta distribution has uses in applied statistics
- Zipf's law or the Zipf distribution. A discrete power-
- The Zipf–Mandelbrot law is a discrete power law dis

Supported on semi-infinite intervals, usually [0,∞)

- The Beta prime distribution
- The Birnbaum–Saunders distribution, also known a
- The chi distribution
  - The noncentral chi distribution
- The chi-squared distribution, which is the sum of t
  - The inverse-chi-squared distribution
  - The noncentral chi-squared distribution
  - The Scaled-inverse-chi-squared distribution
- The Dagum distribution
- The exponential distribution, which describes the t
- The F-distribution, which is the distribution of the ratio of two chi-squared variates which are not no
  - The noncentral F-distribution
- Fisher's z-distribution
- The folded normal distribution
- The Fréchet distribution
- The Gamma distribution, which describes the time
  - The Erlang distribution, which is a special case
  - The inverse-gamma distribution
- The generalized Pareto distribution
- The Gamma/Gompertz distribution
- The Gompertz distribution
- The half-normal distribution

- The Cauchy distribution, an example of a resonance energy distribution, impact and
- Chernoff's distribution
- The Exponentially modified Gaussian distri
- The Fisher–Tippett, extreme value, or log-
  - The Gumbel distribution, a special case
- Fisher's z-distribution
- The generalized logistic distribution
- The generalized normal distribution
- The geometric stable distribution
- The Holtsmark distribution, an example of
- The hyperbolic distribution
- The hyperbolic secant distribution
- The Johnson SU distribution
- The Landau distribution
- The Laplace distribution
- The Lévy skew alpha-stable distribution or distribution, Lévy distribution and normal d
- The Linnik distribution
- The logistic distribution
- The map-Airy distribution
- The normal distribution, also called the Ga independent, identically distributed variabl
- The Normal-exponential-gamma distributior
- The Pearson Type IV distribution (see Pea
- The skew normal distribution

- The negative multinomial distribution, a genera
- The generalized multivariate log-gamma distrib

Matrix-valued distributions [edit source | edit b

- The Wishart distribution
- The inverse-Wishart distribution
- The matrix normal distribution
- The matrix t-distribution

Non-numeric distributions [edit source | edit b

- The categorical distribution

newton distribution

Miscellaneous distributions [edit source | edit

- The Cantor distribution
- The generalized logistic distribution family
- The Pearson distribution family
- The phase-type distribution

And surely more!

Probability Density Functions

The number of PDFs is infinite, and nearly so is the list of known ones:

Discrete distributions [edit source | edit beta]

With finite support [edit source | edit beta]

- The Bernoulli distribution, which takes value 1 with
- The Rademacher distribution, which takes value 1
- The binomial distribution, which describes the numb
- The beta-binomial distribution, which describes the
- The degenerate distribution at x_0, where X is certa random variables in the same formalism.
- The discrete uniform distribution, where all element shuffled deck
- The h there
- The F
- Fishe
- Walle
- Benfo

With infi

- The b
- The E anal
  - Th
  - Th
- The E
- The e
- The e
- The generalized log-series distribution
- The generalized normal distribution
- The geometric distribution, a discrete distribution w
- The hypergeometric distribution
- The logarithmic (series) distribution
- The negative binomial distribution or Pascal distribu
- The parabolic fractal distribution
- The Poisson distribution, which describes a very la Poisson, the hyper-Poisson, the general Poisson b
  - The Conway–Maxwell–Poisson distribution, a tw
- The Polya-Eggenberger distribution
- The Skellam distribution, the distribution of the diff
- The skew elliptical distribution
- The skew normal distribution
- The Yule–Simon distribution
- The zeta distribution has uses in applied statistics
- Zipf's law or the Zipf distribution. A discrete power-
- The Zipf–Mandelbrot law is a discrete power law dis

Continuous distributions [edit source | edit beta]

Supported on a bounded interval [edit source | edit

- The Arcsine distribution on [a,b], which is a speci
- The Beta distribution on [0,1], of which the unifor
- The Logitnormal distribution on (0,1).
- The Dirac delta function although not strictly a fu but the notation treats it as if it were a continuous
- The continuous uniform distribution on [a,b], wher
  - The rectangular distribution is a uniform distrib
- The Irwin-Hall distribution is the distribution of the
- The chi-squared distribution, which is the sum of t
  - The inverse-chi-squared distribution
  - The noncentral chi-squared distribution
  - The Scaled-inverse-chi-squared distribution
- The Dagum distribution
- The exponential distribution, which describes the t
- The F-distribution, which is the distribution of the ratio of two chi-squared variates which are not no
  - The noncentral F-distribution
- Fisher's z-distribution
- The folded normal distribution
- The Fréchet distribution
- The Gamma distribution, which describes the time
  - The Erlang distribution, which is a special case
  - The inverse-gamma distribution
- The generalized Pareto distribution
- The Gamma/Gompertz distribution
- The Gompertz distribution
- The half-normal distribution

- Hotelling's T-squared distribution
- The inverse Gaussian distribution, also kn
- The Lévy distribution
- The log-Cauchy distribution
- The log-gamma distribution
- The log-Laplace distribution
- The log-logistic distribution
- The log-normal distribution, describing vari
- The Mittag–Leffler distribution
- The Nakagami distribution
- The Pareto distribution, or "power law" dist
- The Pearson Type III distribution
- Fisher's z-distribution
- The generalized logistic distribution
- The generalized normal distribution
- The geometric stable distribution
- The Holtsmark distribution, an example of
- The hyperbolic distribution
- The hyperbolic secant distribution
- The Johnson SU distribution
- The Landau distribution
- The Laplace distribution
- The Lévy skew alpha-stable distribution or distribution, Lévy distribution and normal d
- The Linnik distribution
- The logistic distribution
- The map-Airy distribution
- The normal distribution, also called the Ga independent, identically distributed variabl
- The Normal-exponential-gamma distributior
- The Pearson Type IV distribution (see Pea
- The skew normal distribution

- Student's t-distribution, useful for estimating u
  - The noncentral t-distribution
- The type-1 Gumbel distribution
- The Voigt distribution, or Voigt profile, is the c
- The Gaussian minus exponential distribution is

With variable support [edit source | edit beta]

- The generalized extreme value distribution has parameter
- The generalized Pareto distribution has a sup
- The Tukey lambda distribution is either suppo
- The Wakeby distribution

butions [edi

on replaces n

edit beta]

variables the

the same sar

eralization of t

s a probability

eneralization c

ion, a generali

ution, a genera

-gamma distrib

source | edit b

- The matrix t-distribution

Non-numeric distributions [edit source | edit b

- The categorical distribution

newton distribution

Miscellaneous distributions [edit source | edit

- The Cantor distribution
- The generalized logistic distribution family
- The Pearson distribution family
- The phase-type distribution

"Essentially, all models are wrong, but some are useful"

[George E. P. Box, British Statistician, 1919-2013]

And surely more!

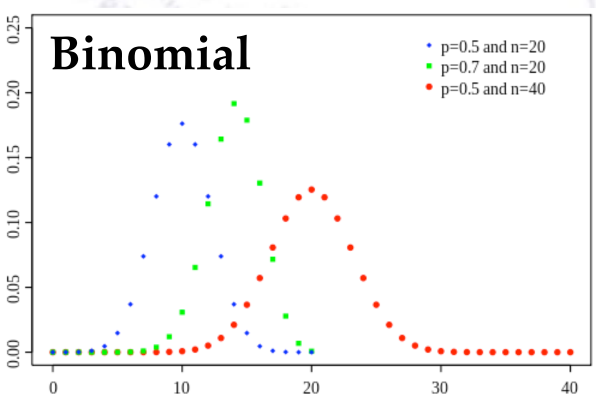
Probability Density Functions

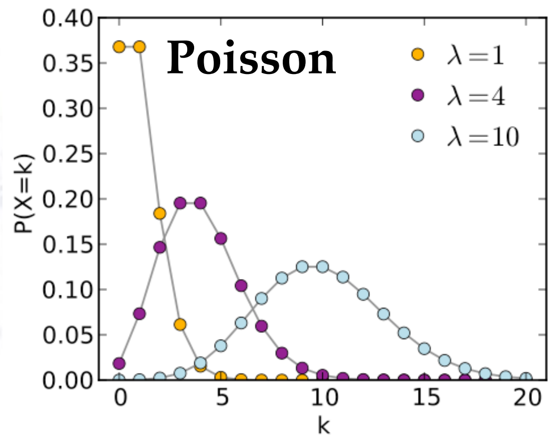
An almost complete list of those we will deal with in this course is:

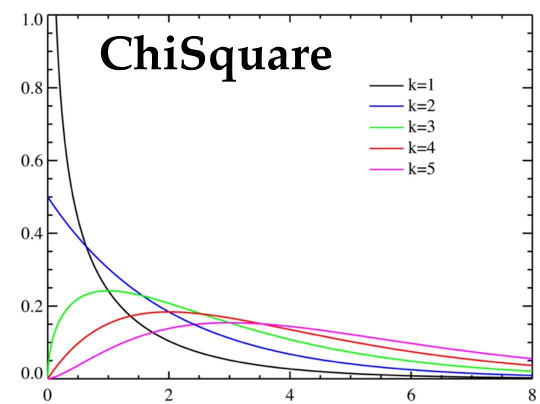
- **Gaussian** (aka. Normal)
- **Poisson**
- **Binomial** (and also Multinomial)
- Students t-distribution
- Uniform
- ChiSquare
- Exponential
- Error function (integral of Gaussian)

See Barlow chap.3 and Cowan chap.2

You should already know most of these, and the rest will be explained.







Binomial, Poisson, Gaussian

$$f(n;N,p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Given **N trials** each with **p chance of success**, how many **successes n** should you expect in total?

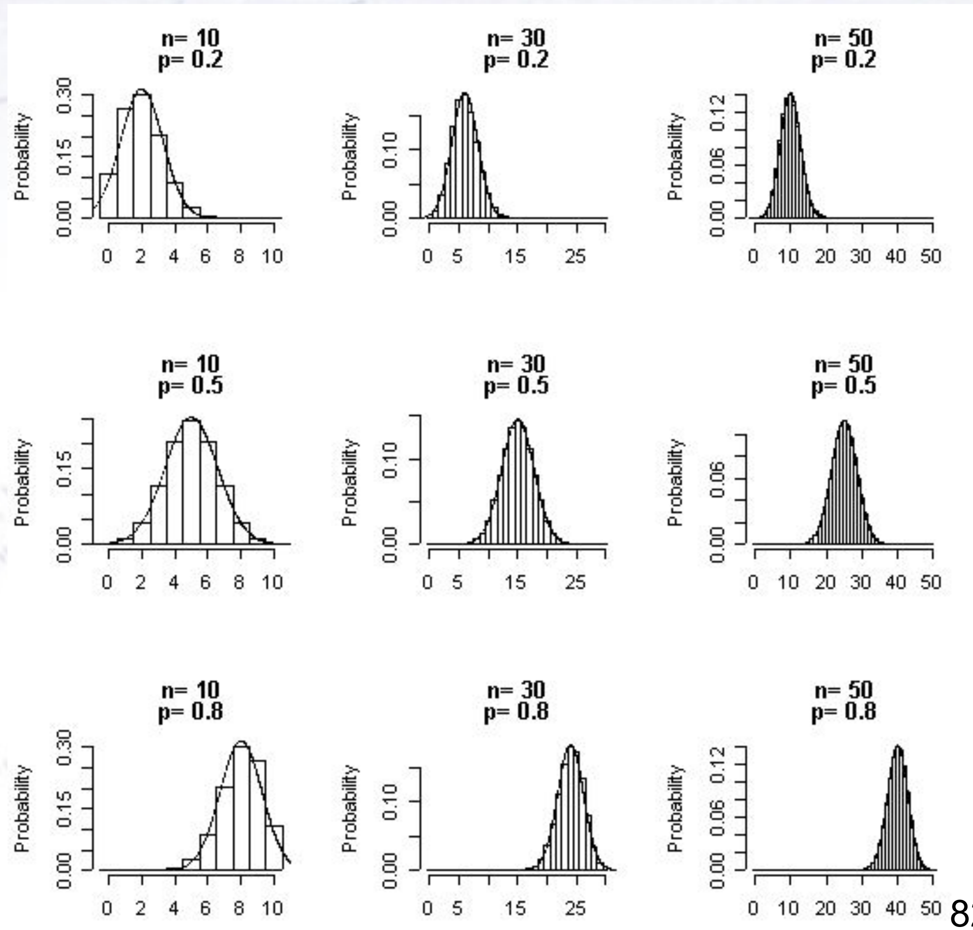
This distribution is… **Binomial**, with

Mean = Np

Variance = Np(1-p)

This means, that the error on a fraction f = n/N is:

$$\sigma(f) = \sqrt{\frac{f(1-f)}{N}}$$



Binomial, Poisson, Gaussian

$$f(n;N,p) = \boxed{\frac{N!}{n!(N-n)!}}\boxed{p^n(1-p)^{N-n}}$$

The binomial distribution was first introduced by Jacob Bernoulli in 1713 (posthumously).

$$(x+y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4.$$

0:	1								
1:	1	1							
2:	1	2	1						
3:	1	3	3	1					
4:	1	4	6	4	1				
5:	1	5	10	10	5	1			
6:	1	6	15	20	15	6	1		
7:	1	7	21	35	35	21	7	1	
8:	1	8	28	56	70	56	28	8	1

The binomial distribution basically consists of two elements: The binomial coefficient (green) and the probabilities of exactly n such events (blue).

Even though a system has many outcomes, it is typically possible to refer to either "success" of "failure".

Assume the probability to have COVID19 is 1%. In a sample of 50 people the chance to have 1 or more infected is: 1-p(0) = 1 - 0.99^{50} = 0.60

Binomial, Poisson, Gaussian

<u>Requirements to be Binomial:</u>

- Fixed number of trials, N
- Independent trials.
- Only two outcomes (success/failure).
- Constant probability of success/failure.

If number of possible outcomes is more than two ⇒ **Multinomial distribution**.

<u>Examples of Binomial experiments:</u>

- Tossing a coin 20 times counting number of tails.
- Asking 200 people if they watch sports on TV.
- Rolling a die to see if a 6 appear (Multinomial for all outcomes!).
- Asking 100 die-hards from Enhedslisten, if they would vote for Konservative at the next election!

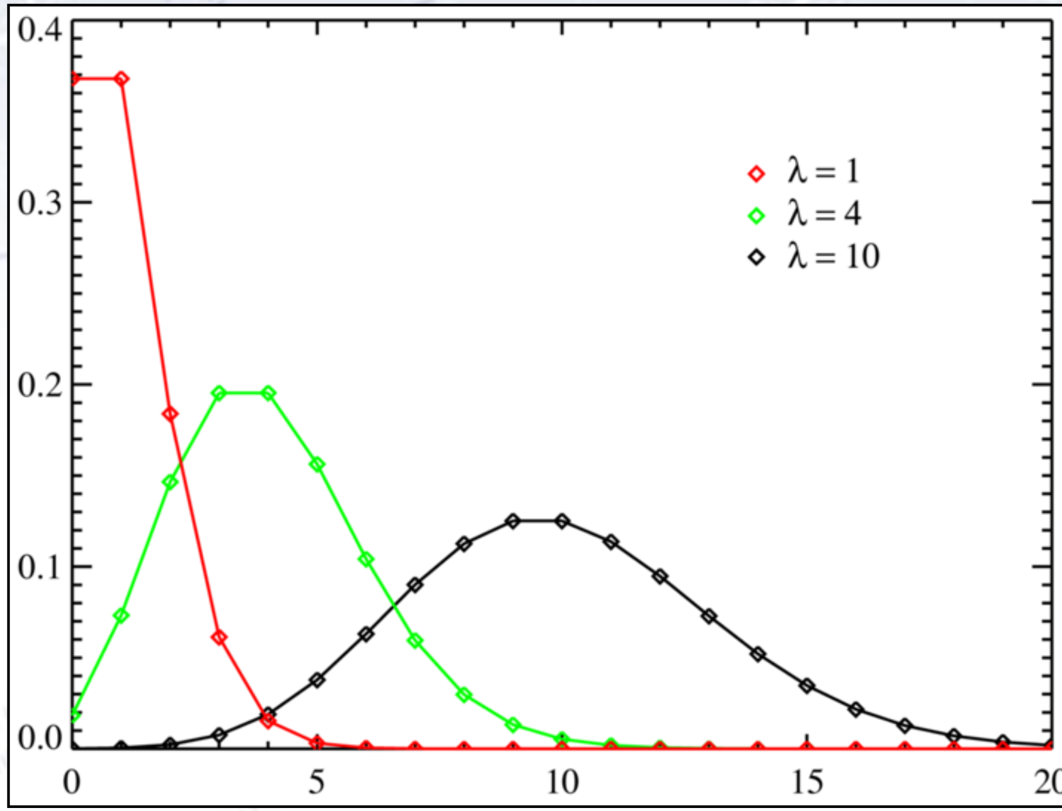
<u>Examples which aren't Binomial experiments:</u>

- Rolling a die until a 6 appears (not fixed number of trials).
- Drawing 5 cards for a poker hand (no replacement ⇒ not independent)

Binomial, **Poisson**, Gaussian

If N → ∞ and p → 0, but Np → λ then a Binomial approaches a Poisson: (see Barlow 3.3.1)

$$f(n,\lambda) = \frac{\lambda^n}{n!}e^{-\lambda}$$



In reality, the approximation is already quite good at e.g. N=50 and p=0.1.

The Poisson distribution only has one parameter, namely λ.
Mean = λ
Variance = λ

So the error on a number is...

...the square root of that number!

Binomial, **Poisson**, Gaussian

The error on a (Poisson) number... is **the square root** of that number!!!

Binomial, **Poisson**, Gaussian

The error on a (Poisson) number...

A very useful case of this is the error to assign a bin in a histogram, if there is reasonable statistics ($N_i > 5\text{-}20$) in each bin.

is **the square root** of that number!!!

The error on a (Poisson) number... is **the square root** of that number!!!

Note: The sum of two Poissons with λ_a and λ_b is a new Poisson with $\lambda = \lambda_a + \lambda_b$.
(See Barlow pages 33-34 for proof)

Binomial, **Poisson**, Gaussian

The Poisson distribution has the advantage that **neither the number of trials N nor the probability of succes p has to be known** - just their product.

A typical use is when dealing with **rates** in a given interval of time, distance, area, volume, etc.

Binomial, **Poisson**, Gaussian

The Poisson distribution has the advantage that **neither the number of trials N nor the probability of succes p has to be known** - just their product.

A typical use is when dealing with **rates** in a given interval of time, distance, area, volume, etc.

<u>Example (real from 1898):</u>
There were 122 deaths by horse kicks over 10 different regiments, over 20 years. What is the predicted number of deaths in a specific regiment and year?

First we estimate the mean value:

$$\boxed{\mu = \frac{122}{20 * 10} = 0.61}$$

This means that the probability that 0 will die is given by:

$$\boxed{P(0) = e^{0}\frac{0.61^{0}}{0!} = 0.54}$$

Quick Quiz

You need to know the efficiency of your PID system for positrons.

Find 1000 data events where two electron candidates have a combined mass of 91.2 GeV (Z^0) and the negative candidate is identified as an electron ("Tag-and-probe" technique).

In 900 events the positive candidate is also identified as an electron.
In 100 events it is not. Efficiency is 90%, but what about the uncertainty?

Colleague A says sqrt(900) = 30, thus 90.0 ± 3.0 %
Colleague B says sqrt(100) = 10, thus 90.0 ± 1.0 %

Which is right?

Quick Quiz

You need to know the efficiency of your PID system for positrons.

Find 1000 data events where two electron candidates have a combined mass of 91.2 GeV (Z^0) and the negative candidate is identified as an electron ("Tag-and-probe" technique).

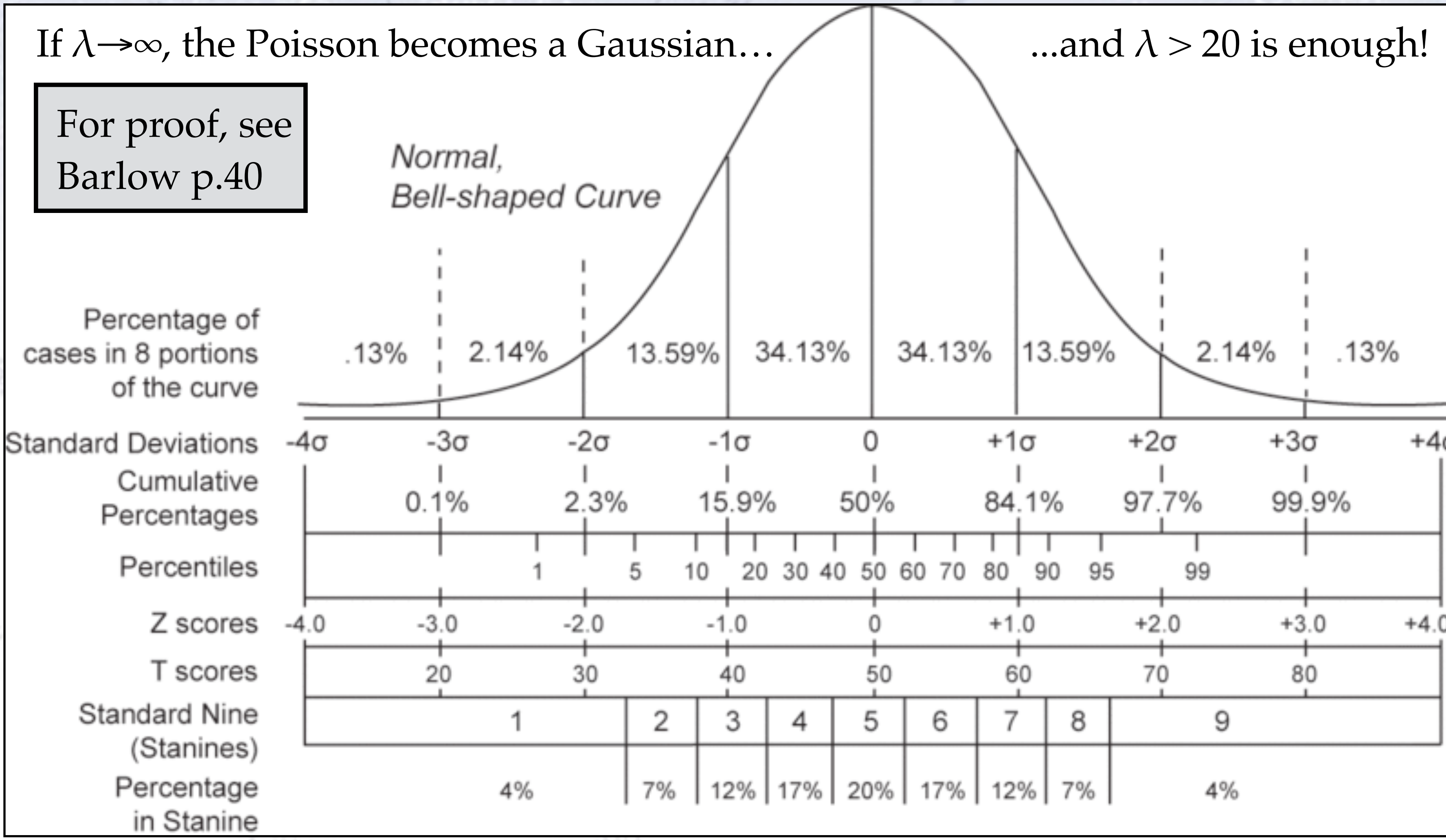
In 900 events the positive candidate is also identified as an electron.
In 100 events it is not. Efficiency is 90%, but what about the uncertainty?

Colleague A says sqrt(900) = 30, thus 90.0 ± 3.0 %
Colleague B says sqrt(100) = 10, thus 90.0 ± 1.0 %

Which is right? **Neither!**
This is not a Poisson but a Binomial (N = 1000 trials, p = 0.9 of success)
Uncertainty is sqrt(N*p*(1-p)) = 9.49, thus 90.0 ± 0.9 %

From previous page: $$\sigma(f) = \sqrt{\frac{f(1-f)}{N}}$$

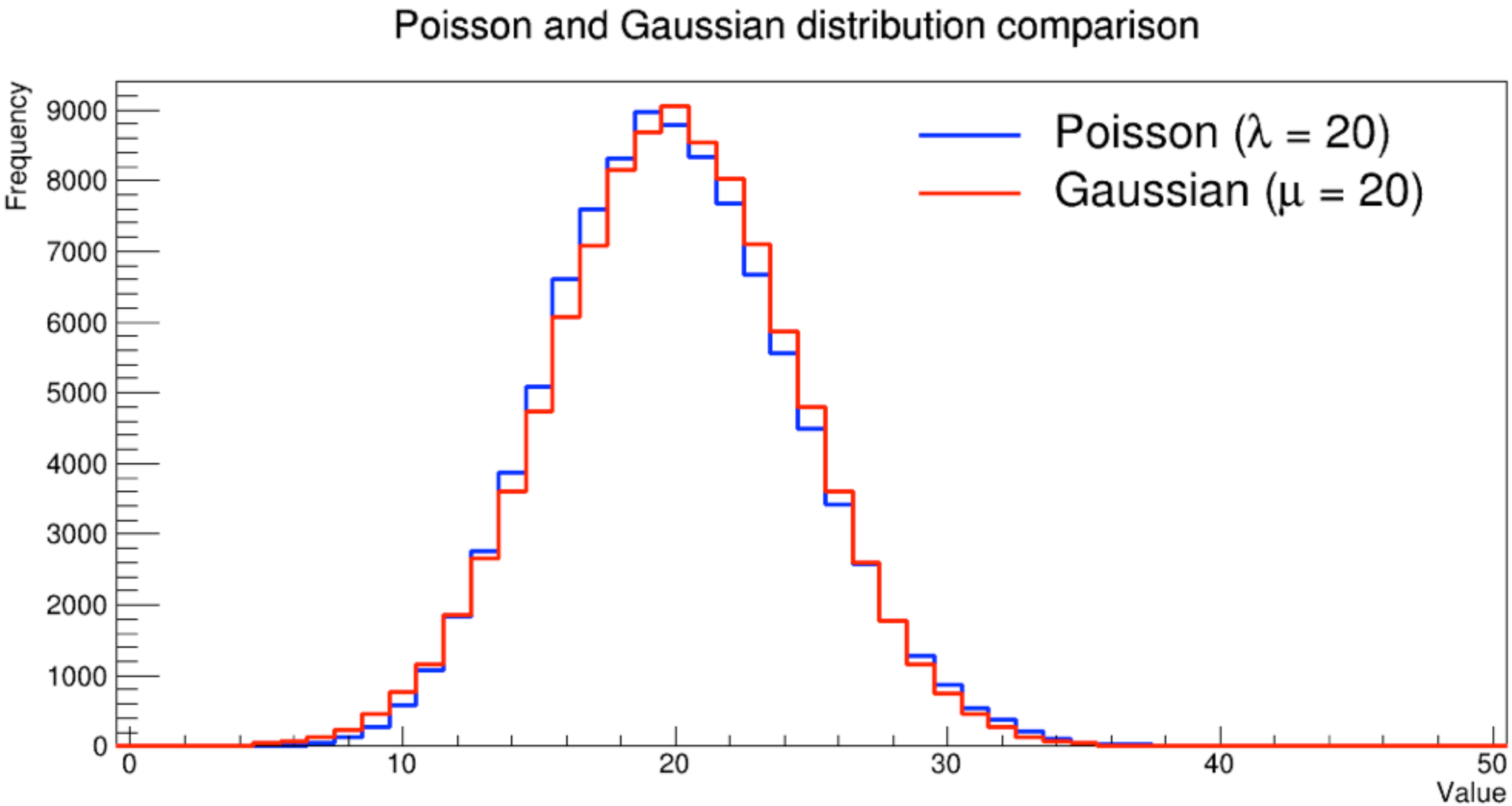
Binomial, Poisson, **Gaussian**



All fields encounter the Gaussian, and for this reason, its scale has many names!

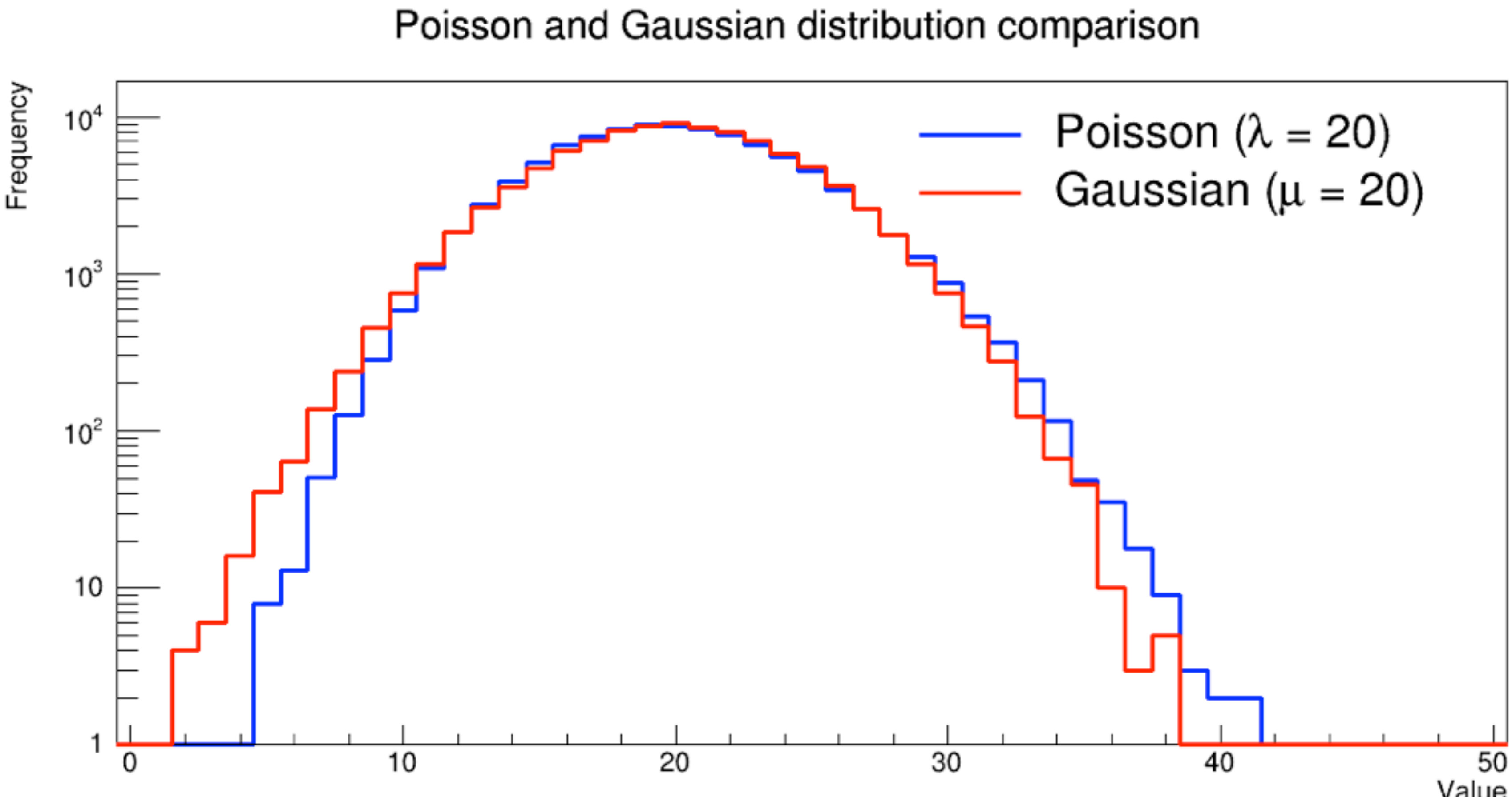
Binomial, Poisson, **Gaussian**

If $\lambda \to \infty$, the Poisson becomes a Gaussian… ...and $\lambda > 20$ is enough!



Binomial, Poisson, **Gaussian**

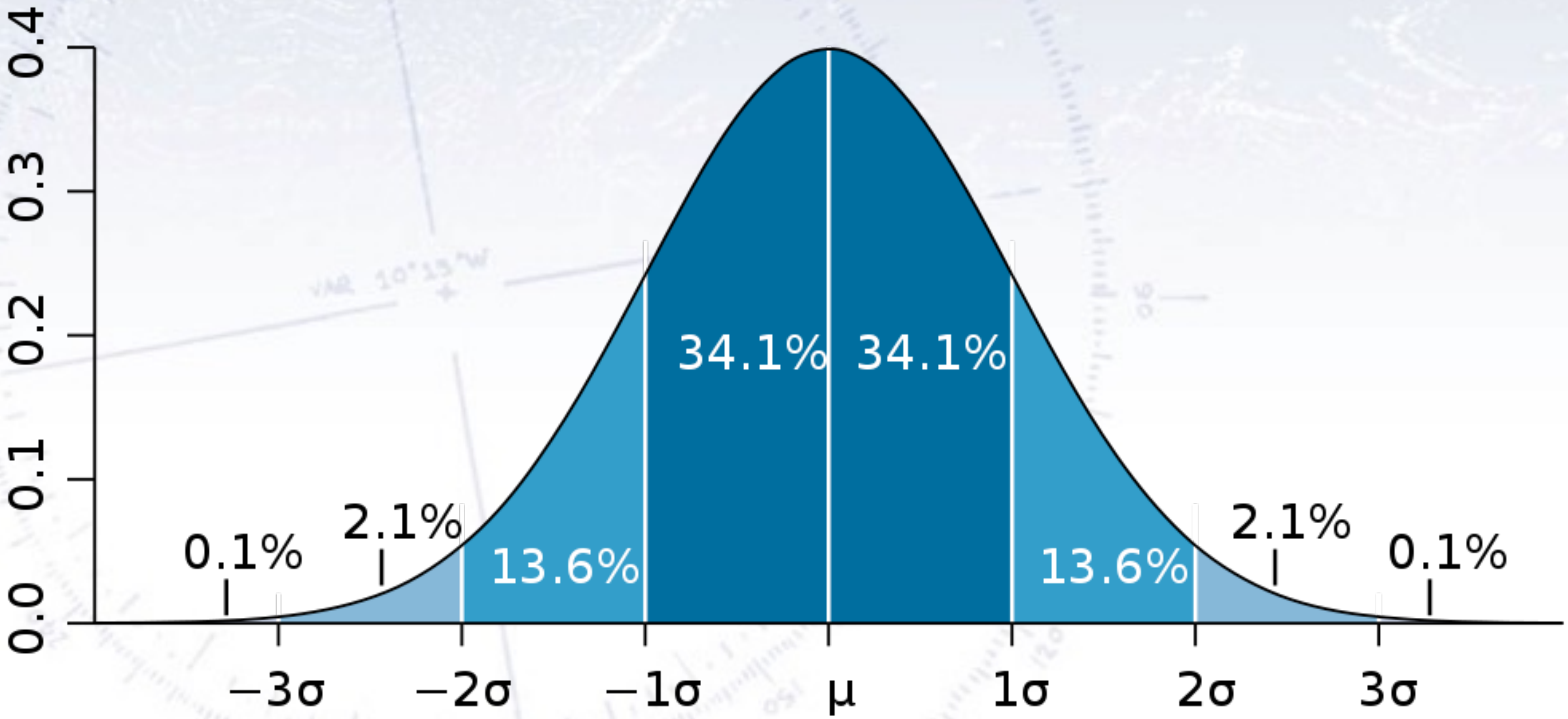
If $\lambda \to \infty$, the Poisson becomes a Gaussian… ...and $\lambda > 20$ is enough!



However, note that the TAILS are not quite the same!!!
This is the very reason for the difference between Chi2 and (binned) likelihood!

Binomial, Poisson, **Gaussian**

"If the Greeks had known it, they would have deified it."

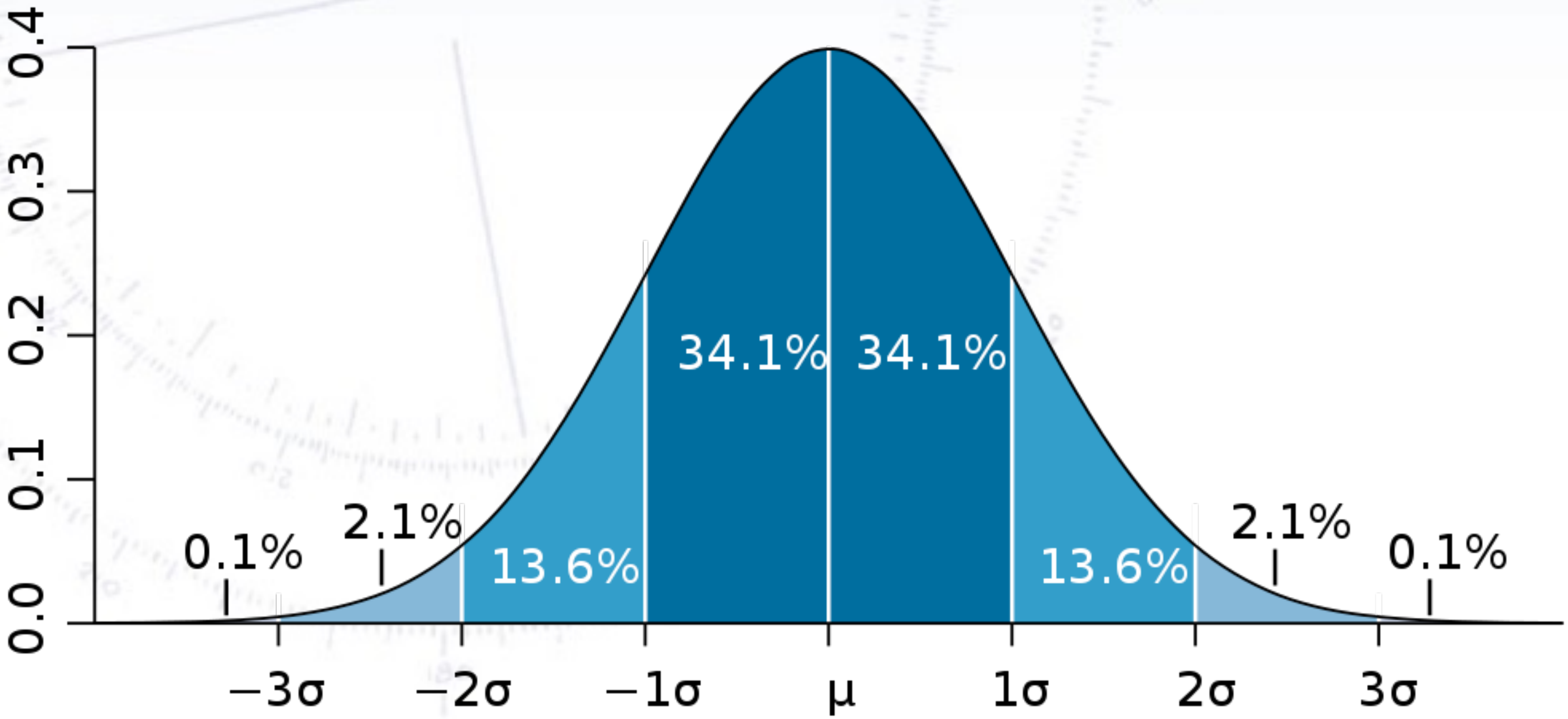


"If the Greeks had known it, they would have deified it. It reigns with serenity and in complete self-effacement amids the wildest confusion. ***The more huge the mob and the greater the apparent anarchy, the more perfect is its sway****. It is the supreme Law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to be latent all along." [Karl Pearson]*

Binomial, Poisson, **Gaussian**

The Gaussian **defines** the way we consider uncertainties.

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 %



Student's t-distribution

Given only a small (n obs.) sample (still assumed Gaussian), we don't know the mean μ and width σ well - we only know estimates of them! This changes the PDF to:

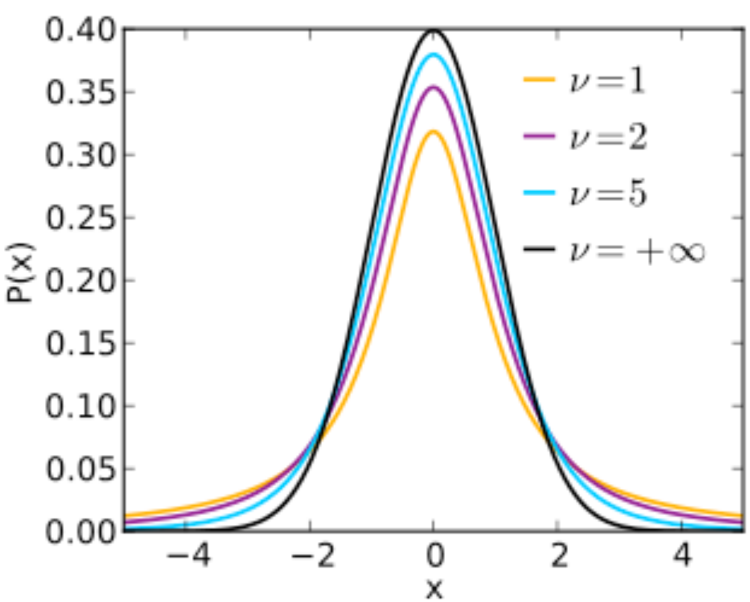
$$p(x\,|\,\nu,\hat{\mu},\hat{\sigma}^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\hat{\sigma}^2}}\left(1+\frac{1}{\nu}\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right)^{-\frac{\nu+1}{2}} \qquad \nu = N_{\mathrm{DoF}} = n-1$$

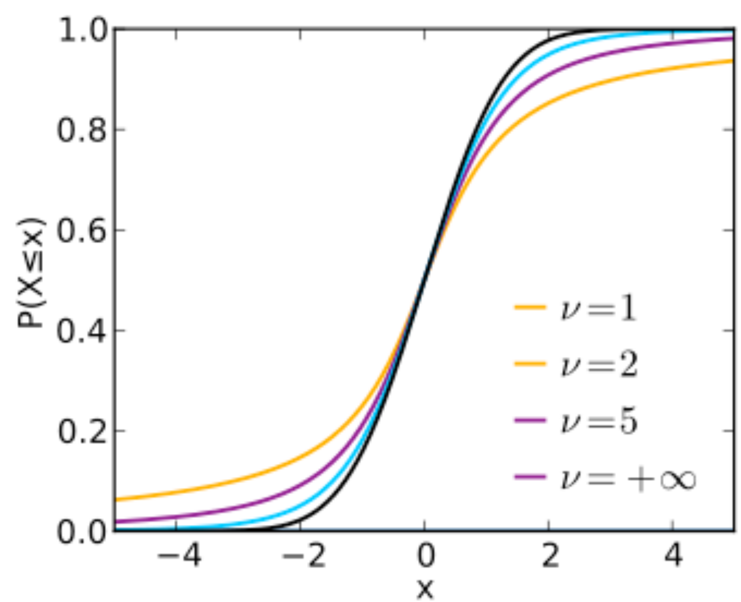
Student's t-distribution

Given only a small (n obs.) sample (still assumed Gaussian), we don't know the mean μ and width σ well - we only know estimates of them! This changes the PDF to:

$$p(x\,|\,\nu,\hat{\mu},\hat{\sigma}^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\,\sqrt{\pi\nu\hat{\sigma}^2}} \left(1+\frac{1}{\nu}\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right)^{-\frac{\nu+1}{2}} \qquad \nu = N_{\mathrm{DoF}} = n-1$$

"Discovered" by William Gosset, student's t-distribution takes into account the **lacking knowledge of the mean and variance** (as is the case for small samples).



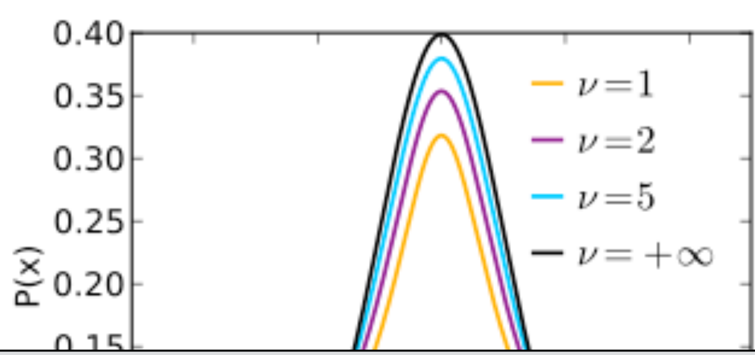


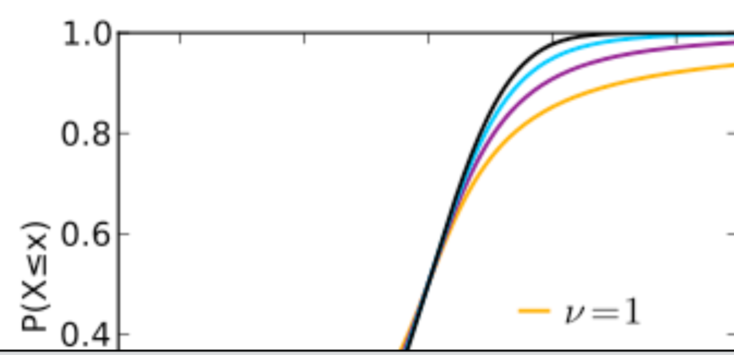
Student's t-distribution

Given only a small (n obs.) sample (still assumed Gaussian), we don't know the mean μ and width σ well - we only know estimates of them! This changes the PDF to:

$$p(x\,|\,\nu,\hat{\mu},\hat{\sigma}^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\,\sqrt{\pi\nu\hat{\sigma}^2}}\left(1+\frac{1}{\nu}\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right)^{-\frac{\nu+1}{2}} \qquad \nu = N_{\mathrm{DoF}} = n-1$$

"Discovered" by William Gosset, student's t-distribution takes into account the **lacking knowledge of the mean and variance** (as is the case for small samples).



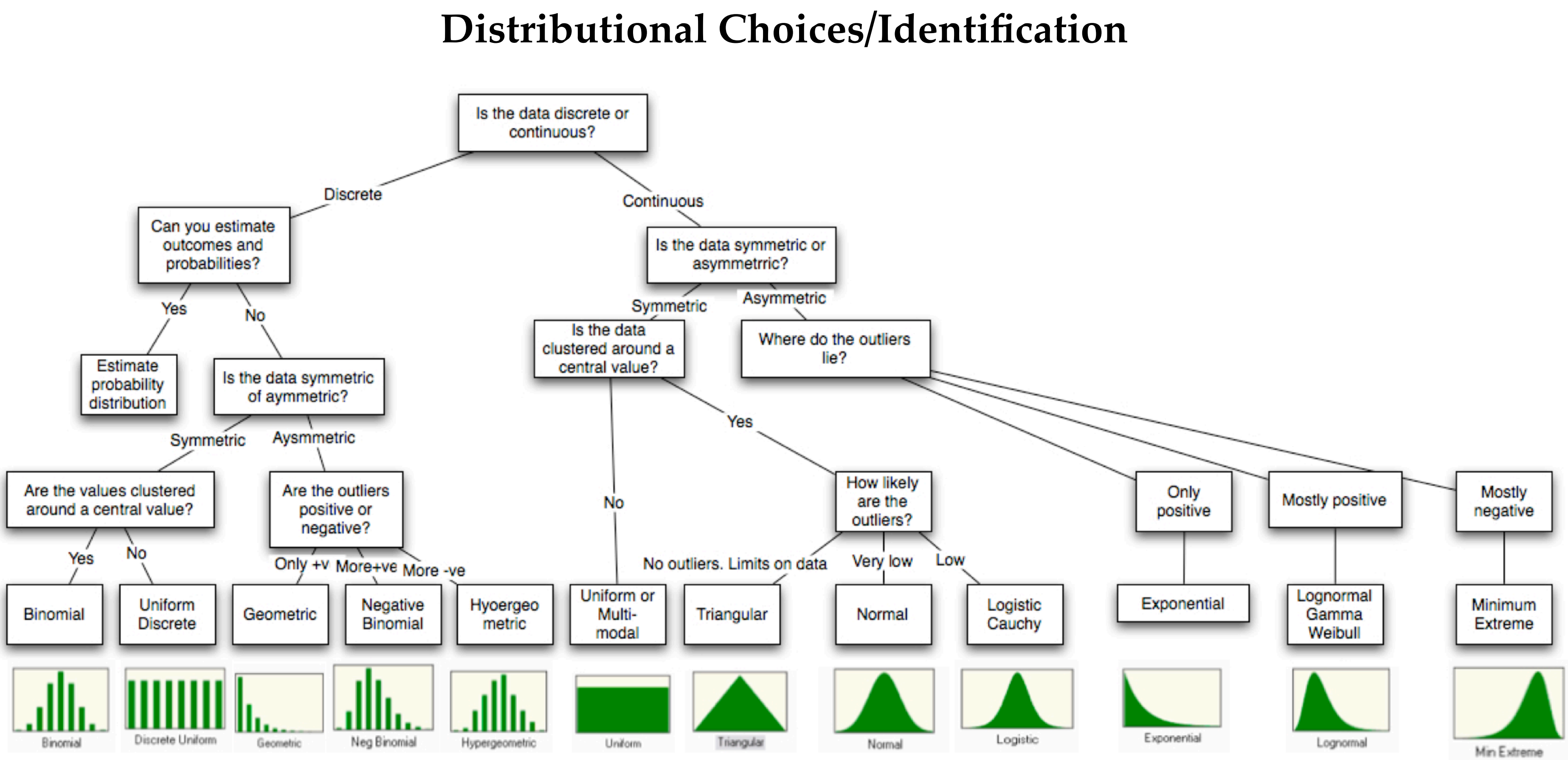


When mean and width are poorly known, estimating it from sample gives:

Gaussian: $z = \dfrac{x-\mu}{\sigma}$ **Student's:** $t = \dfrac{x-\hat{\mu}}{\hat{\sigma}}$

Distribution Overview

I like the following overview of the most common PDFs, though it is far from perfect. However, it shows what makes the essential differences between PDFs.



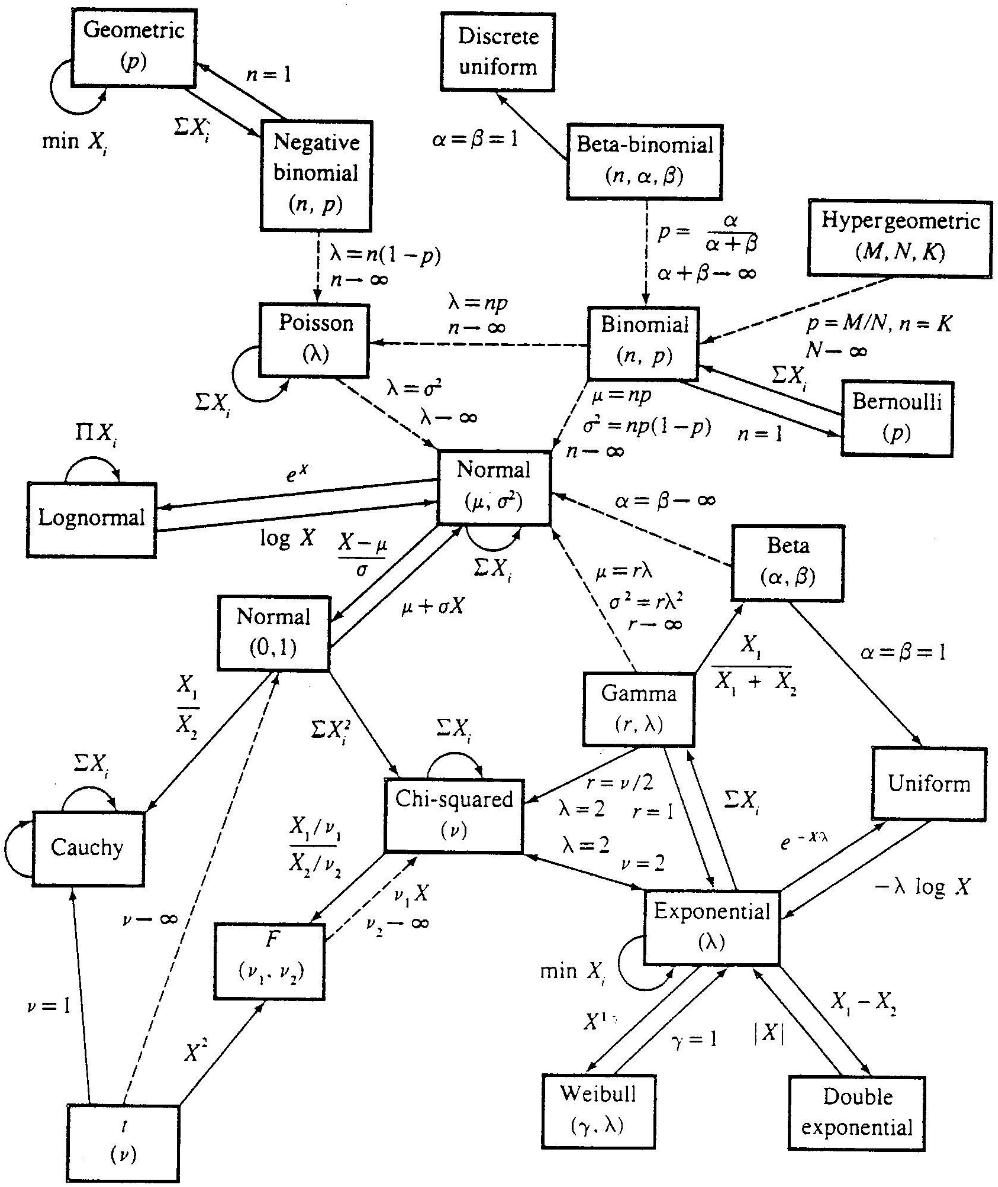
From: A. Damodaran

Distribution Relationship

The different PDFs are related.

As can be seen, essentially all PDFs "converges" towards the Gaussian (normal) distribution.

630 *Table of Common Distributions*



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

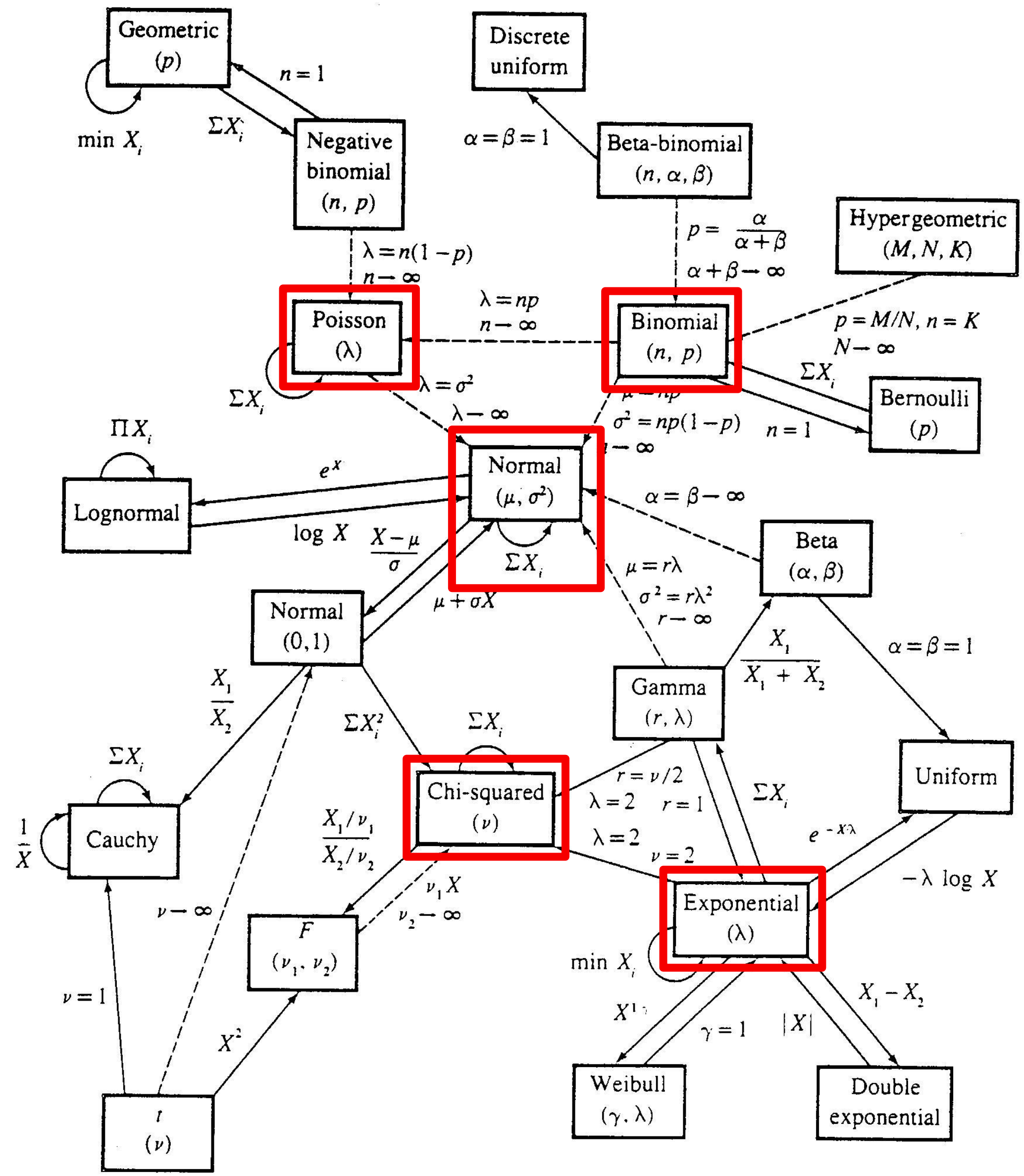
Distribution Relationship

The different PDFs are related.

As can be seen, essentially all PDFs "converges" towards the Gaussian (normal) distribution.

Don't worry about knowing them all…. Through a long life in statistics, I have still yet to encounter all of these in use!

630 Table of Common Distributions



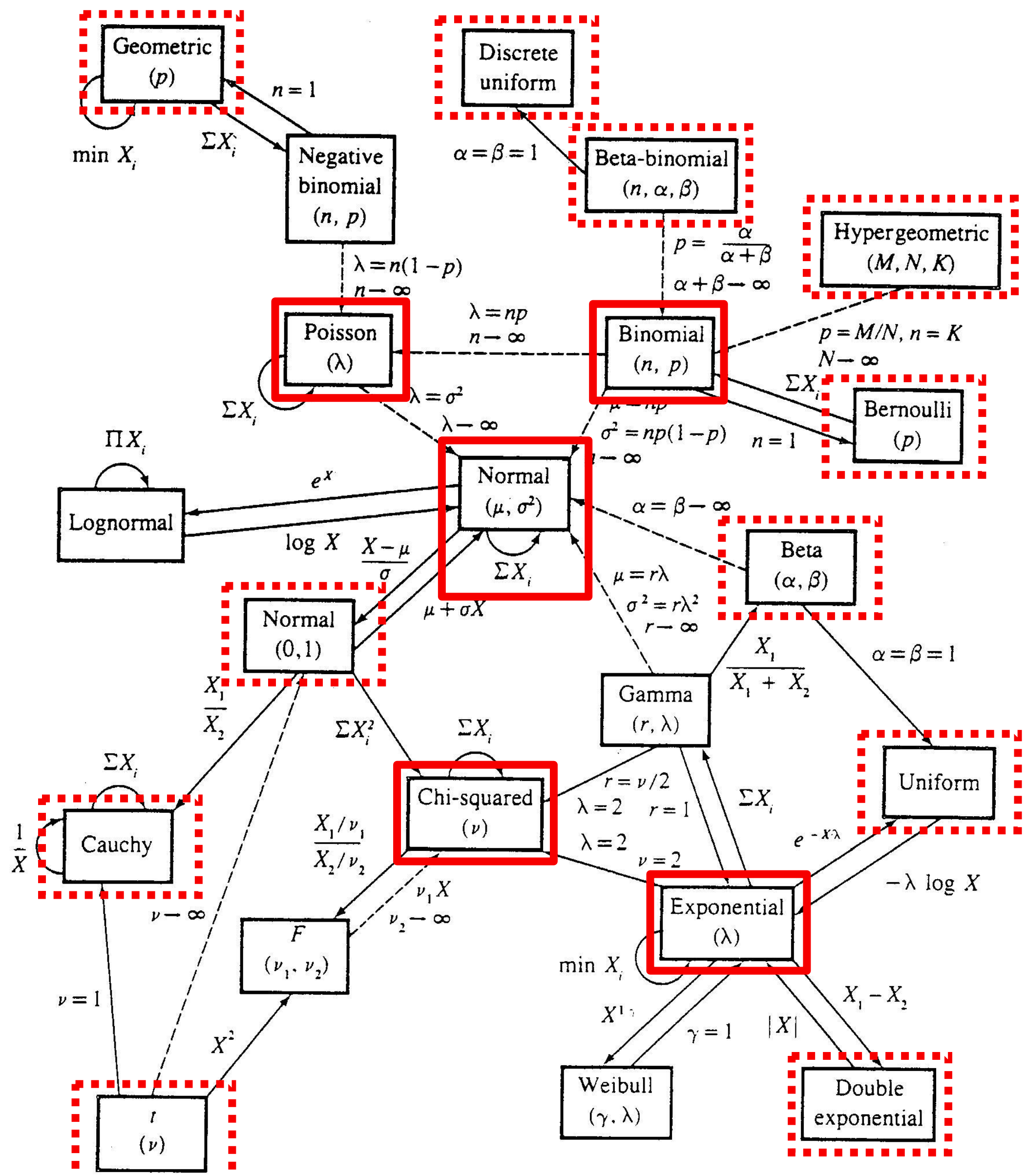
Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

Distribution Relationship

The different PDFs are related.

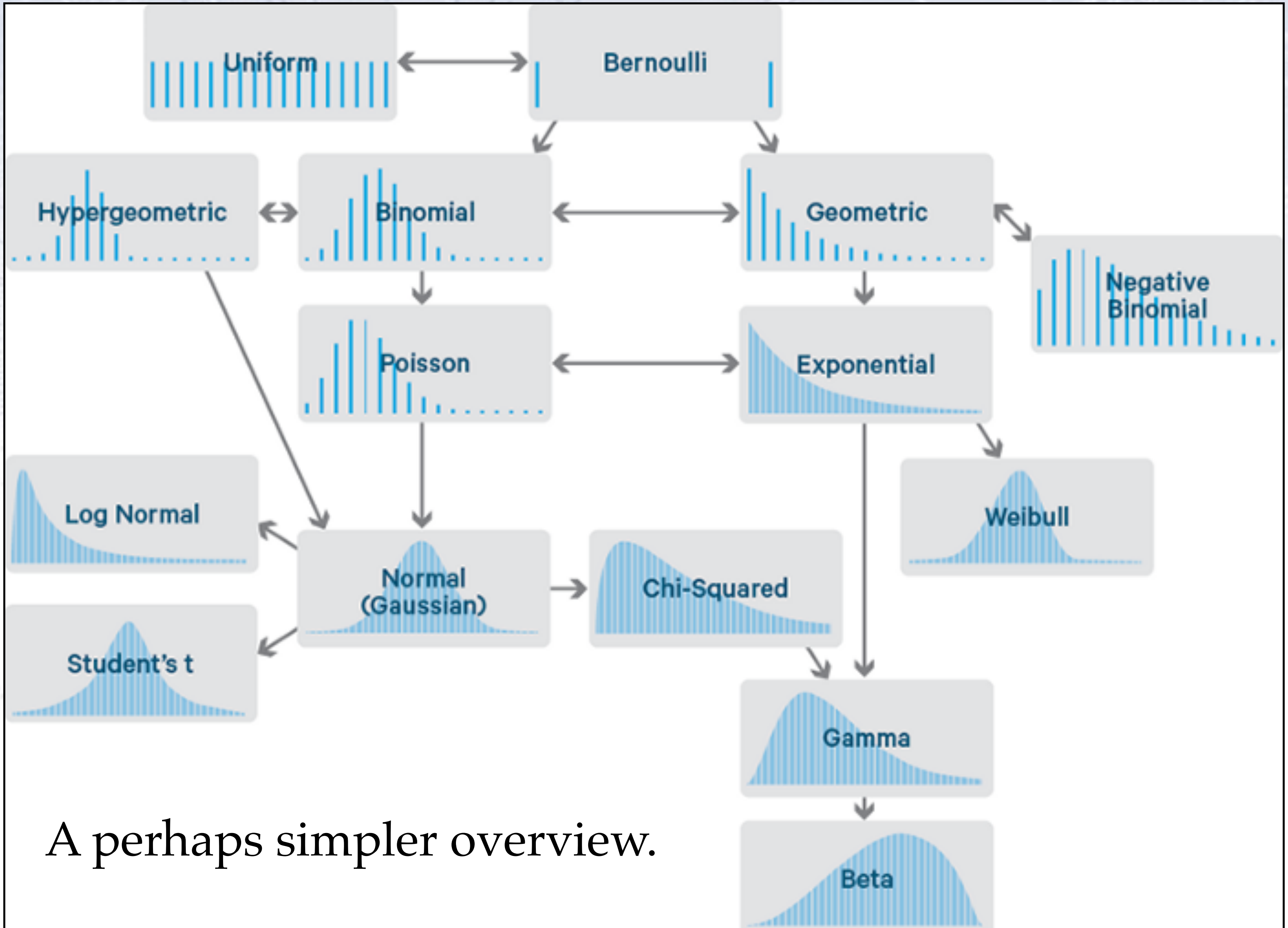
As can be seen, essentially all PDFs “converges” towards the Gaussian (normal) distribution.

Don’t worry about knowing them all…. Through a long life in statistics, I have still yet to encounter all of these in use!



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

Distribution Overview

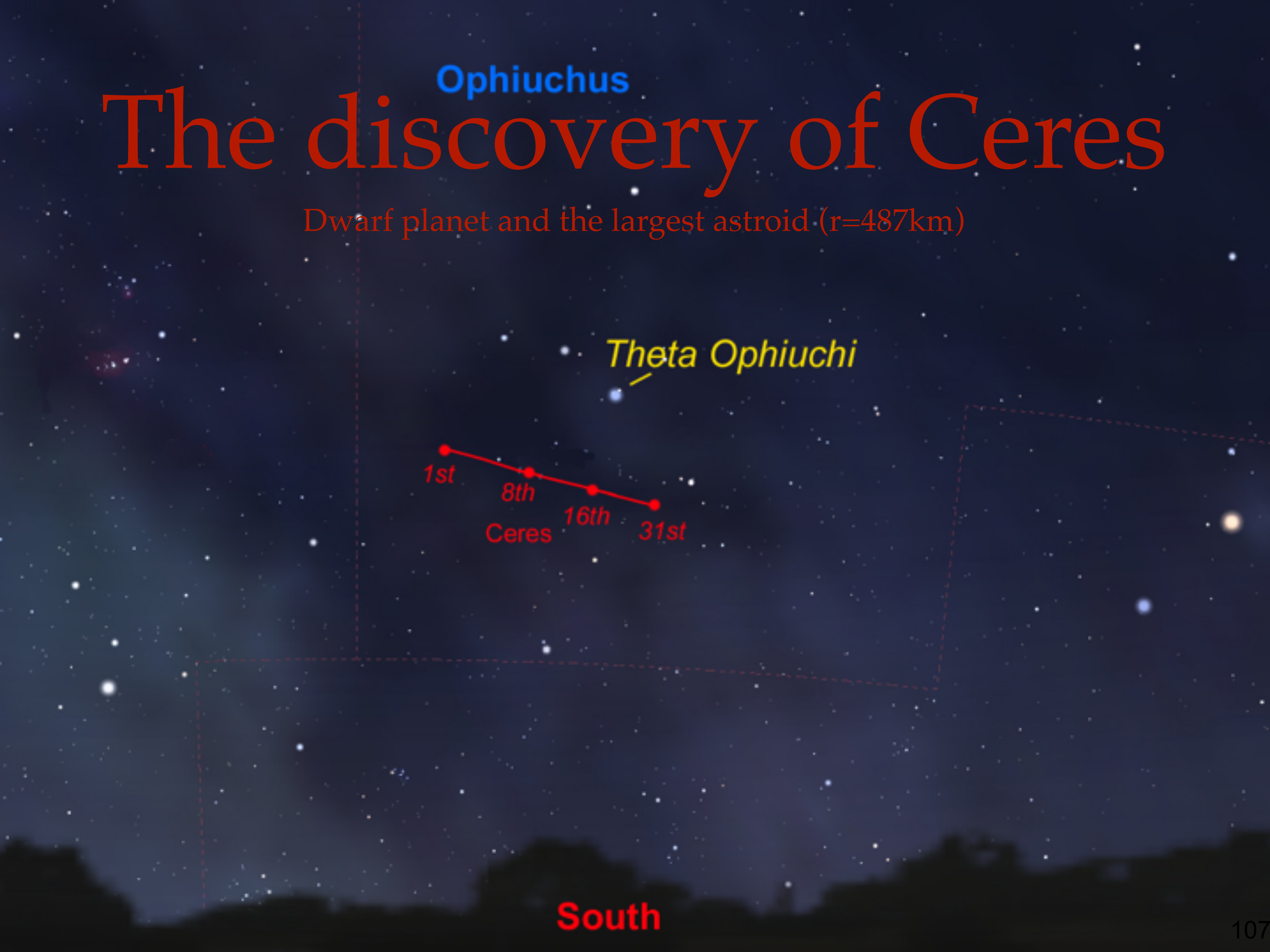


A perhaps simpler overview.

The ChiSquare

The discovery of Ceres

Dwarf planet and the largest astroid (r=487km)



The discovery of Ceres

Dwarf planet and the largest astroid (r=487km)

On the 1st of January 1801 Giuseppe Piazzi discovered “new light” and could follow this comet/planet until 11th of February. He published the positions, but due to Ceres being behind the sun, it would be out of sight until the following winter. Following the calculations of a 24 year old mathematician/physicist, it was recovered on the 31st of December 1801 by von Zach and H. Olbers.
The young man’s name was Carl Friedrich Gauss, and the method he used/invented for this was…

The discovery of Ceres

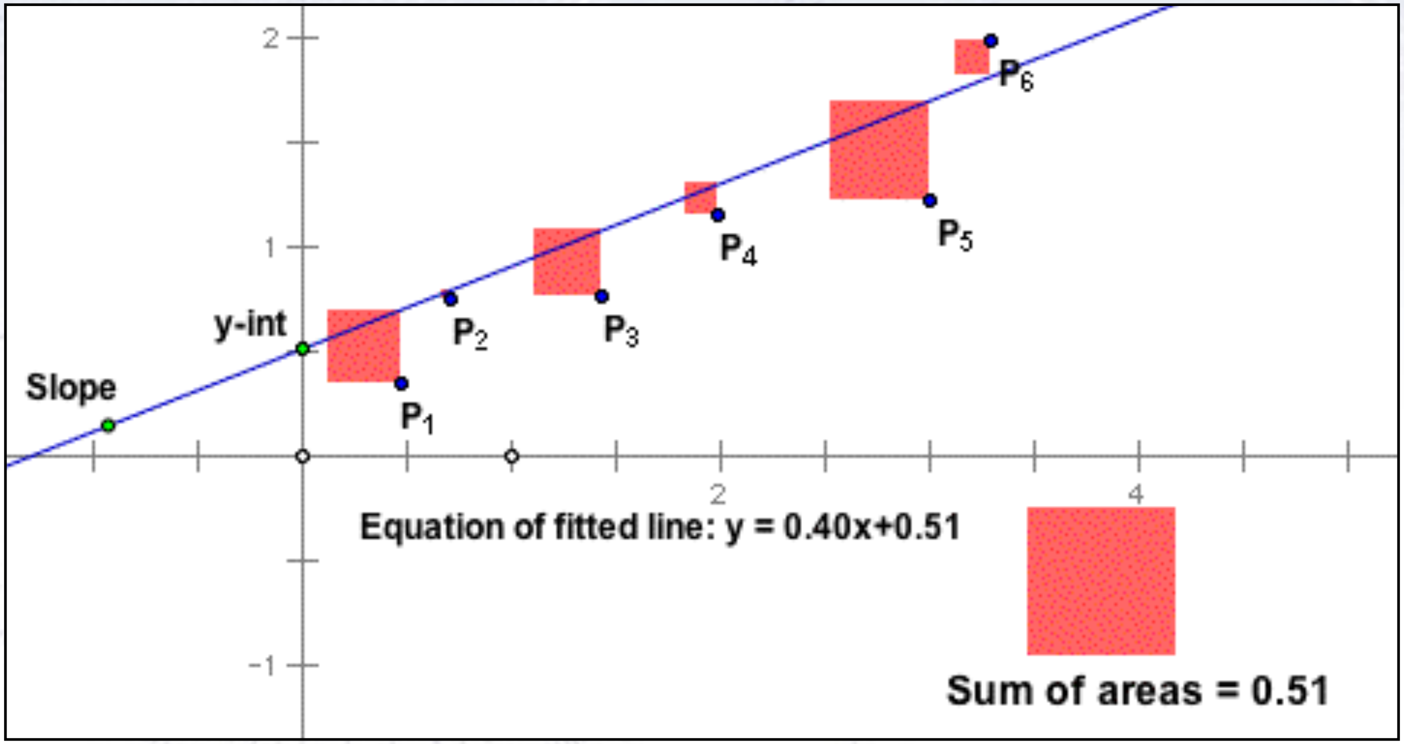
Dwarf planet and the largest astroid (r=487km)

On the 1st of January 1801 Giuseppe Piazzi discovered “new light” and could follow this comet/planet until 11th of February. He published the positions, but due to Ceres being behind the sun, it would be out of sight until the following winter. Following the calculations of a 24 year old mathematician/physicist, it was recovered on the 31st of December 1801 by von Zach and H. Olbers.
The young man’s name was Carl Friedrich Gauss, and the method he used/invented for this was…

…method of least squares!

Method of Least Squares

The problem at hand is determining the curve that best fitted data:

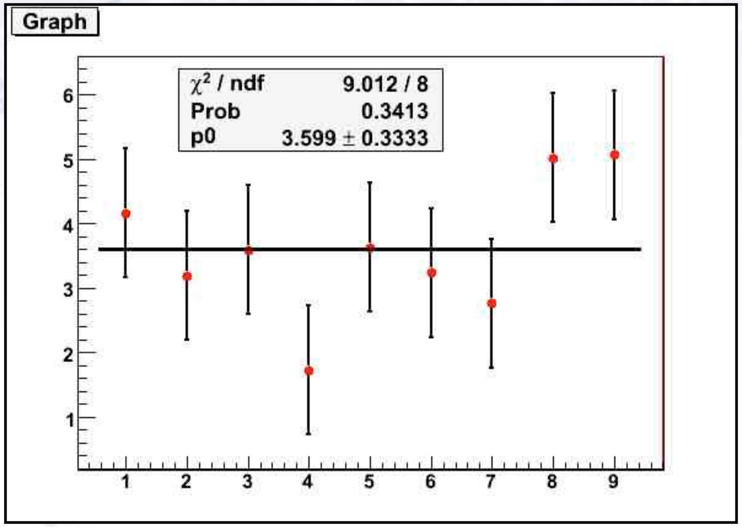


The "best fit" is found by minimising the sum of the squares…

Originally, uncertainties were not included (not "invented" yet!)

Method of Least Squares

The method of least squares is a standard approach to the approximate solution of **overdetermined systems**, i.e. sets of equations in which there are **more equations than unknowns**.

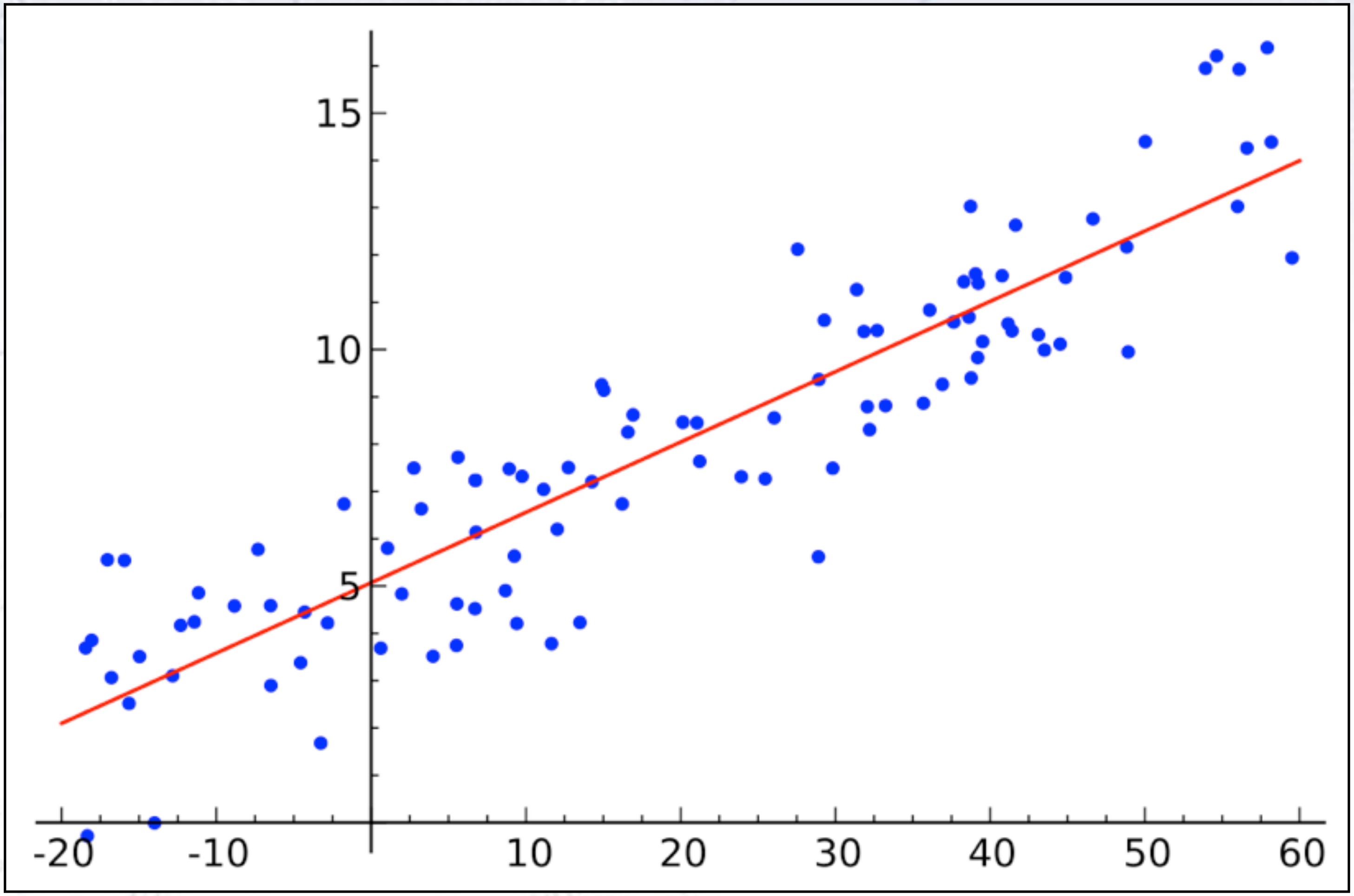


“Least squares” means that the overall solution minimises the **sum of the squares** of the errors made in solving every single equation.

The most important application is in **data fitting**. The best fit in the least-squares sense minimises the sum of squared residuals, a residual being the difference between an observed value and the fitted value provided by a model.

Method of Least Squares

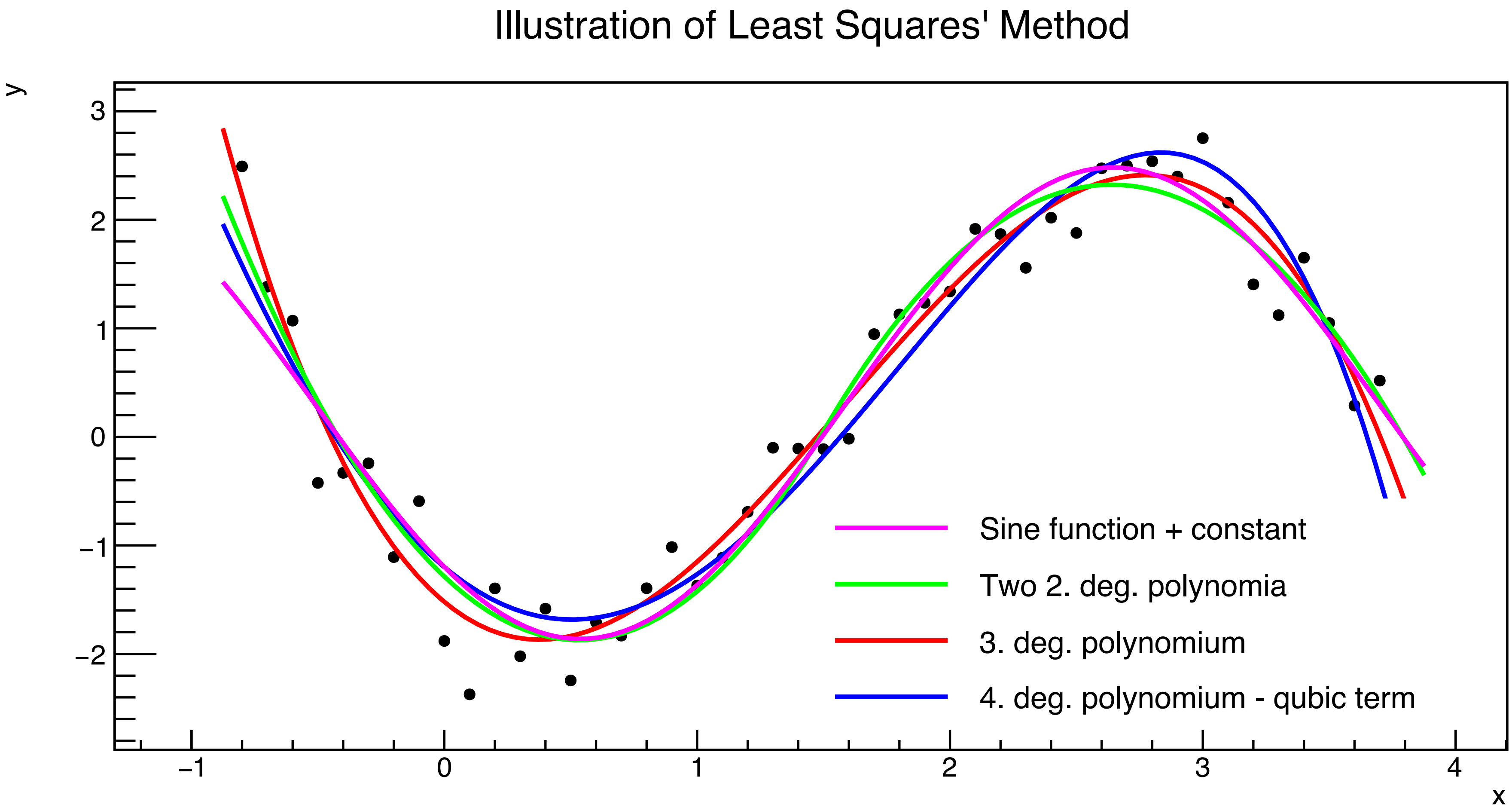
The problem at hand is determining the curve that best fitted data:



Originally, uncertainties were not included (not “invented” yet!)

Method of Least Squares

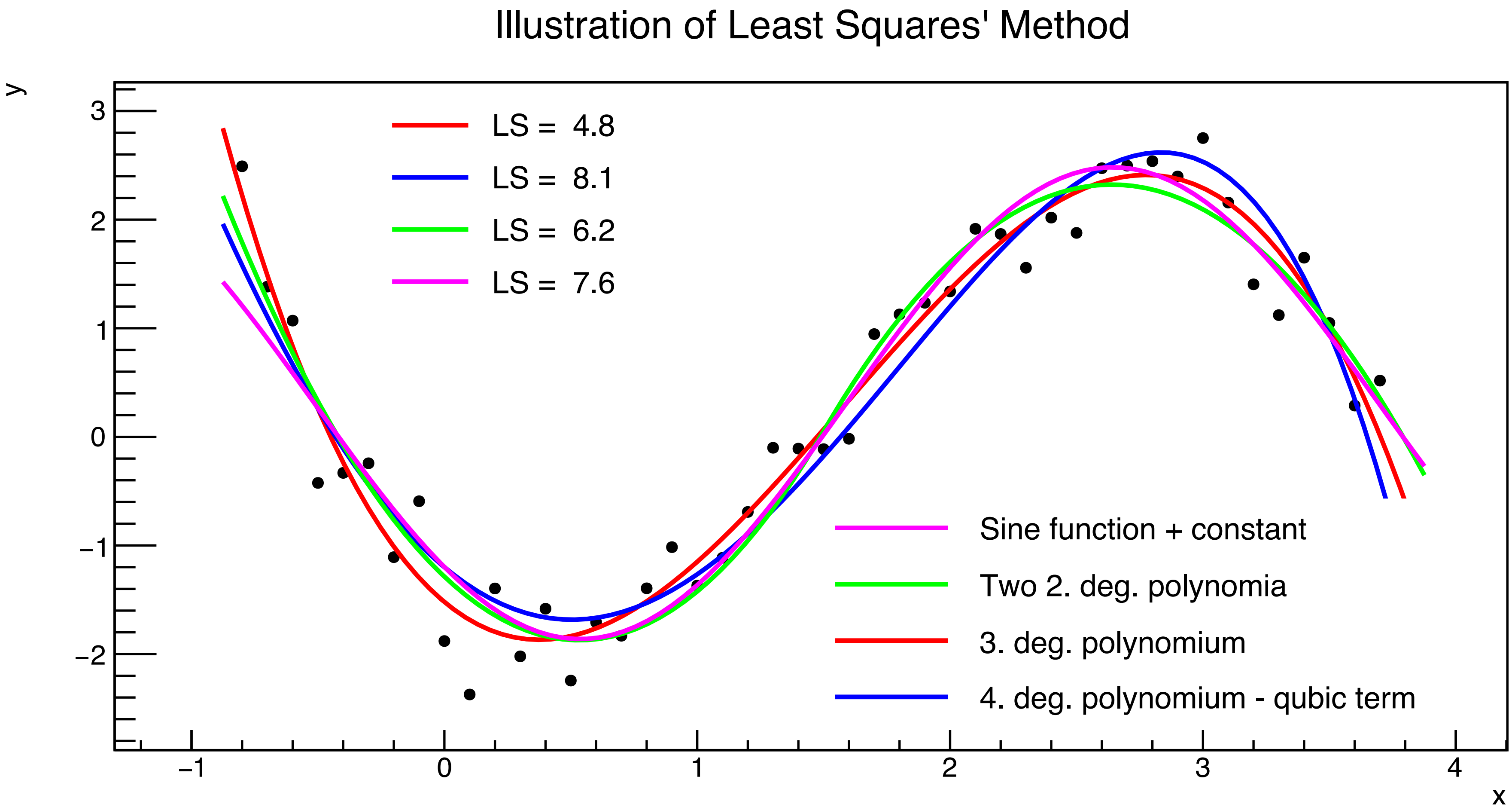
Look at the figure below, and determine which curve fits best…



Well, what do you define as "best"?

Method of Least Squares

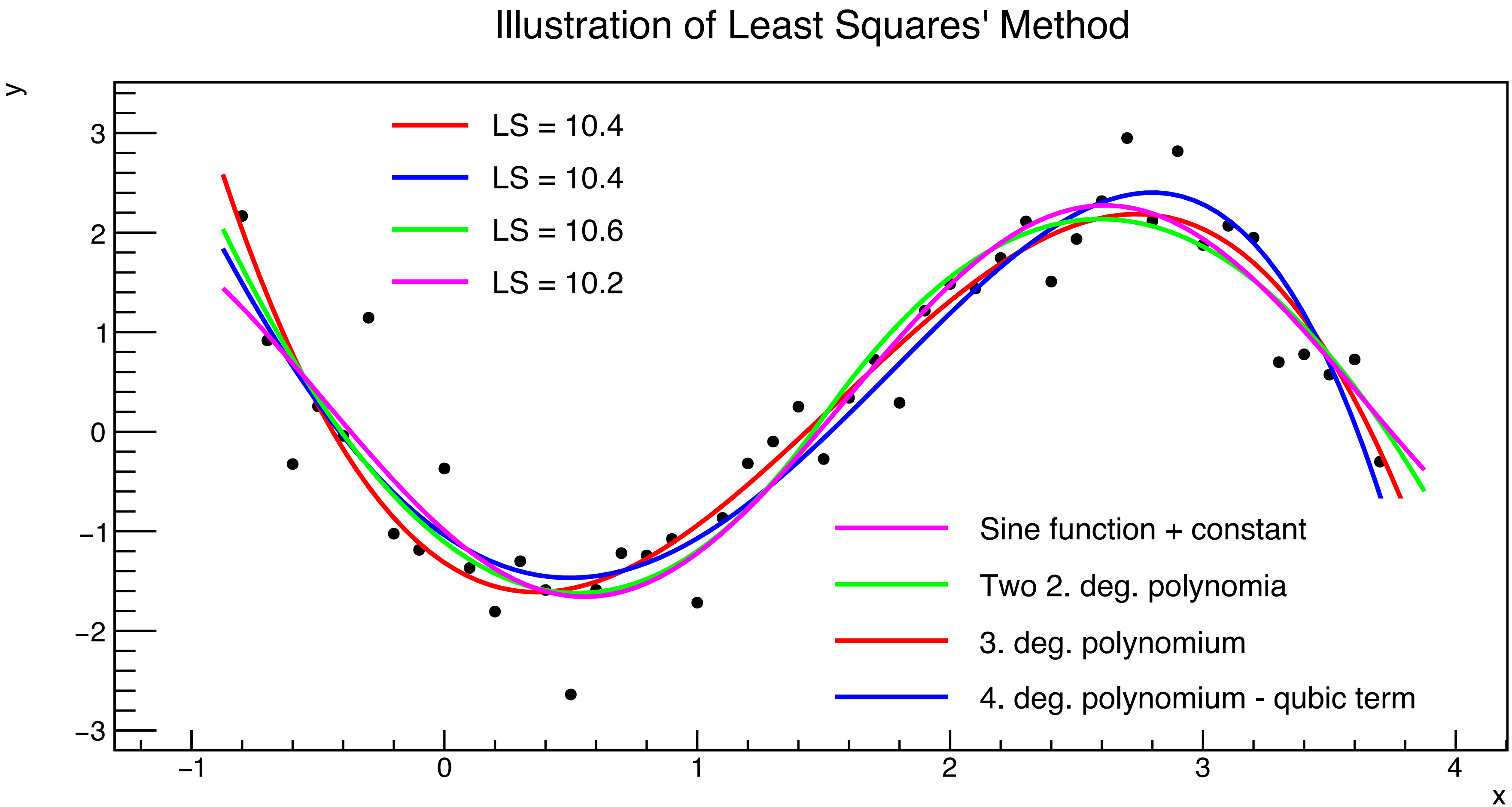
Look at the figure below, and determine which curve fits best...



Well, what do you define as "best"? And how good is it?!?

Method of Least Squares

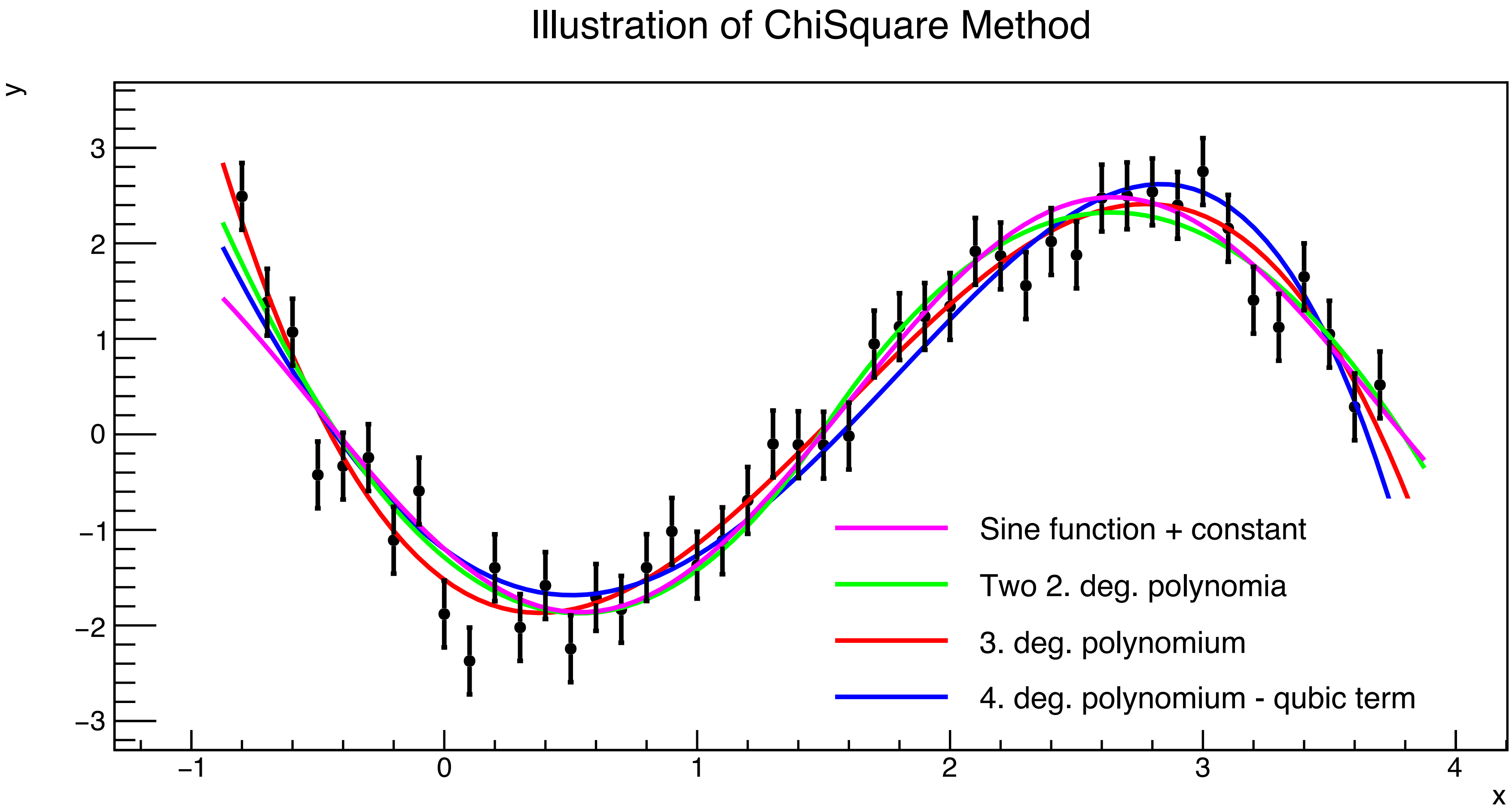
Look at the figure below, and determine which curve fits best...



Well, what do you define as “best”? And how good is it?!?

Chi-Square method

Look at the figure below, and determine which curve fits best...



Well, what do you define as "best"?

Defining the Chi-Square

<u>Problem Statement:</u> Given N data points (x, y, σ_y), adjust the parameter(s) θ of a model, such that it fits data best.

The best way to do this, given uncertainties σ_i on y_i is by minimising:

$$\chi^2(\theta) = \sum_i^N \frac{(y_i - f(x_i, \theta))^2}{\sigma_i^2}$$

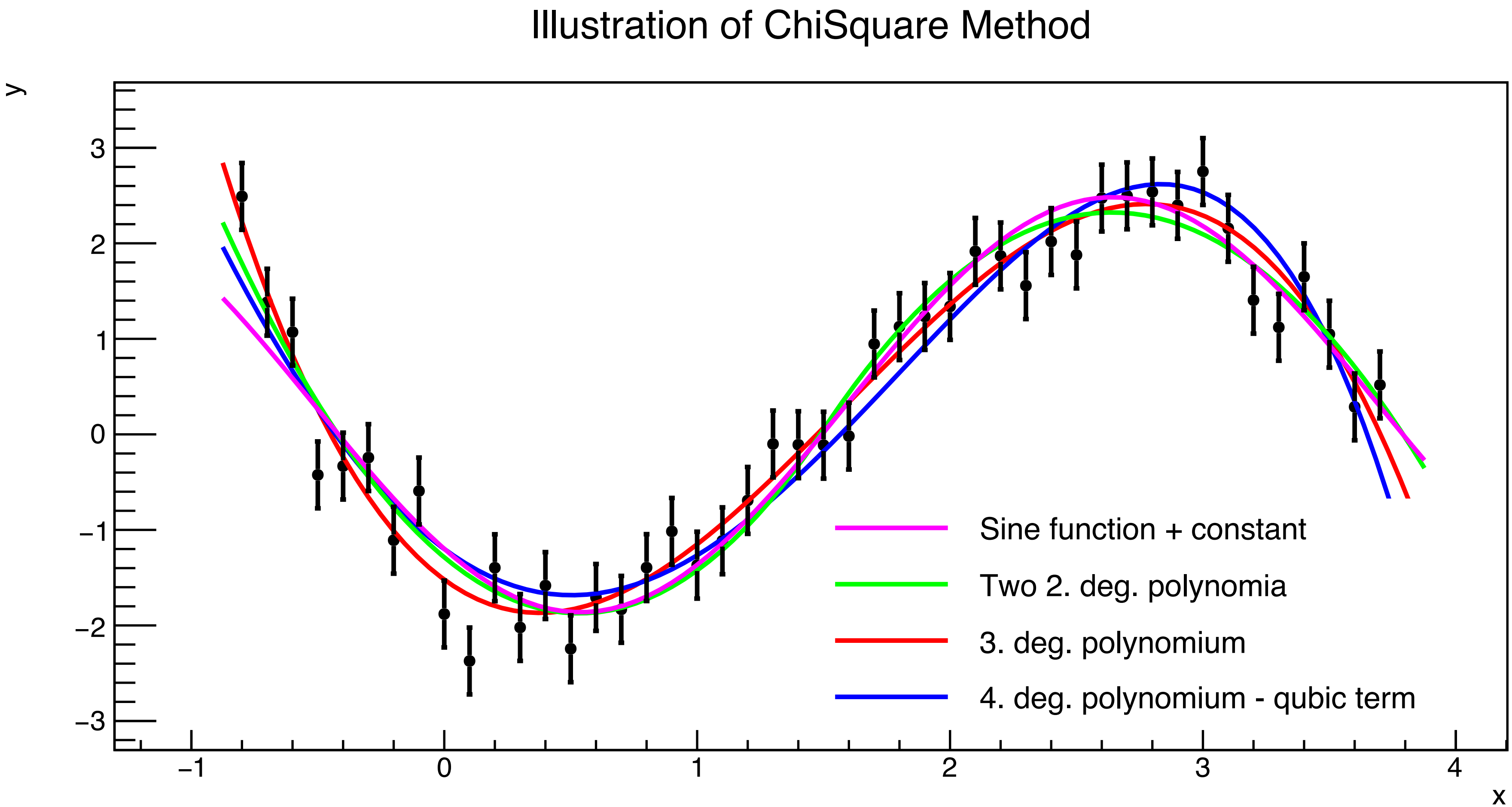
The power of this method is hard to overstate!

Not only does it provide a simple, elegant and unique way of fitting data, but more importantly it provides a **goodness-of-fit measure**.

This is the Chi-Square test!

Chi-Square method

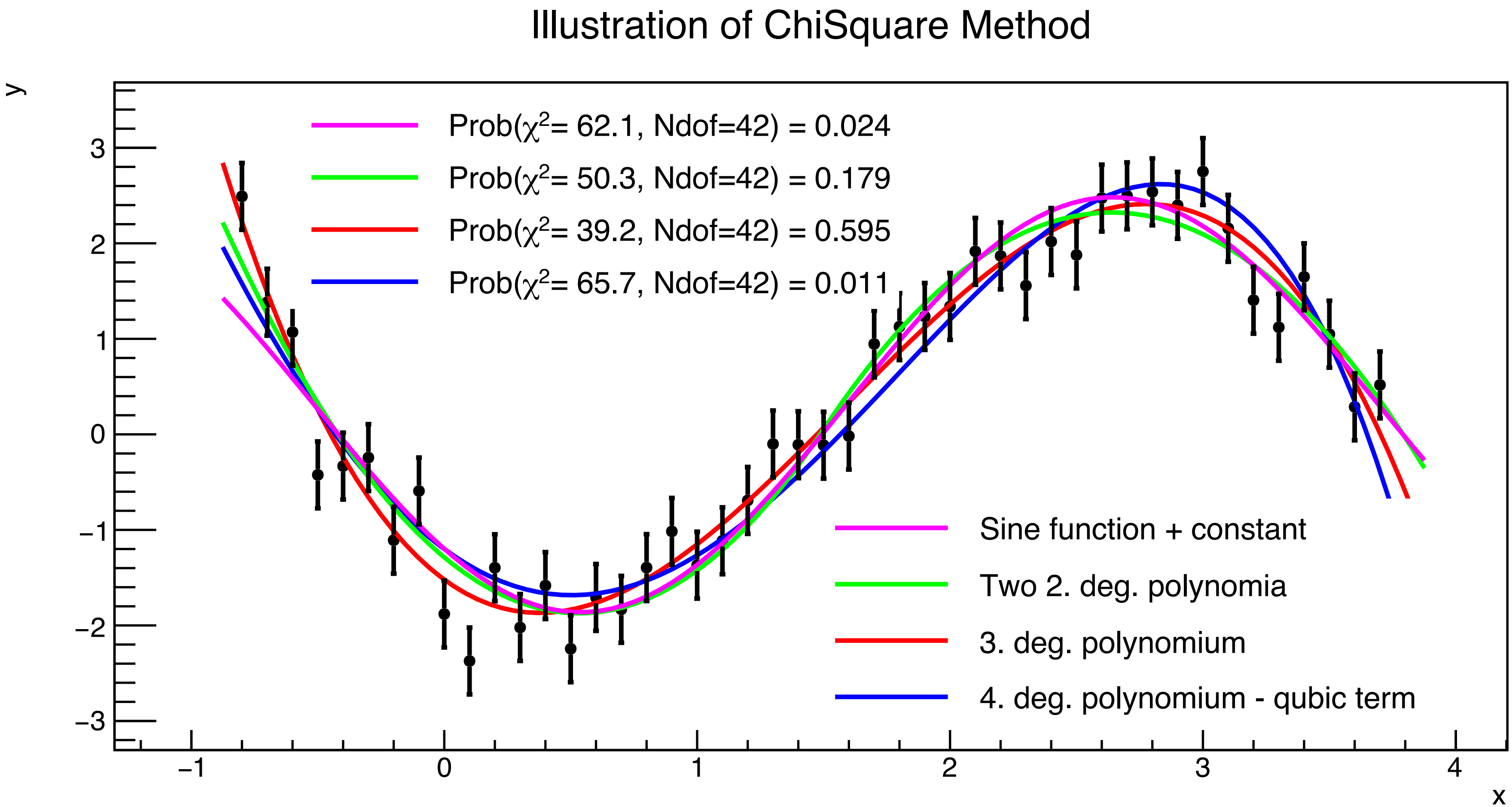
Look at the figure below, and determine which curve fits best...



Well, what do you define as "best"?

Chi-Square method

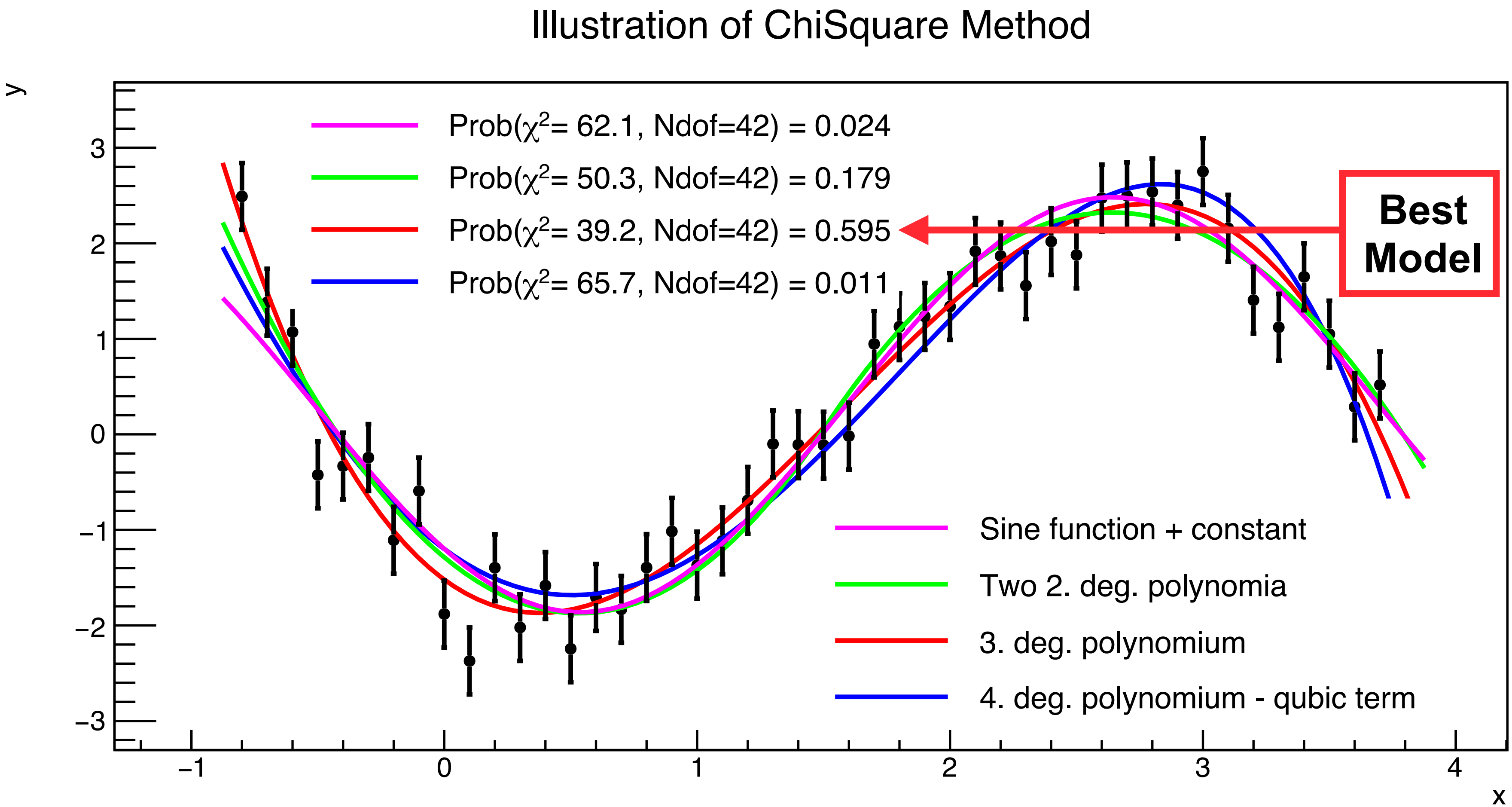
Look at the figure below, and determine which curve fits best…



Well, what do you define as "best"? The Chi2 quantifies this!

Chi-Square method

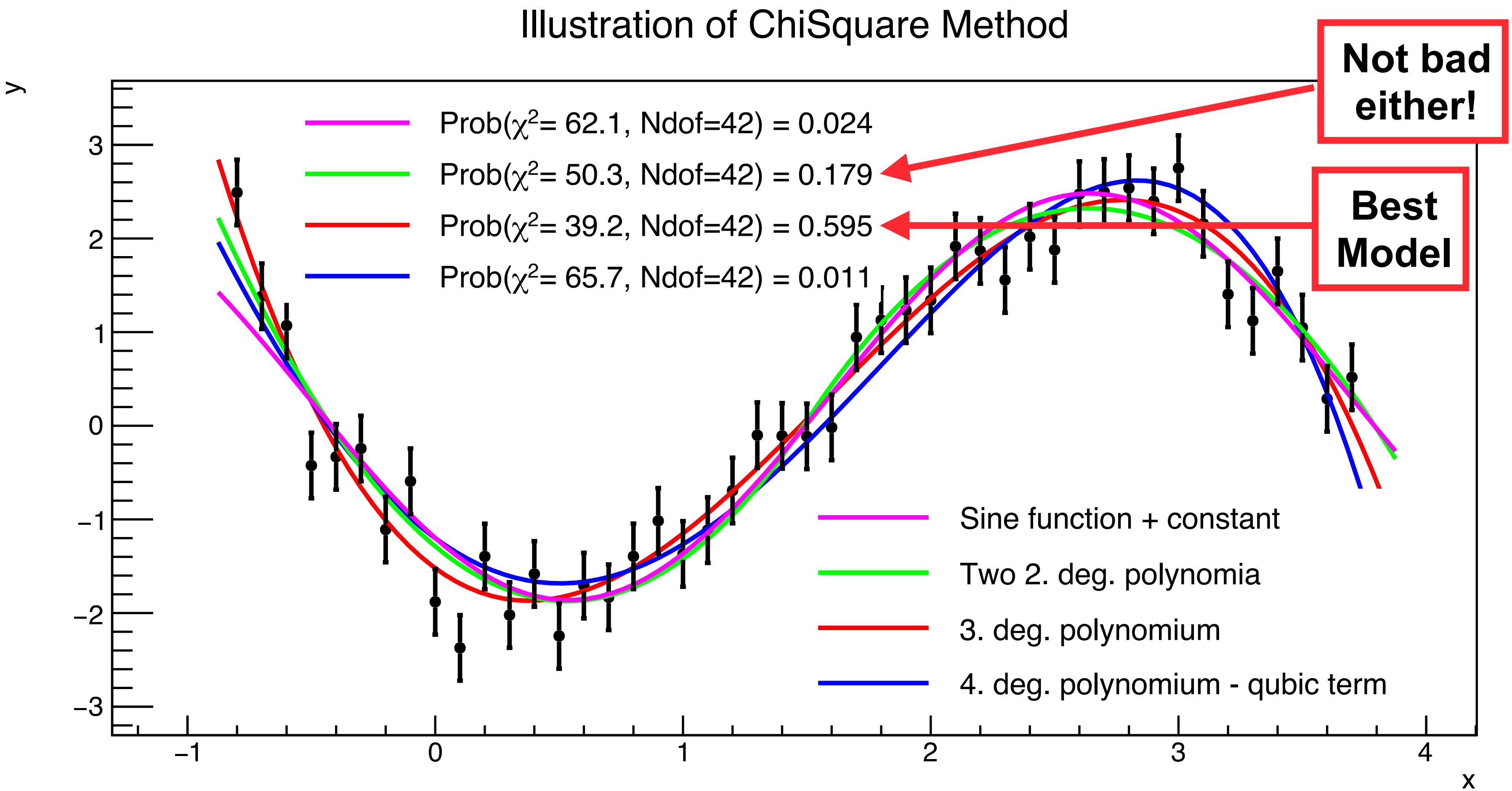
Look at the figure below, and determine which curve fits best...



Well, what do you define as "best"? The Chi2 quantifies this!

Chi-Square method

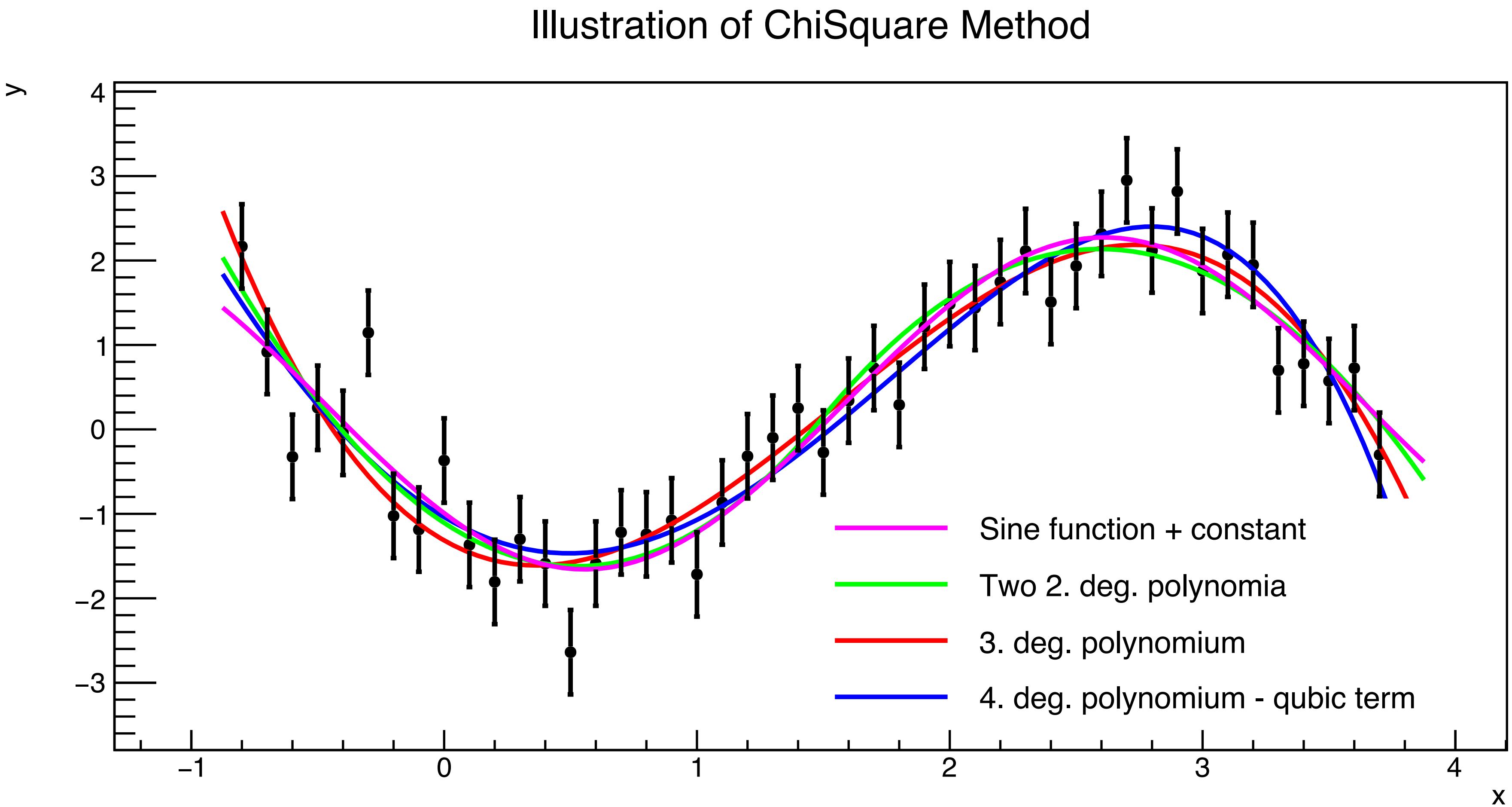
Look at the figure below, and determine which curve fits best...



Well, what do you define as "best"? The Chi2 quantifies this!

Chi-Square method

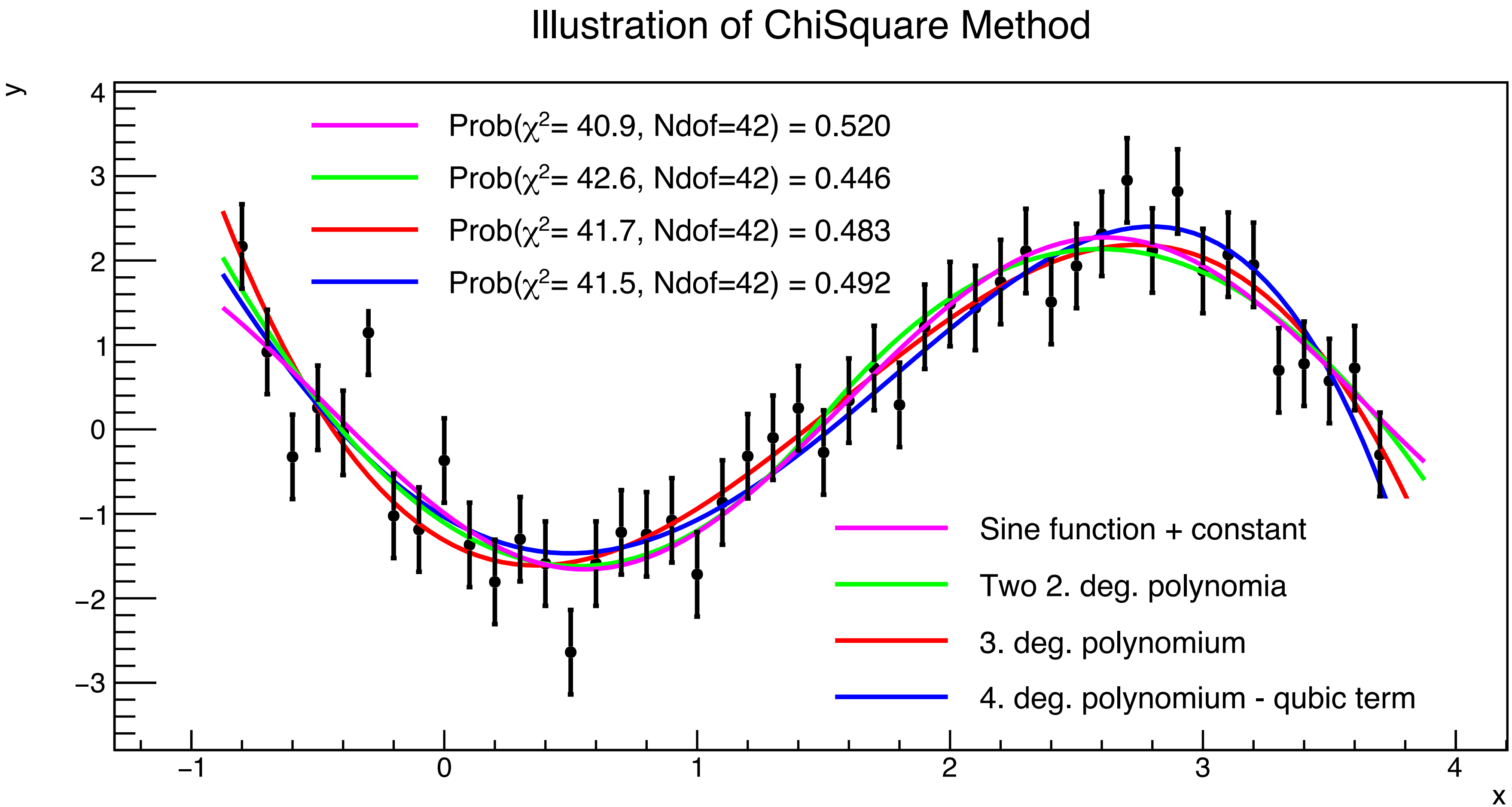
Look at the figure below, and determine which curve fits best…



What about now with **larger** errors?

Chi-Square method

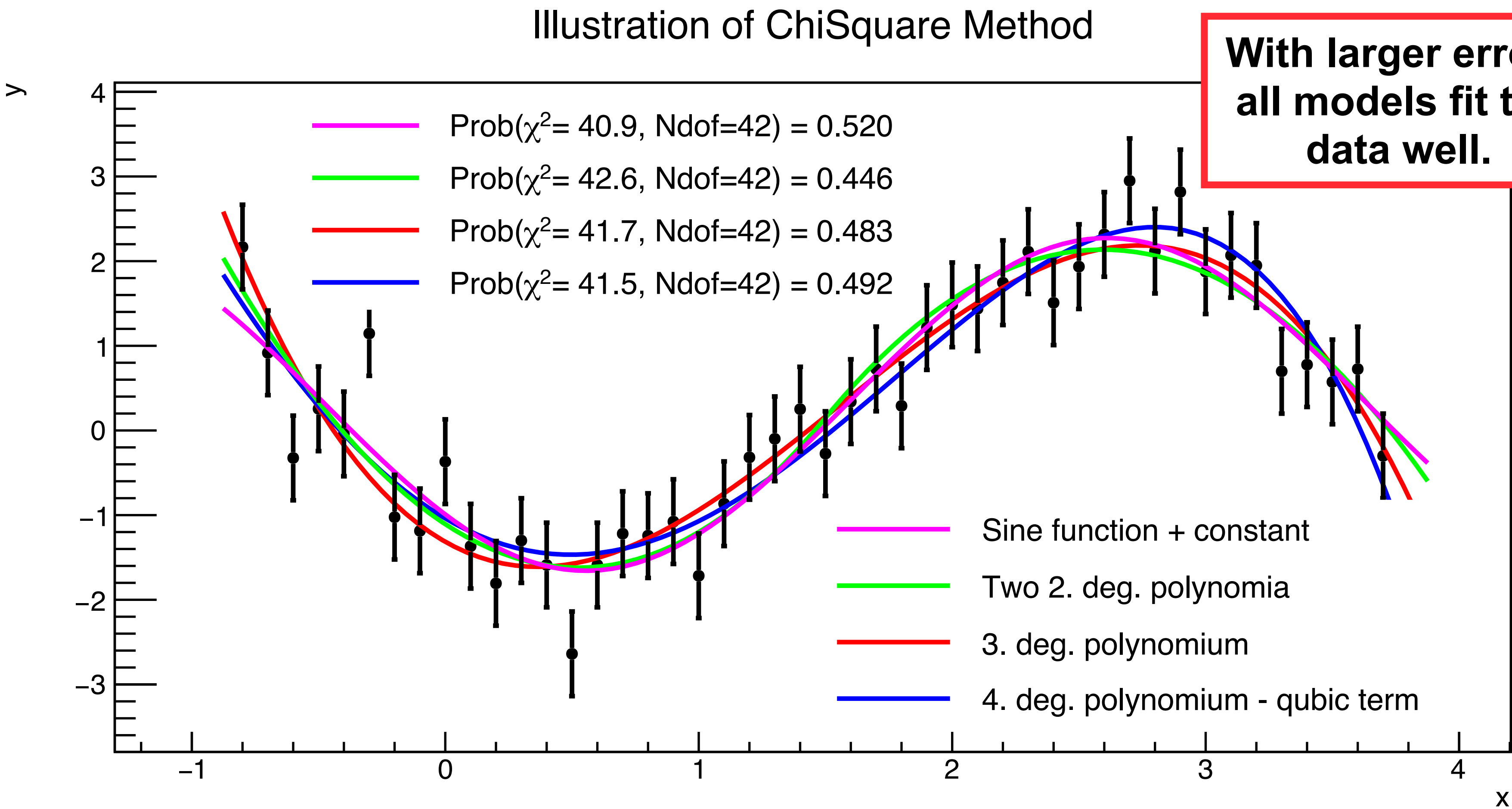
Look at the figure below, and determine which curve fits best...



What about now with **larger** errors?

Chi-Square method

Look at the figure below, and determine which curve fits best...

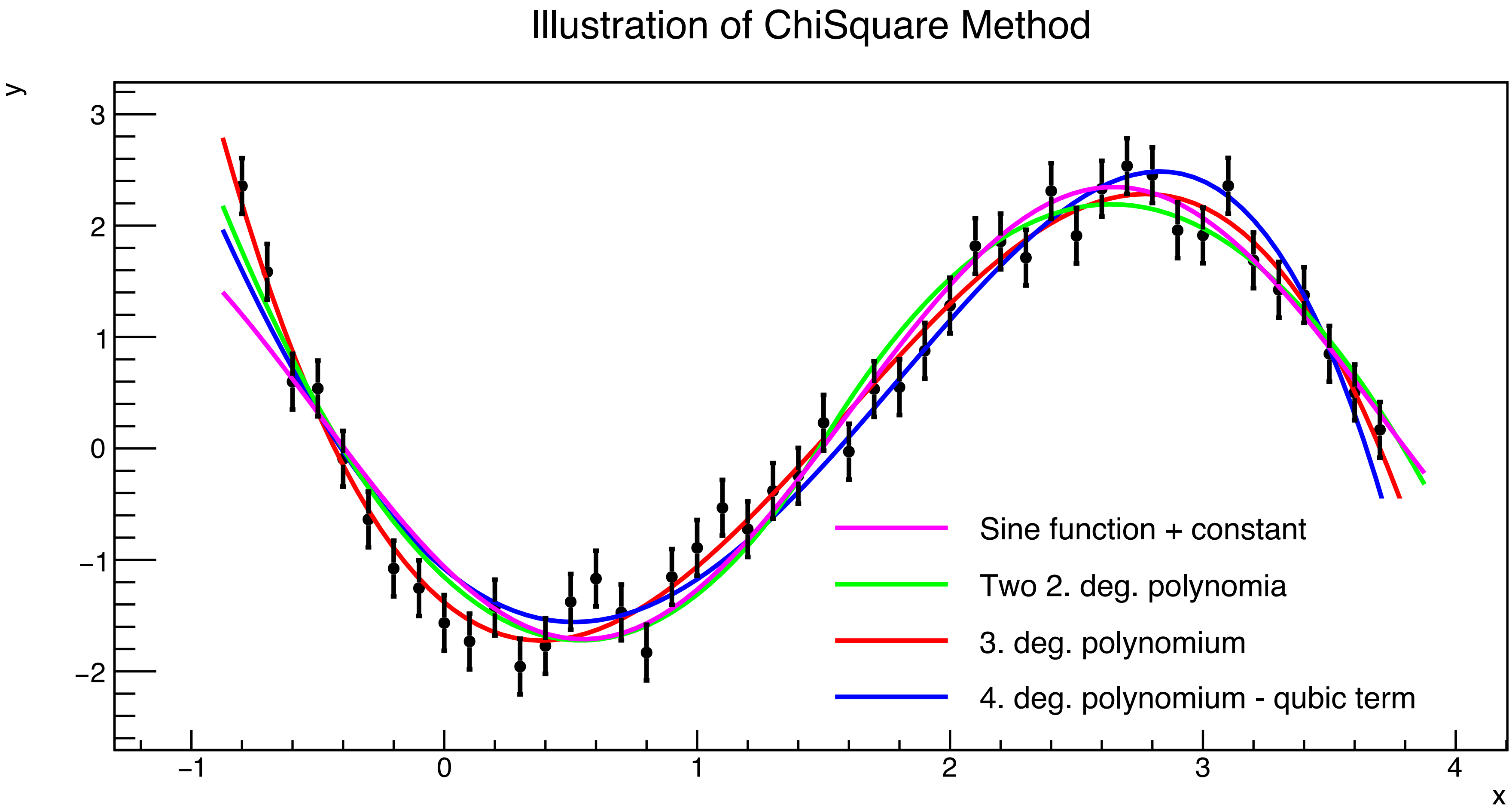


With larger errors all models fit the data well.

What about now with **larger** errors?

Chi-Square method

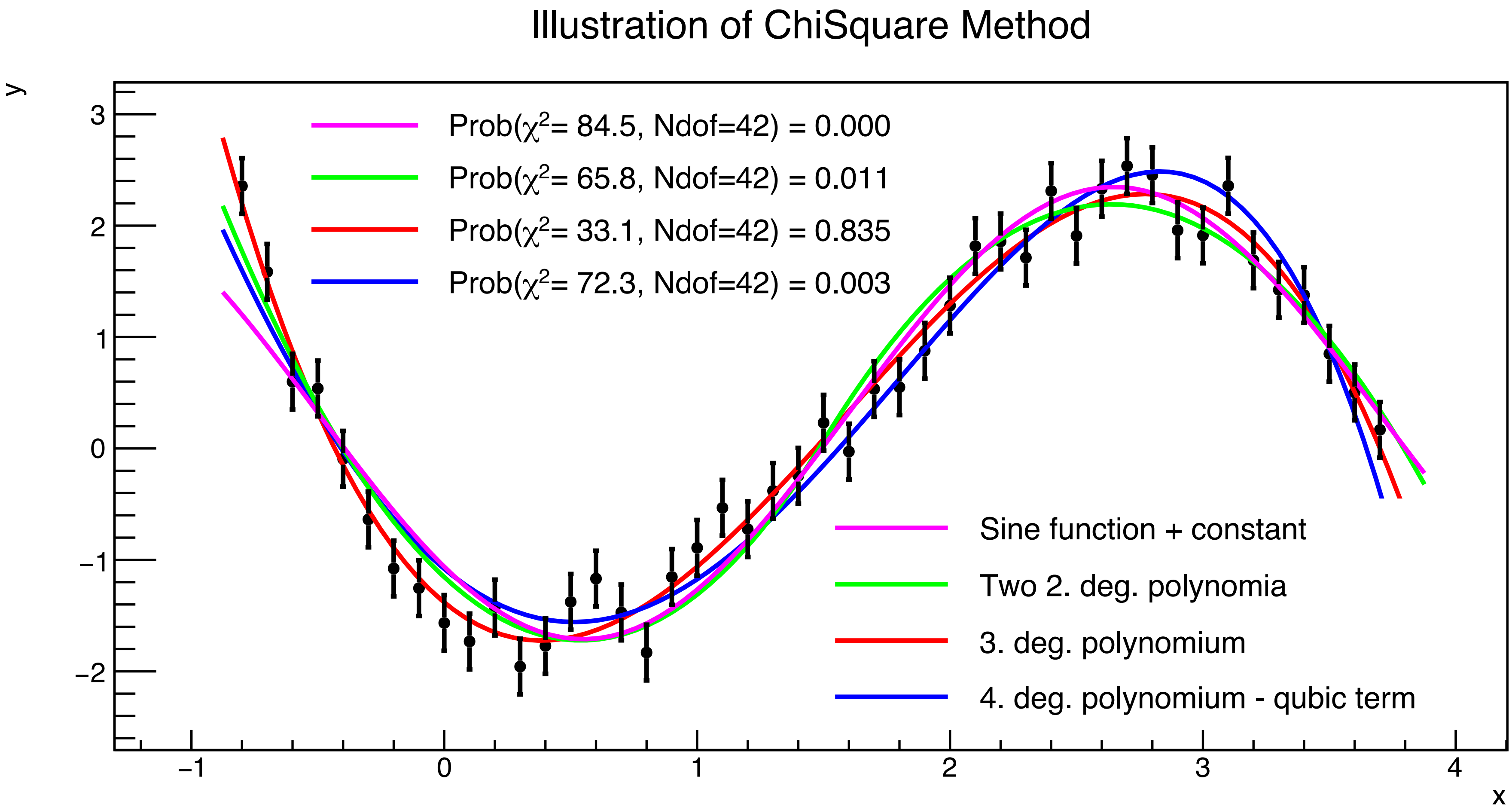
Look at the figure below, and determine which curve fits best...



What does **smaller** errors do?

Chi-Square method

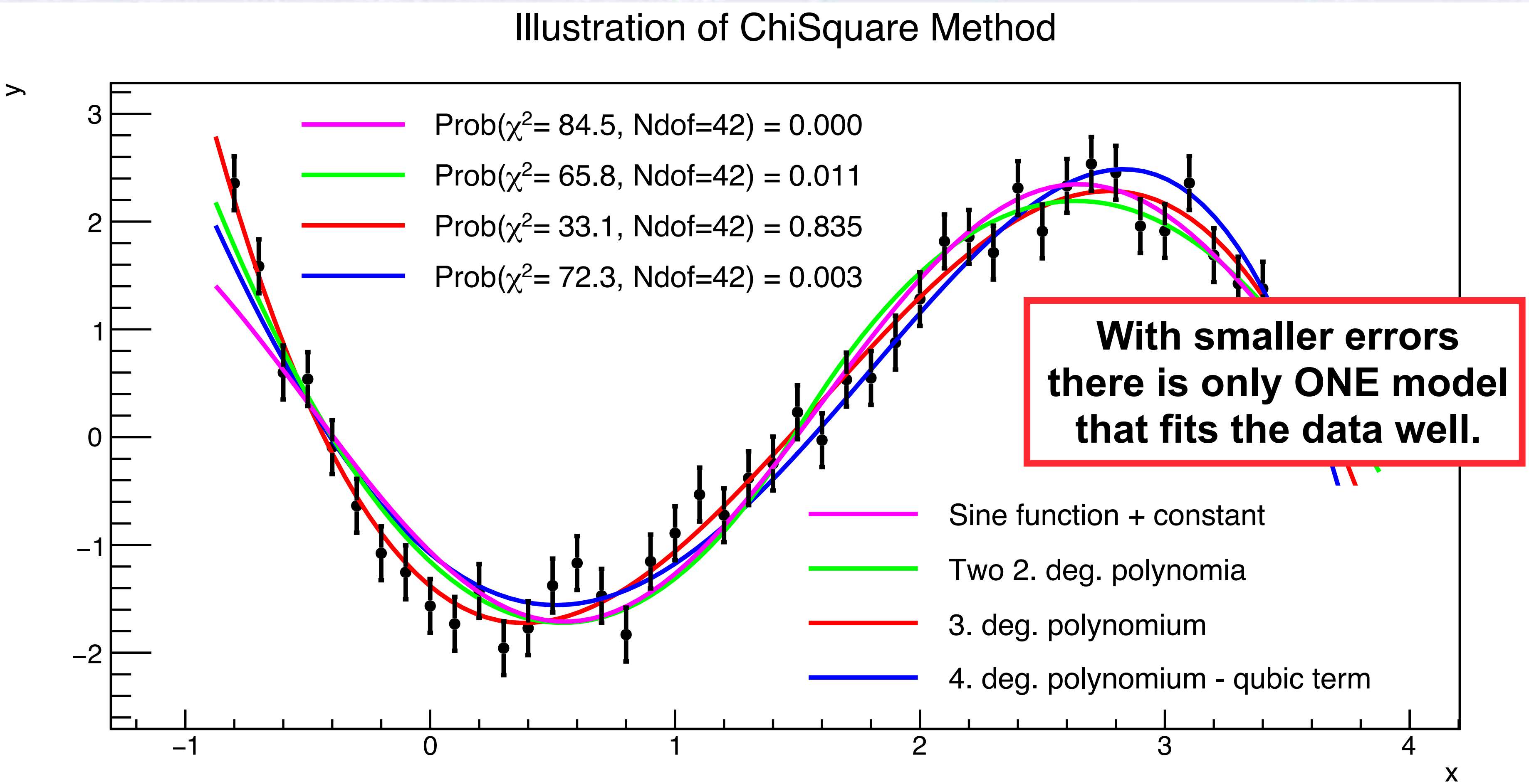
Look at the figure below, and determine which curve fits best...



What does **smaller** errors do?

Chi-Square method

Look at the figure below, and determine which curve fits best...



What does **smaller** errors do?

Defining the Chi-Square

Problem Statement: Given N data points (x,y,σ_y), adjust the parameter(s) θ of a model, such that it fits data best.

The best way to do this, given uncertainties σ_i on y_i is by minimising:

$$\chi^2(\theta) = \sum_i^N \frac{(y_i - f(x_i,\theta))^2}{\sigma_i^2}$$

The power of this method is hard to overstate!

Not only does it provide a simple, elegant and unique way of fitting data, but more importantly it provides a **goodness-of-fit measure**.

This is the Chi-Square test!

Defining the Chi-Square

Problem St

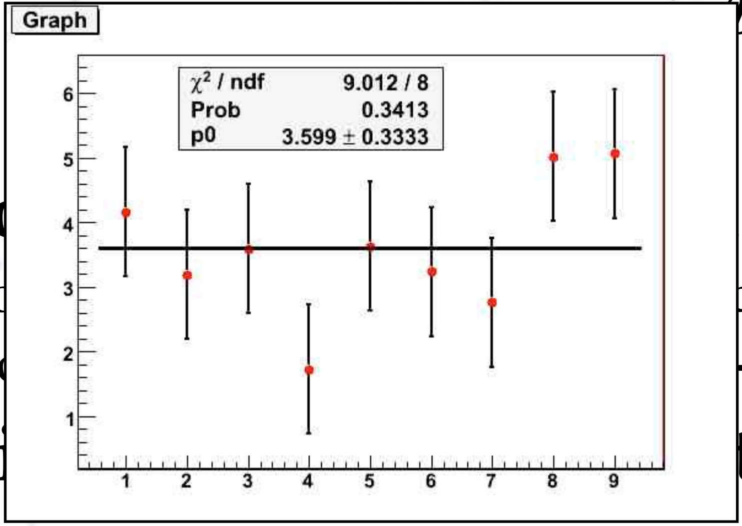
Note that when doing a weighted mean, one should check if the measurements agree with each other!
This can be done with a ChiSquare test.

arameter(s)

The best w

imising:

$$\chi^2(\theta) = \sum^{N} \frac{(y_i - f(x_i, \theta))^2}{\sigma_i^2}$$



The power **to overstate!**

Not only does it pro que way of fitting data, but more imp **-of-fit measure**.

Thi **test!**

Weighted mean & ChiSquare

The weighted mean is actually an **analytical ChiSquare minimisation to a constant**. The result is the same, and one can then calculate Prob(χ^2, Ndof).

<u>Example:</u>
Data (from pendulum experiment) could be four length measurement (in mm):

d : [17.8 ± 0.5, 18.1 ± 0.3, 17.7 ± 0.5, 17.7 ± 0.2]

The output from the above data is (many digits for *checks only*):

Mean	=	17.8098 mm
Error on mean	=	0.15057 mm
ChiSquare	=	1.28574
Ndof	=	3
Probability	=	0.7325213

NOTE: This seems a very nice (and precise) result, and it may very well be. BUT, it might also be, that we all four estimated it from the same photo or similarly, which could be biased by an angled view. Then we would be fooling ourselves. We will discuss such **"systematic uncertainties"** more!

Weighted mean & ChiSquare

The weighted mean is actually an **analytical ChiSquare minimisation to a constant**. The result is the same, and one can then calculate Prob(χ^2, Ndof).

Example:

Data (from pendulum experiment) could be four length measurement (in mm):

$$\mathbf{d : [17.8 \pm 0.5,\ 18.1 \pm 0.3,\ 17.7 \pm 0.5,\ 17.7 \pm 0.2]}$$

The output from the above data is (many digits for *checks only*):

Mean = 17.8098 mm
Error on mean = 0.15057 mm
ChiSquare = 1.28574
Ndof = 3
Probability = 0.7325213

d = (17.81 ± 0.15) mm
$p(\chi^2=1.3, N_{dof}=3) = 0.73$

NOTE: This seems a very nice (and precise) result, and it may very well be. BUT, it might also be, that we all four estimated it from the same photo or similarly, which could be biased by an angled view. Then we would be fooling ourselves. We will discuss such **"systematic uncertainties"** more!

Why the ChiSquare is great

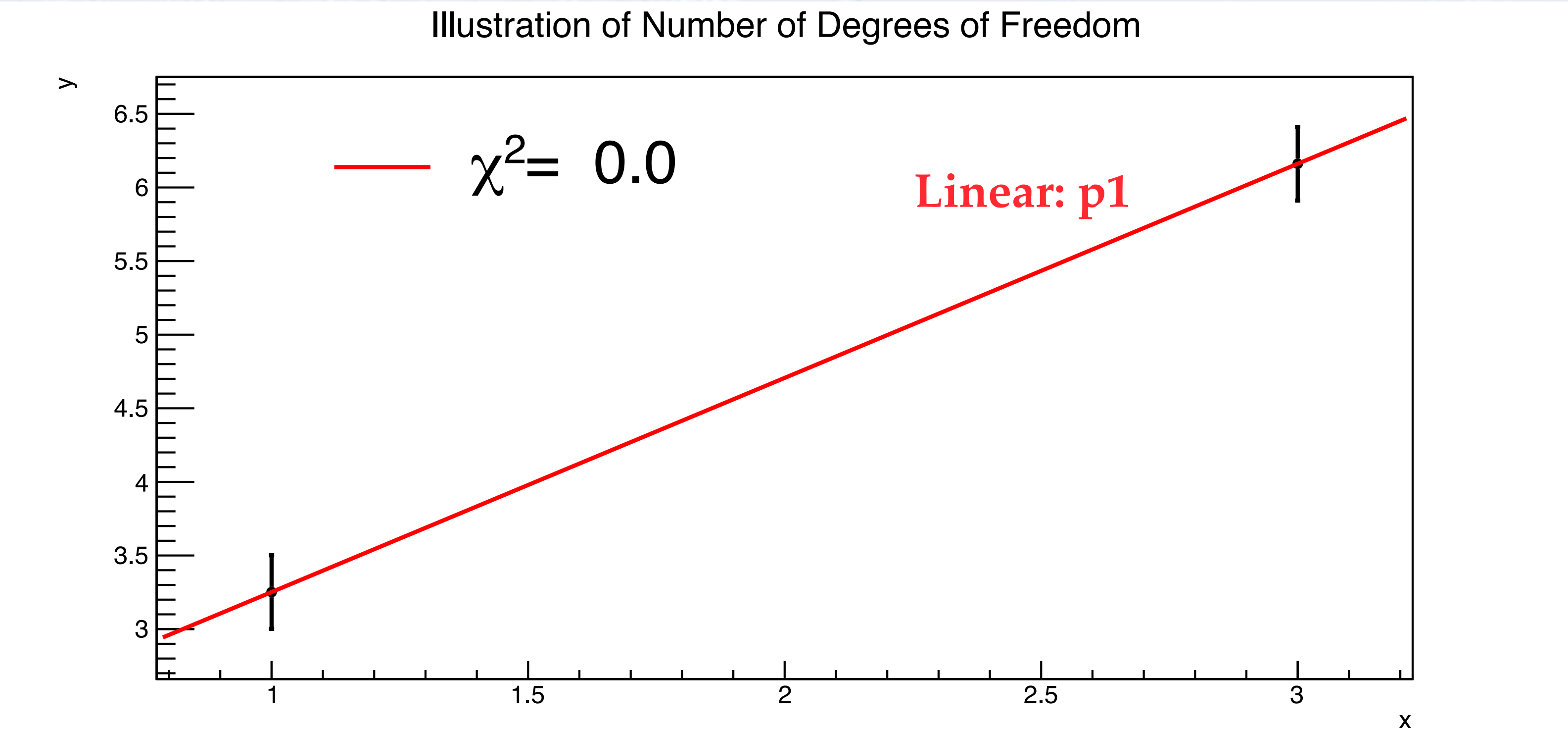
…but not its magic

Number of degrees-of-freedom

How to find/calculate the Number of degrees-of-freedom (Ndof) in a fit?

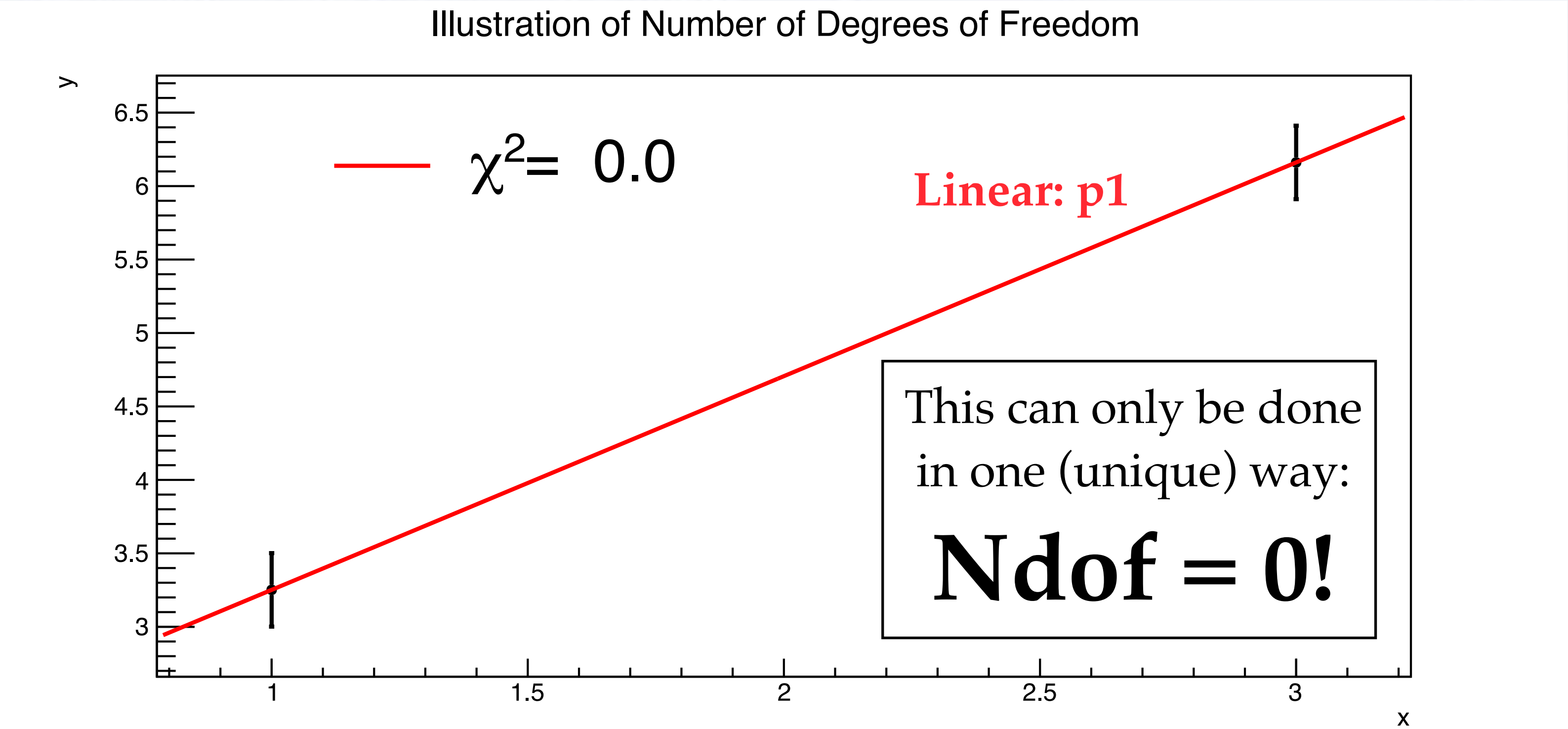
Number of degrees-of-freedom

How to find/calculate the Number of degrees-of-freedom (Ndof) in a fit?



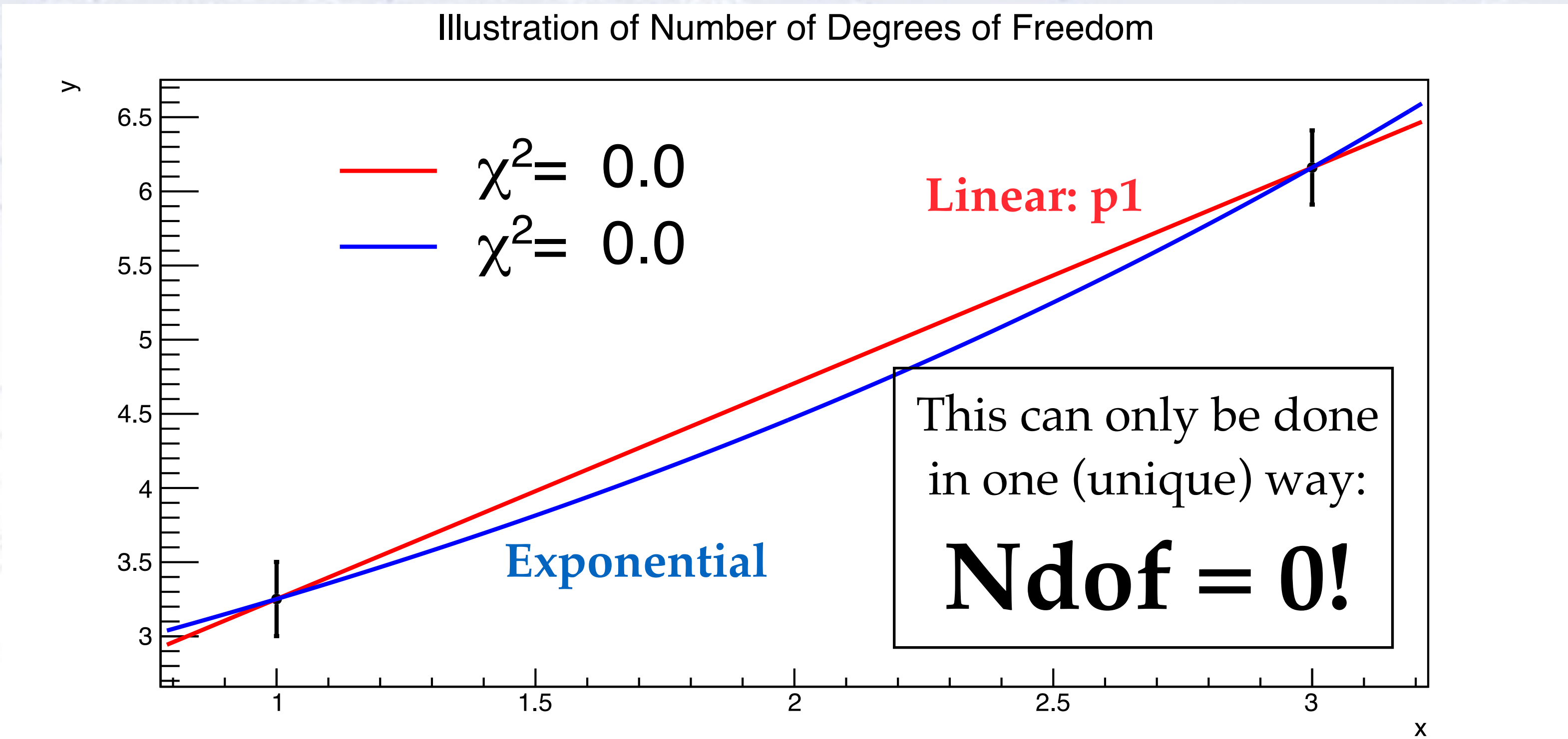
Number of degrees-of-freedom

How to find / calculate the Number of degrees-of-freedom (Ndof) in a fit?



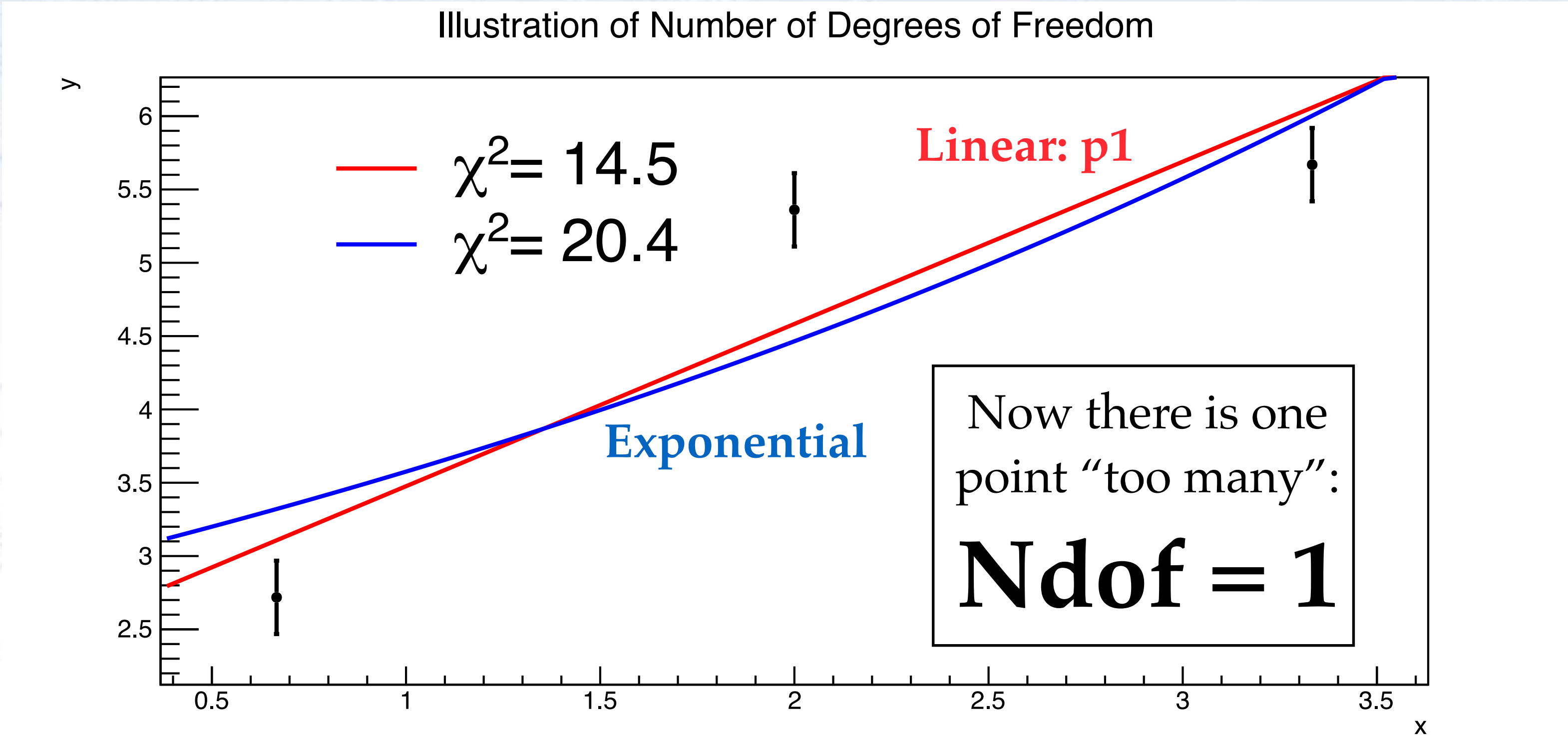
Number of degrees-of-freedom

How to find / calculate the Number of degrees-of-freedom (Ndof) in a fit?



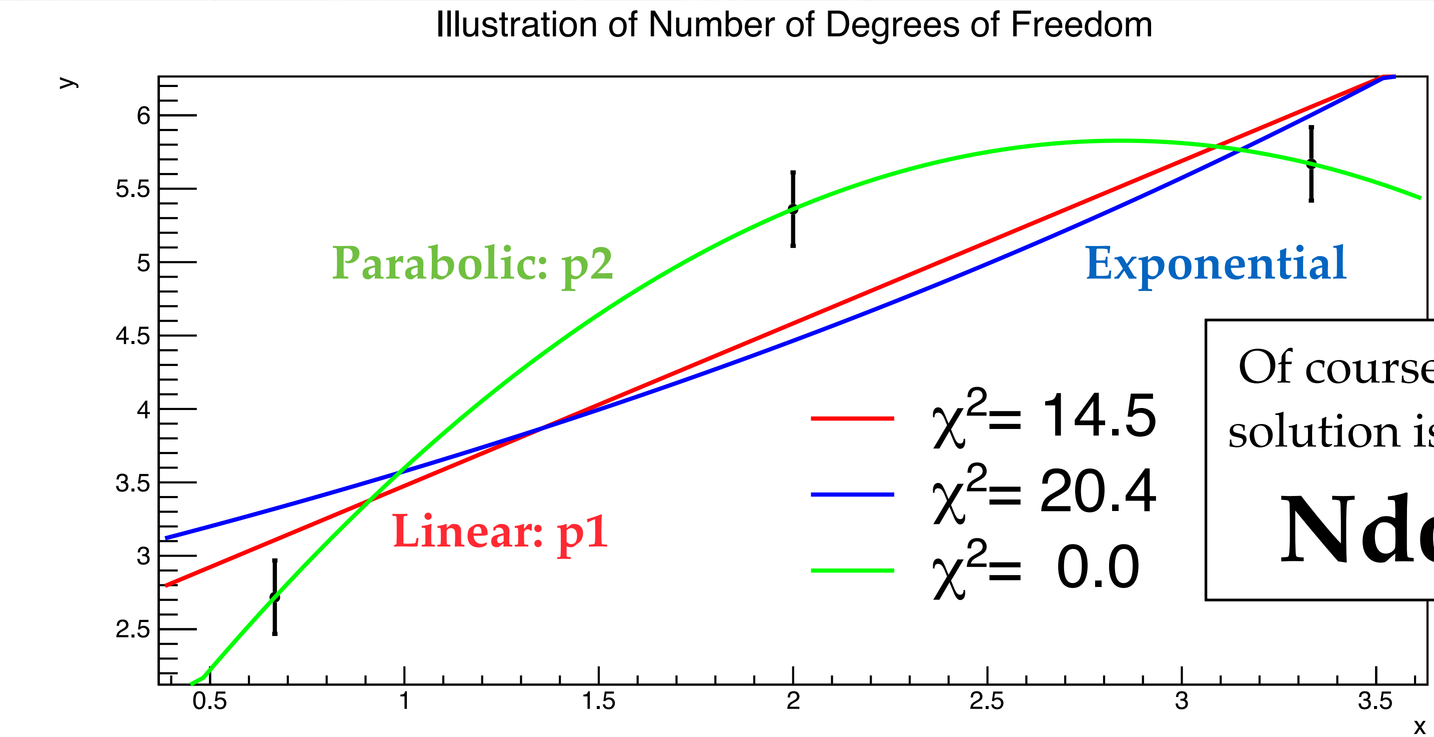
Number of degrees-of-freedom

How to find / calculate the Number of degrees-of-freedom (Ndof) in a fit?



Number of degrees-of-freedom

How to find / calculate the Number of degrees-of-freedom (Ndof) in a fit?



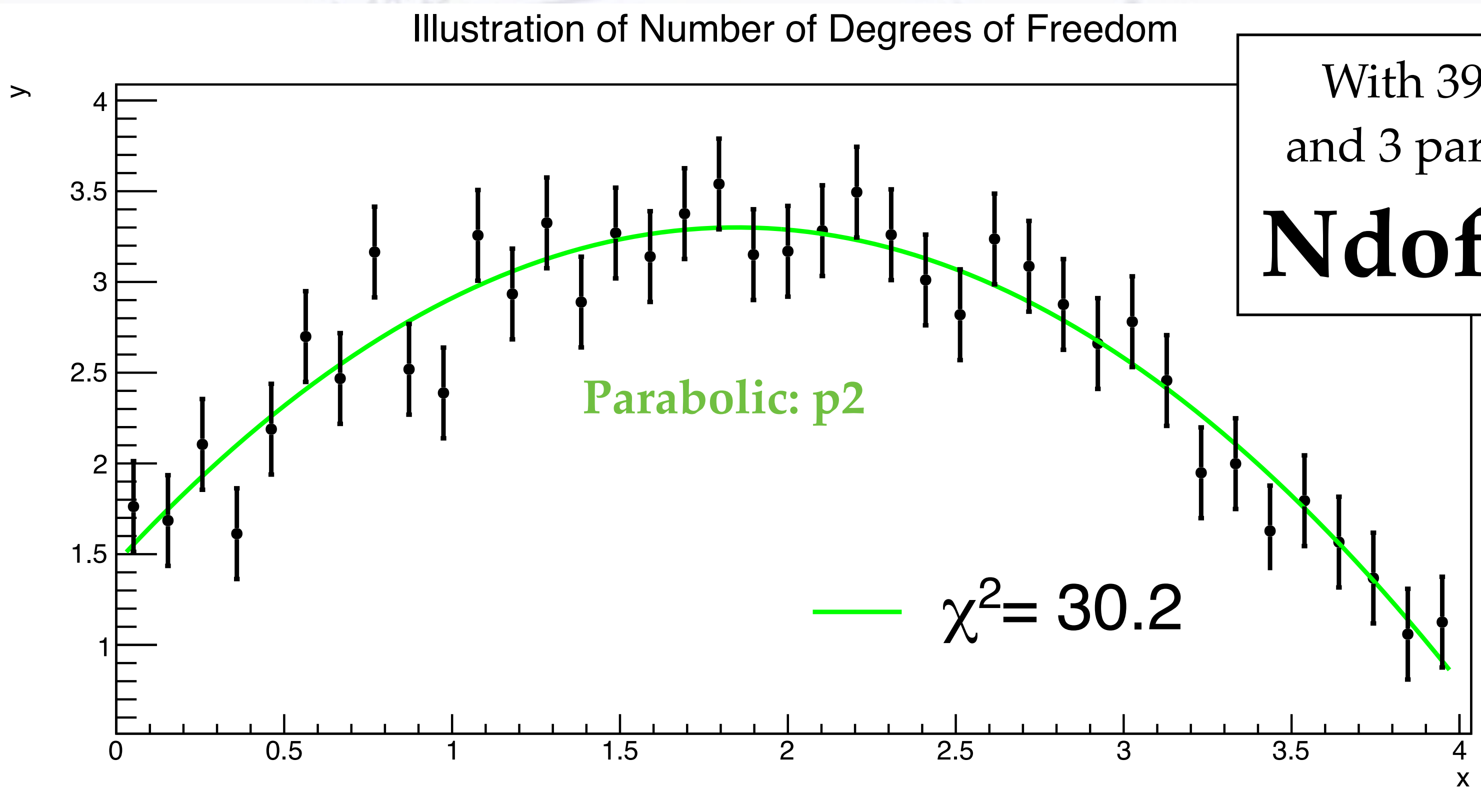
Of course for pol2 the solution is still unique:

Ndof = 0

Number of degrees-of-freedom

The number of degrees-of-freedom, Ndof, can be calculated as the number of points in the fit minus the number of parameters in the fit function:

$$N_{\mathrm{dof}} = N_{\mathrm{data\ points}} - N_{\mathrm{fit\ variables}}$$

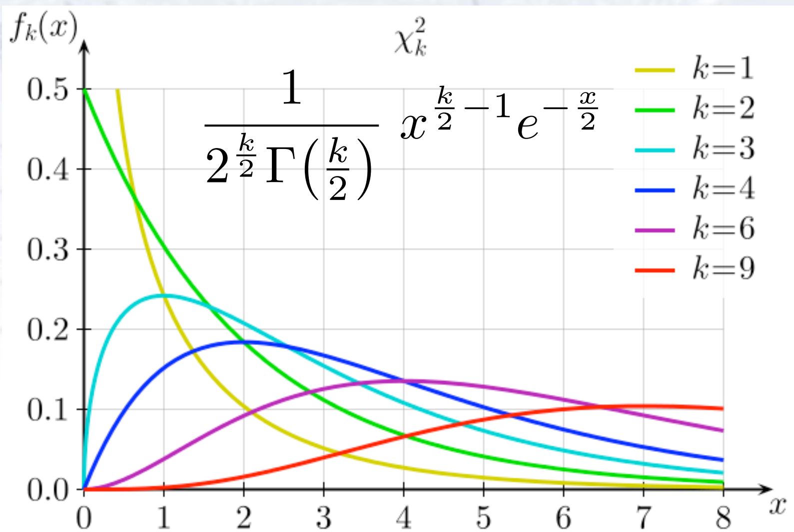


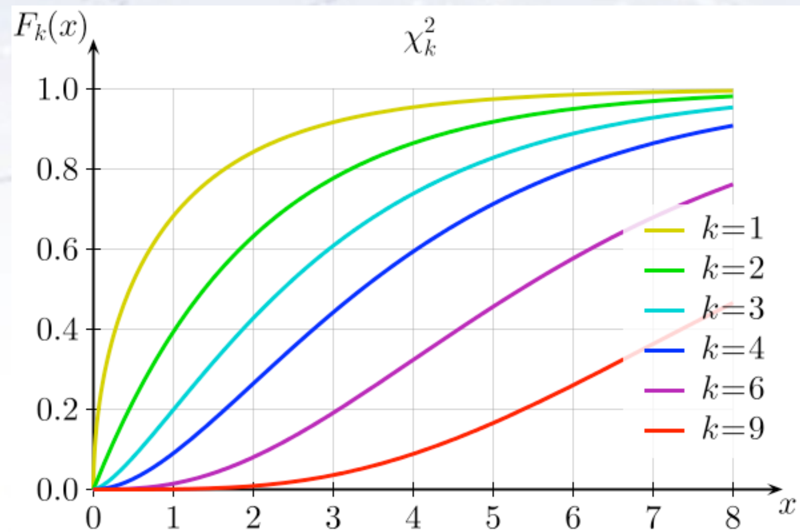
With 39 points and 3 parameters:

Ndof = 36

The Chi-Square distribution and test

The **Chi-Square distribution** for N_{dof} degrees of freedom is the distribution of the sum of the squares of N_{dof} normally distributed random variables.





The **Chi-Square test** consists of comparing the Chi-Square value obtained from a fit with the PDF of expected Chi-Square values. This allows the calculation of the *probability* of observing something with the same Chi-Square value or higher...

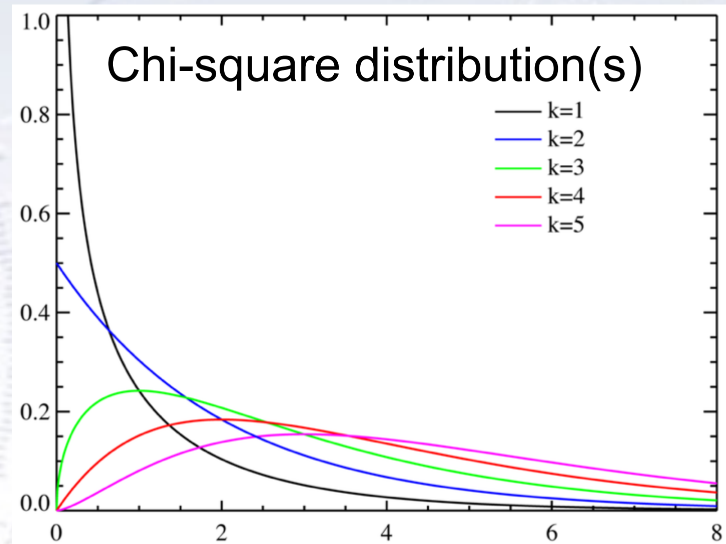
Rule of thumb: Chi-Square should roughly match N_{dof}

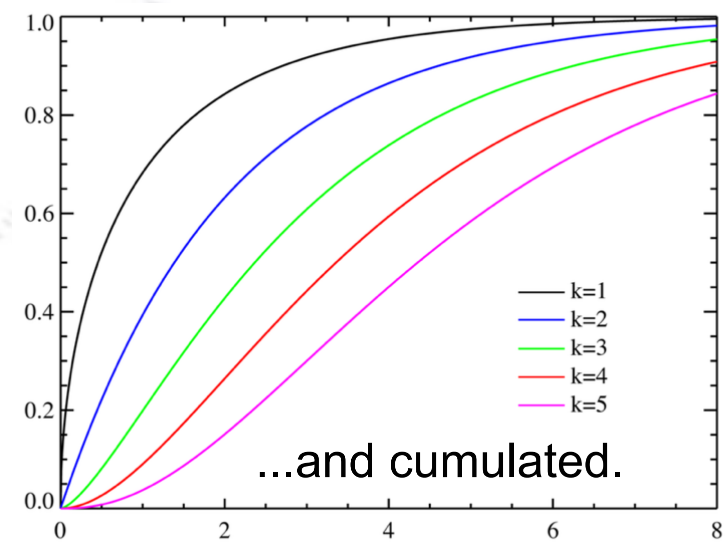
Chi-Square probability calculation

Given a **Chi-square value** and a **number of degrees of freedom** (Ndof), one can obtain a **"goodness-of-fit"**.

It is known, what Chi-square values to expect given the Ndof. One can therefore compare to this (Chi-square) distribution, and see...

what is the probability of getting this Chi-square value or something worse, assuming this is the correct fit function!





Example:

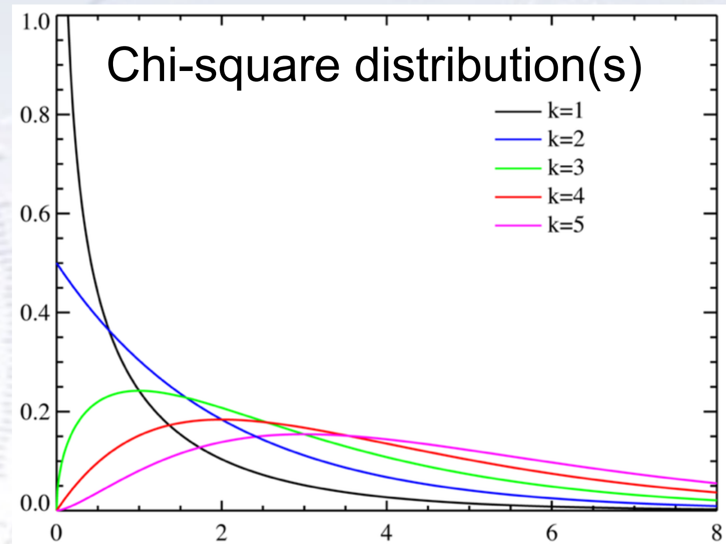
A fit gave the Chi-square 7.1 with 5 dof. The chance of getting this Chi-square or worse is... (reading the pink bottom curve (Ndof = k = 5) at 7.1)...

Chi-Square probability calculation

Given a **Chi-square value** and a **number of degrees of freedom** (Ndof), one can obtain a **"goodness-of-fit"**.

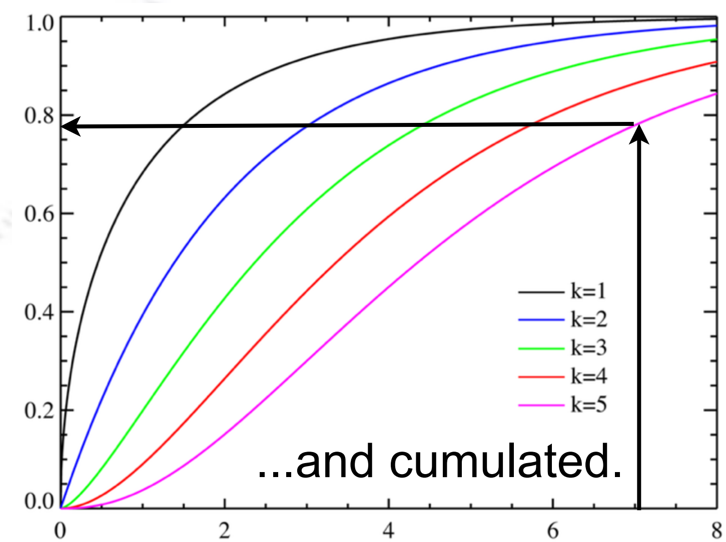
It is known, what Chi-square values to expect given the Ndof. One can therefore compare to this (Chi-square) distribution, and see...

what is the probability of getting this Chi-square value or something worse, assuming this is the correct fit function!



<u>Example:</u>

A fit gave the Chi-square 7.1 with 5 dof. The chance of getting this Chi-square or worse is... (reading the pink bottom curve (Ndof = k = 5) at 7.1)... 1 - 0.78 = **22%**



Chi-Square probability calculation

In the table below, one can get a quick estimate for low N_{dof}.

Degrees of freedom (df)	χ^2 value [16]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
	Non-significant								**Significant**		

Chi-Square probability calculation

In the table below, one can get a quick estimate for low N_{dof}.

Degrees of freedom (df)	χ^2 value [16]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	9.49	13.28	18.47
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
	Non-significant								**Significant**		

Python:

```
chi2_prob = stats.chi2.sf(chi2_value, NDOF)
```

sf (survival function) = 1 - CDF

Chi-Square probability interpretation

The Chi-Square probability can **roughly** be interpreted as follows:

- If χ^2 / **Ndof ≃ 1** or more precisely if **0.01 < p(χ^2,Ndof) < 0.99**, then **all is good**.
- If χ^2 / **Ndof ≫ 1** or more precisely if **p(χ^2,Ndof) < 0.01**, then your fit is probably bad! Four potential reasons: **Hypothesis/model wrong, data is faulty, errors too small or unlucky!**
- If χ^2 / **Ndof ≪ 1** or more precisely if **0.99 < p(χ^2,Ndof)**, then your fit is TOO good! Two potential reasons: **Overestimated uncertainties or lucky!**

If the statistics behind the plot is VERY high (great than 10^6), then you might have a hard time finding a model, which truly describes all the features in the plot (as now tiny effects become visible), and one hardly ever gets a good Chi-Square probability. However, in this case, one should not worry too much, unless very high precision is wanted.

Chi-Square probability interpretation

<u>The Chi-Square probability can **roughly** be interpreted as follows:</u>

- If χ^2 / Ndof ≃ 1 or more precisely if $0.01 < p(\chi^2,\text{Ndof}) < 0.99$, then **all is good**.
- If χ^2 / Ndof ≫ 1 or more precisely if $p(\chi^2,\text{Ndof}) < 0.01$, then your fit is probably bad! Four potential reasons: **Hypothesis/model wrong, data is faulty, errors too small or unlucky!**
- If χ^2 / Ndof ≪ 1 or more precisely if $0.99 < p(\chi^2,\text{Ndof})$, then ...
  Over...

If the st... you
might h... the
feature... hardly
ever ge... he
should ... .

Note: One should only use $\chi^2 \sim N_{dof}$ as a rule-of-thumb, and be cautious anyway:

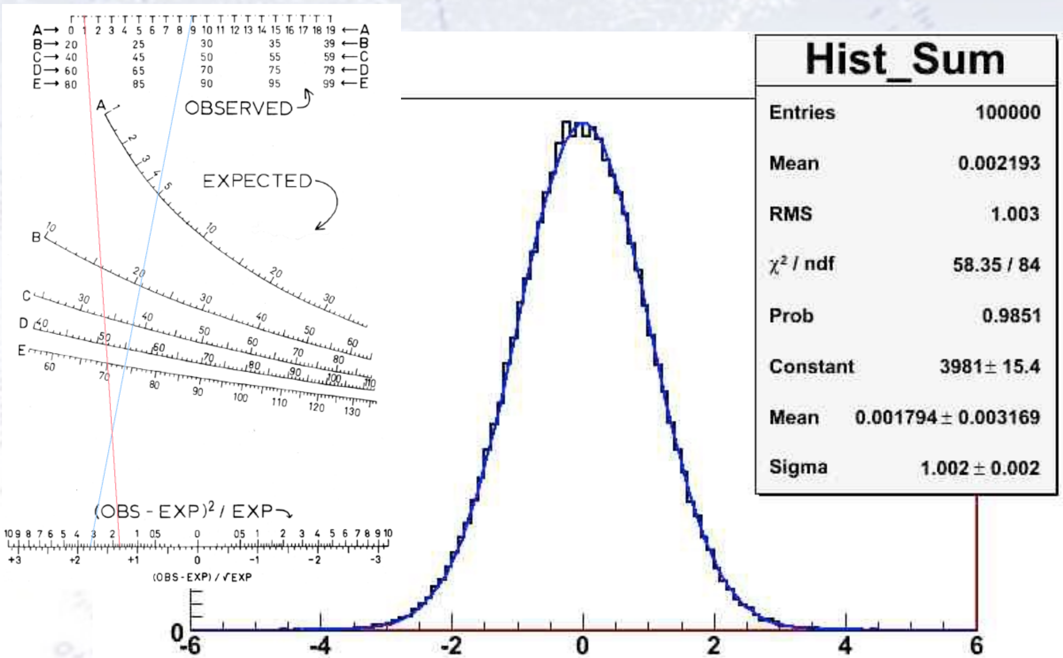
$$\text{Prob}(\chi^2{=}3.0,\ N_{dof}{=}2) = 0.223$$

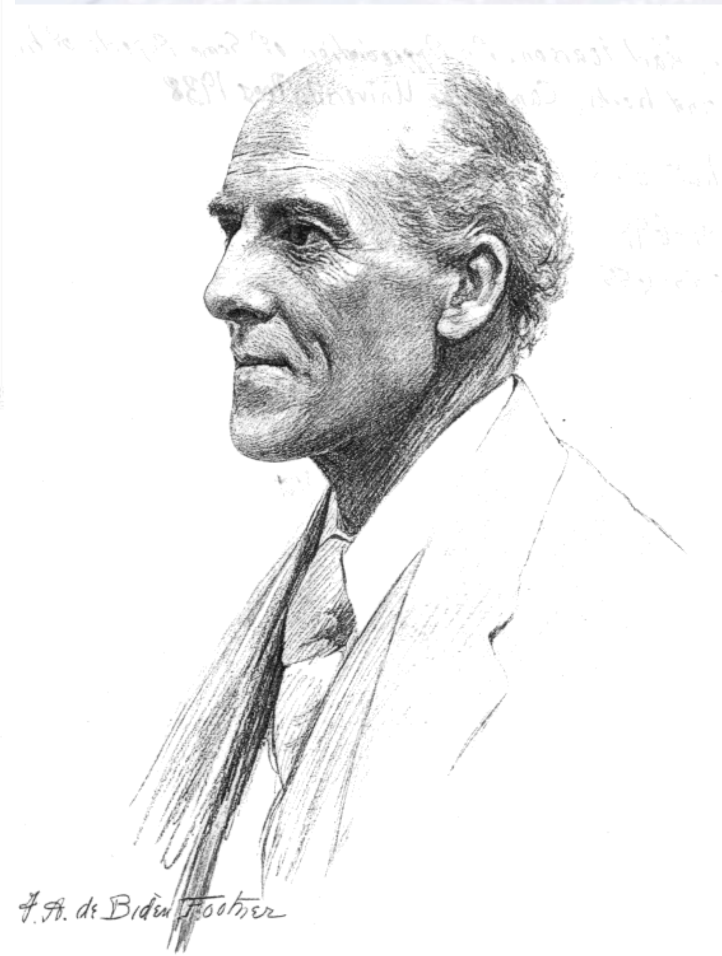
$$\text{Prob}(\chi^2{=}300.0,\ N_{dof}{=}200) = 0.000006$$

Always calculate and consider the probability!

Chi-Square for binned data

If the data is binned (i.e. put into a histogram), then Pearson's Chi-square applies:





The formula (based on Poisson statistics) is:

$$\chi^2 = \sum_{\mathrm{i} \in \mathrm{bin}} \frac{(O_i - E_i)^2}{E_i}$$

Chi-Square for binned data

While Pearson's Chi-square test is quite useful, it has some limitations, especially when some bins have low statistics.

The expected cell count (E_i) should not be too low. Some require 5 or more, and others require 10 or more. A common rule is 5 or more in 80% of bins, but no cells with zero expected count. When this assumption is not met, Yates's Correction can be applied.

One alternative is to divide by O_i when O_i is not 0 (ROOT/Minuit).

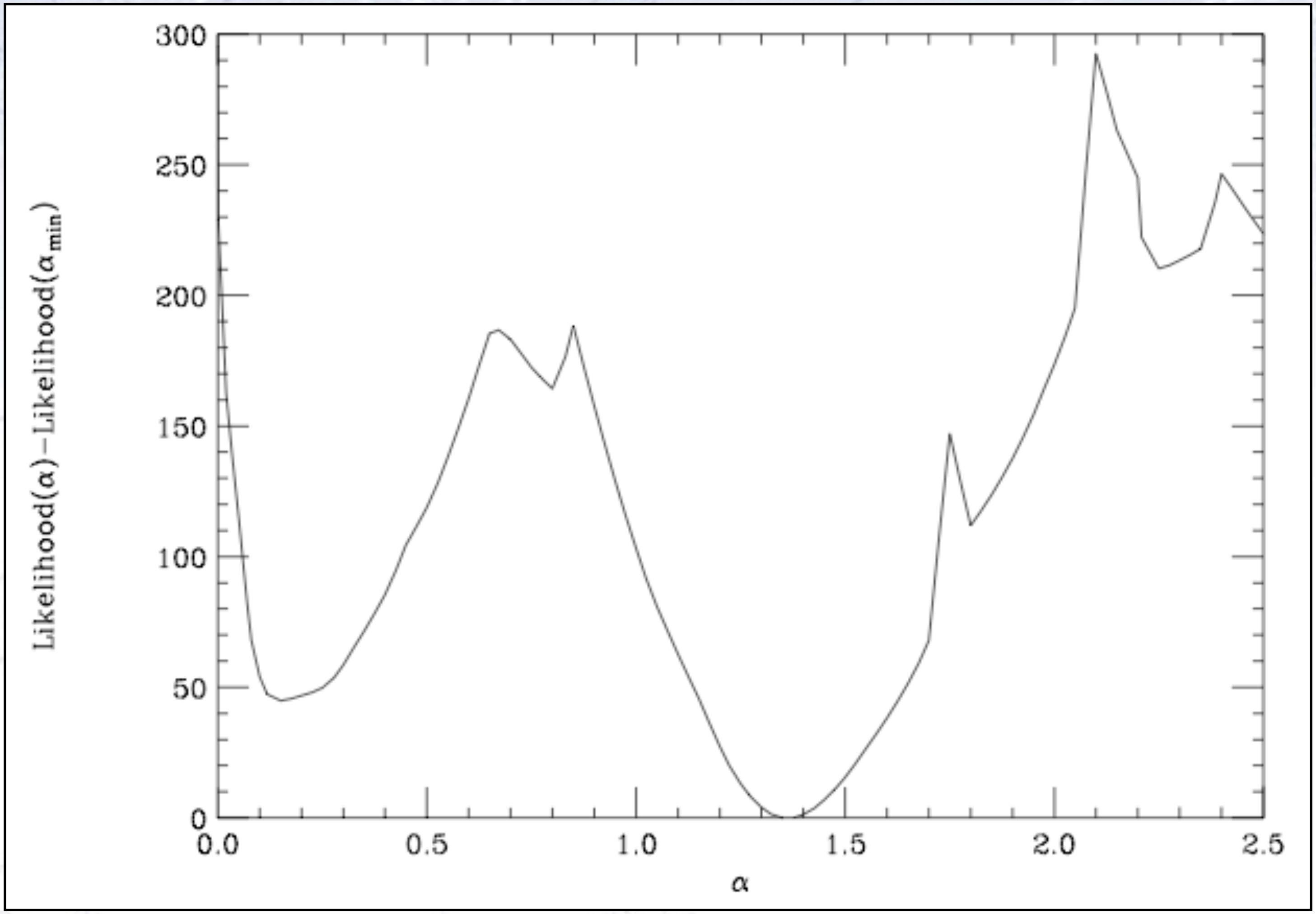
Another alternative is the likelihood fit, which does not suffer under low statistics.

$$\chi^2 = \sum_{\mathrm{i} \in \mathrm{bin}} \frac{(O_i - E_i)^2}{E_i}$$

Yet, another alternative is the G-test, which is more robust at low statistics. However, I've never seen it in use.

$$G = 2 \sum_{\mathrm{i} \in \mathrm{bin}} O_i \ \ln(O_i/E_i)$$

Example of Chi-Square



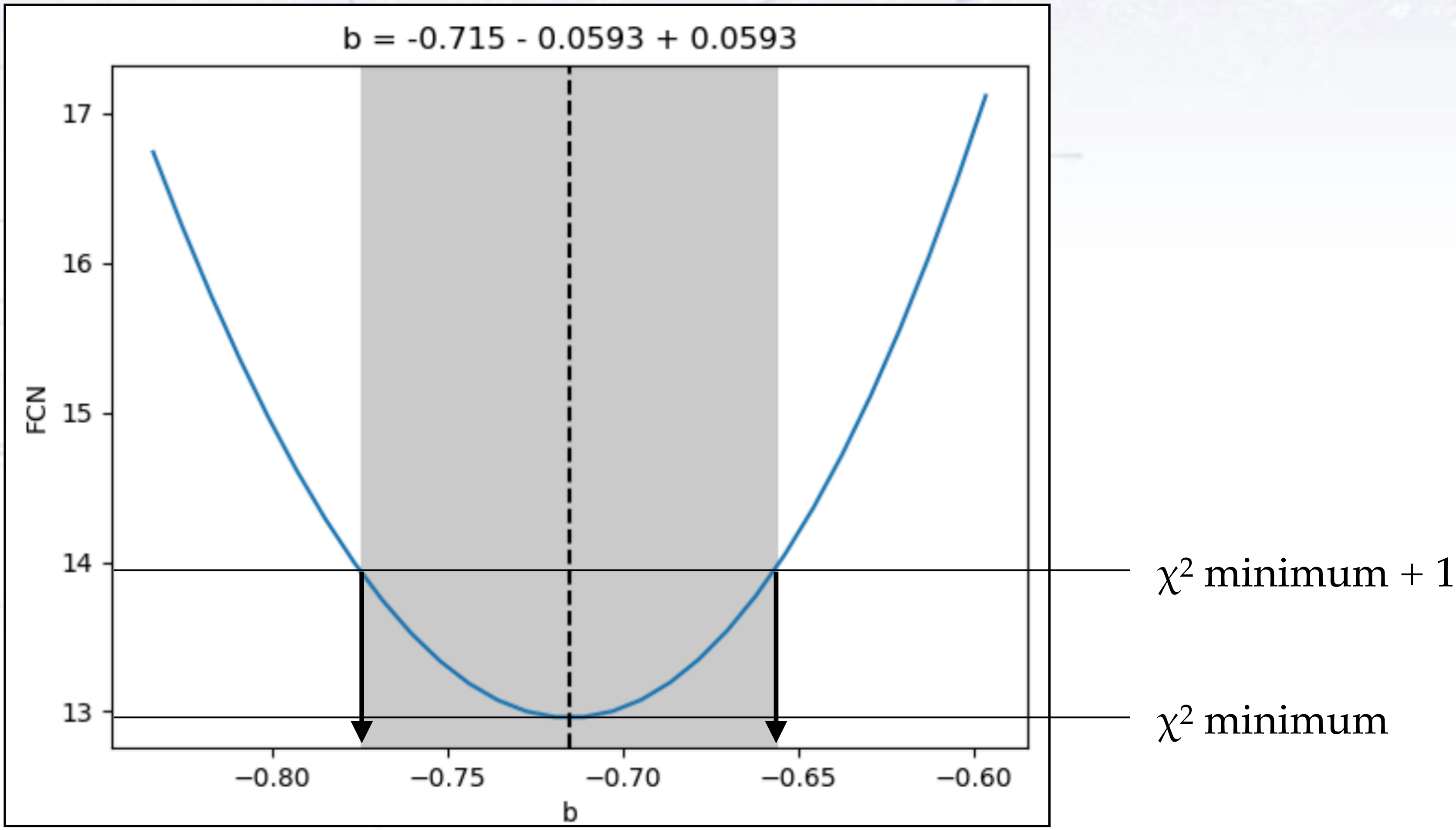
The fact that there are several minima makes fitting difficult/uncertain!

Always give good starting values!!!

Why the ChiSquare is (near) magic

Example of Chi-Square

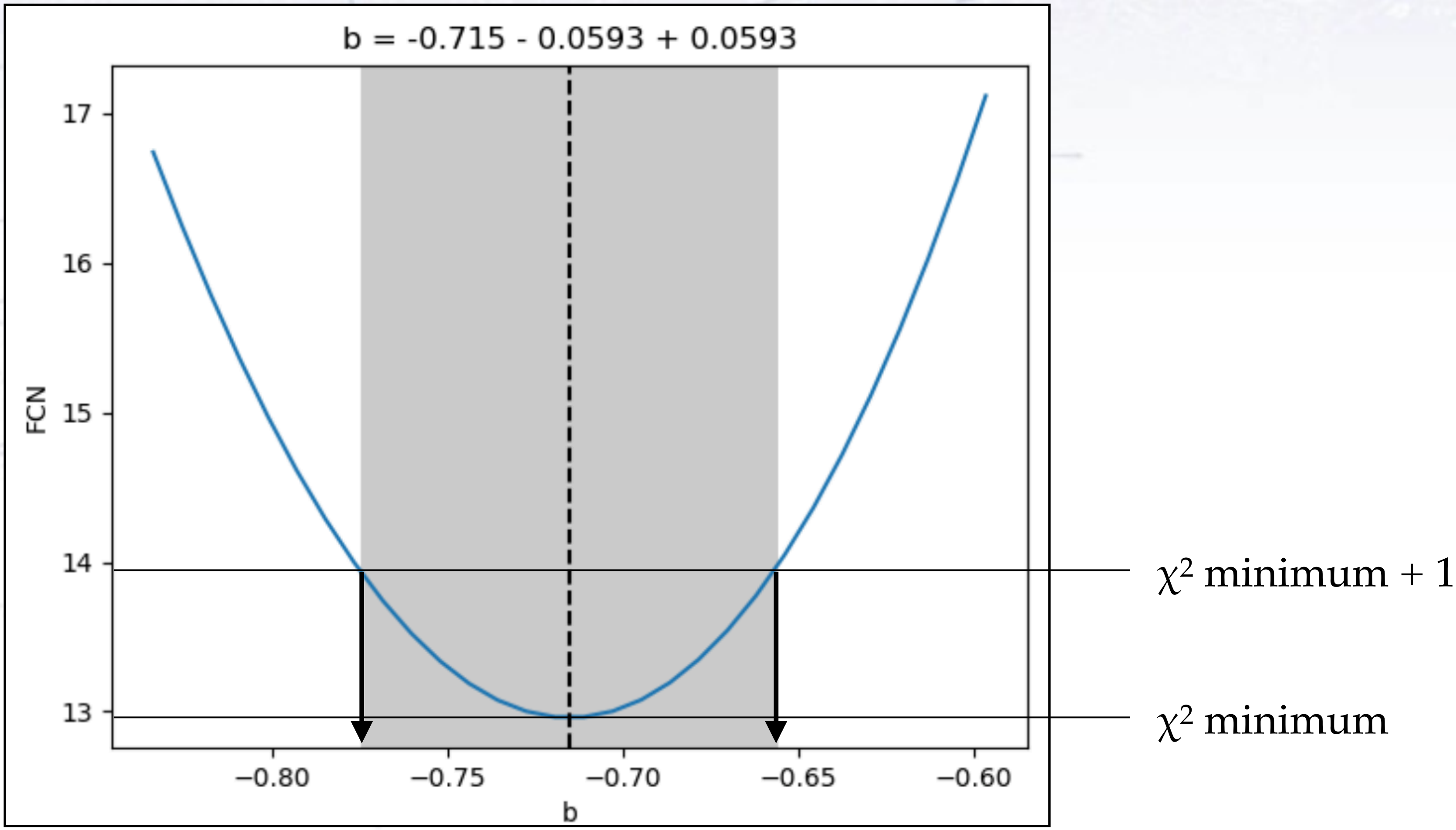
The uncertainty on a parameter is found where the Chi2 has increased by 1 from the minimum.



Example of Chi-Square

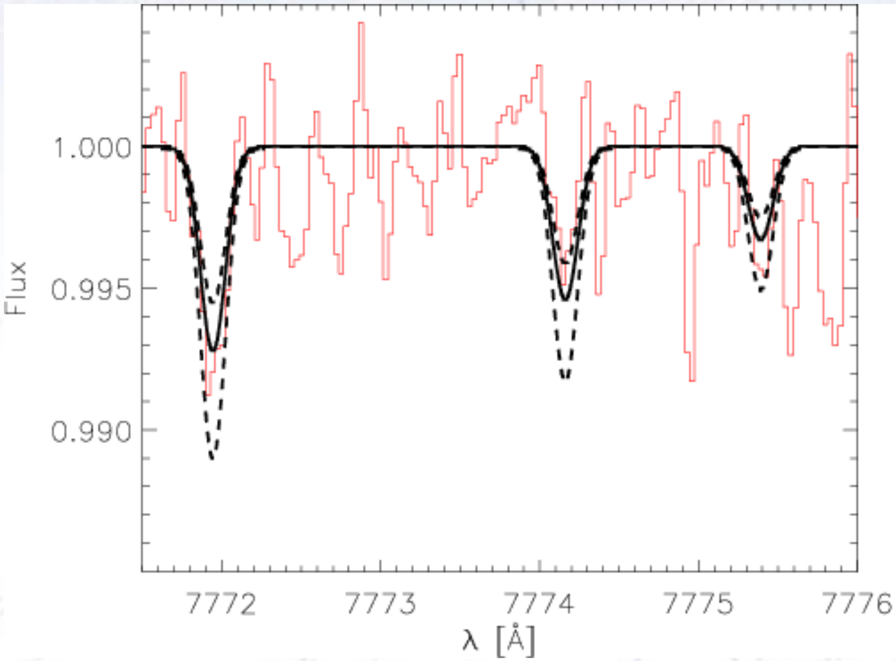
Please commit to memory!

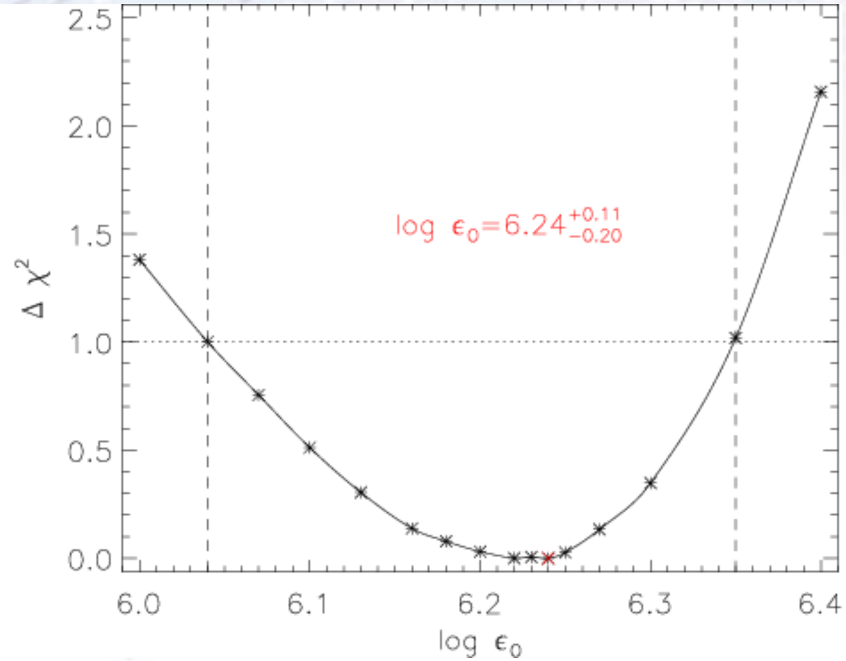
The uncertainty on a parameter is found where the Chi2 has increased by 1 from the minimum.



Example of Chi-Square

Uncertainties need not always be symmetric (though that is usually better!)





Asymmetric uncertainties are tricky to deal with (see later and/or Barlow).

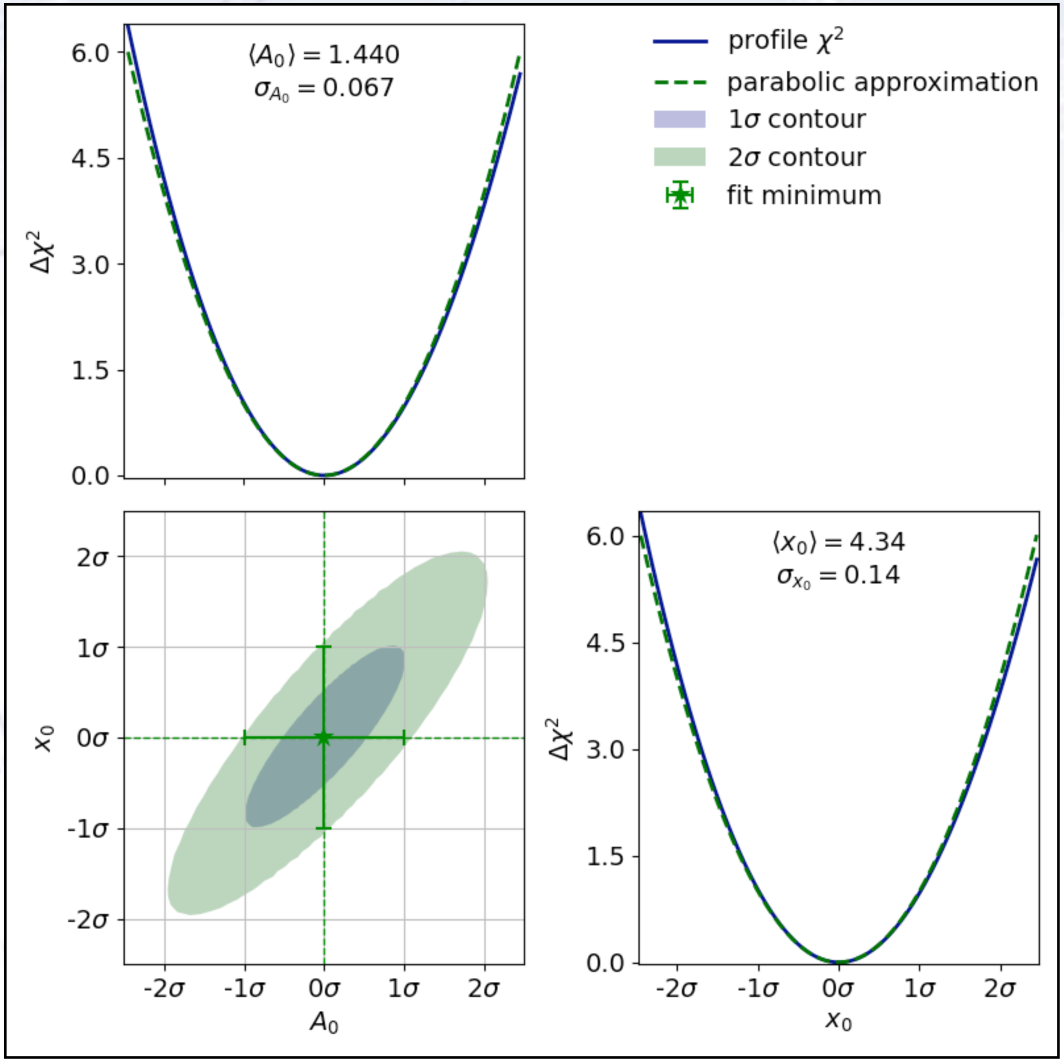
Example of Chi-Square

Fitting with multiple variables, one obtains a multi-dimensional parabola.

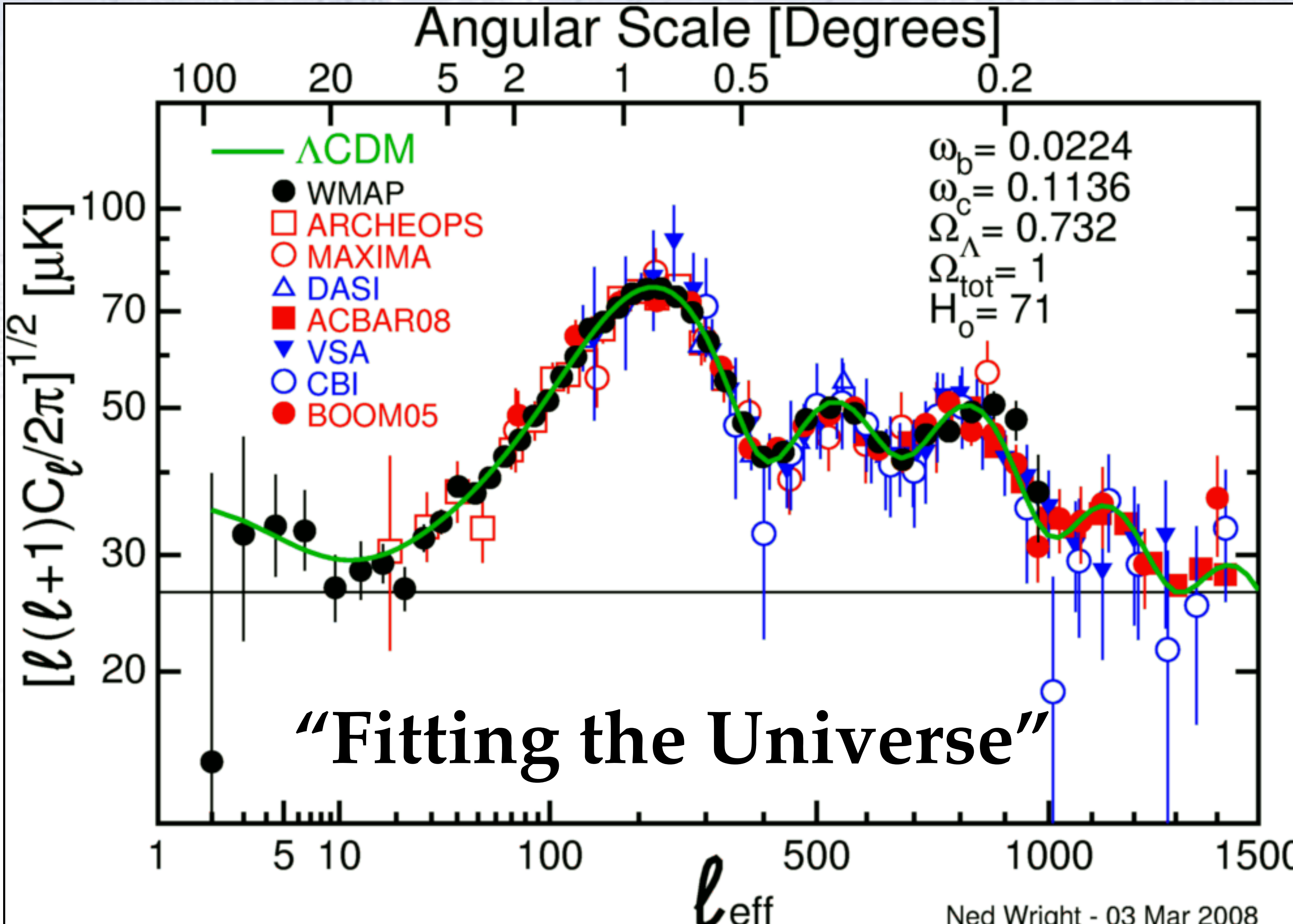
This is summarised in:
- Central fit values, $\hat{\boldsymbol{\mu}}$
- Covariance matrix, $\hat{\mathbf{V}}$

The diagonals of V are the variance (= σ^2) of the fit parameters.
The off-diagonals of V are the co-variances (and thus correlations) between fit parameters.

You should always look at these, as they reveal a lot about your fit (see later).



Example of Chi-Square



"Fitting the Universe"

Notes on the ChiSquare method

"It was formerly the custom, and is still so in works on the theory of observations, to derive the method of least squares from certain theoretical considerations, the assumed normality of the errors of the observations being one such.
It is however, more than doubtful whether the conditions for the theoretical validity of the method are realised in statistical practice, and the student would do well to regard the method as recommended chiefly by its comparative simplicity and by the fact that it has ***stood the test of experience****"*.

[G.U. Yule and M.G. Kendall 1958]

Calibration

Calibration definition

"Operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties (of the calibrated instrument or secondary standard) and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication."

[International Bureau of Weights and Measures]

Calibration definition

"Operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties (of the calibrated instrument or secondary standard) and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication.”

[International Bureau of Weights and Measures]

Personally, I would shorten this to:

“Operation that, under specified conditions:

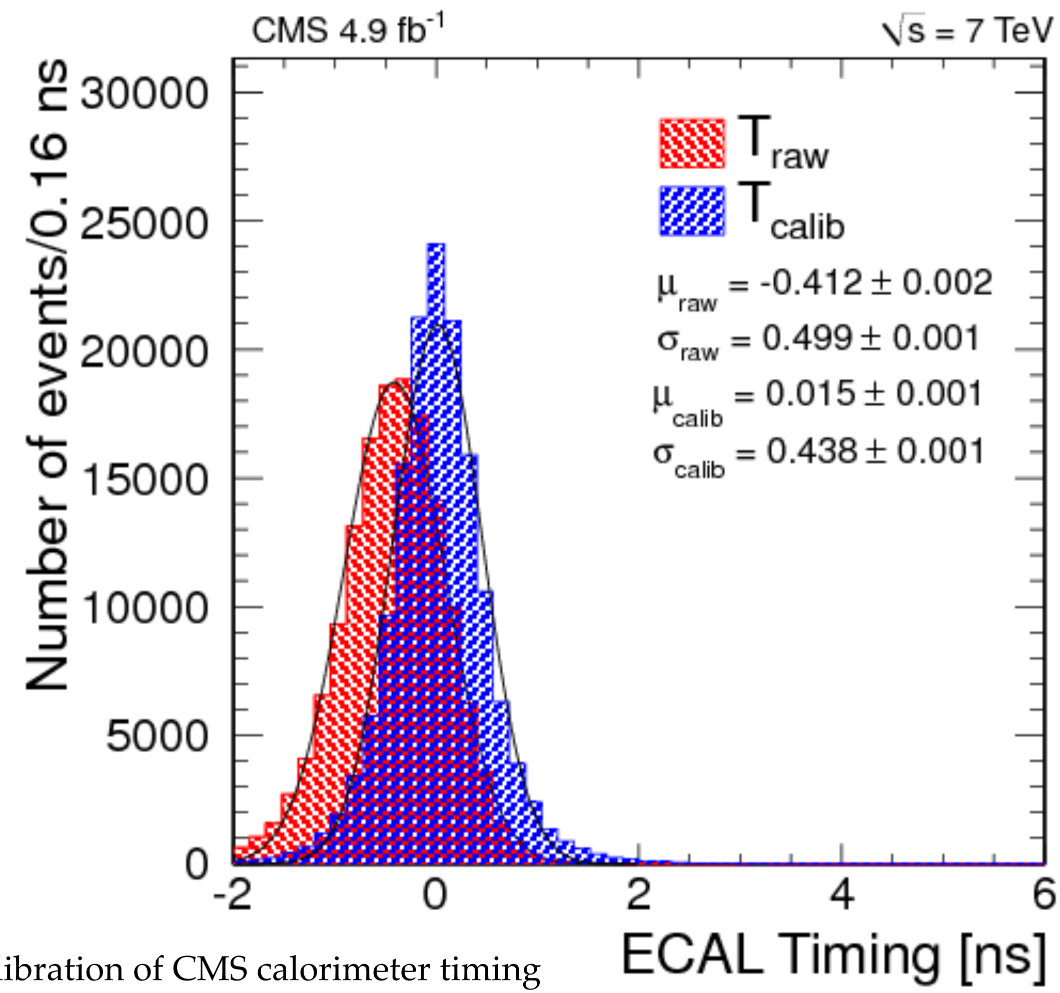
- Establishes a relation between the quantity of interest and associated information
- Uses this information to correct/improve the estimate of the quantity of interest.”

[Shortening of the above]

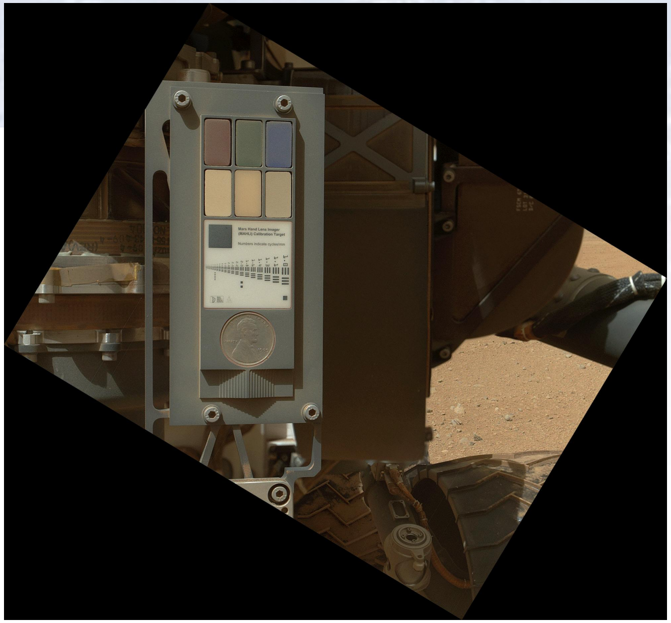
Let’s have a few examples…

Calibration is many things!

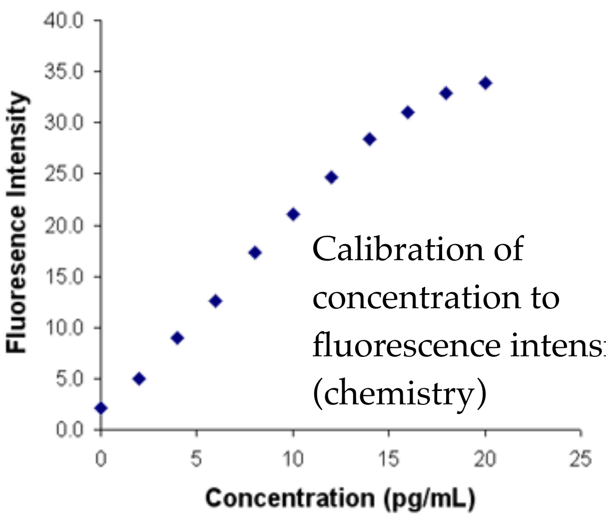
Every field of science involves calibration of some kind.



Calibration of CMS calorimeter timing

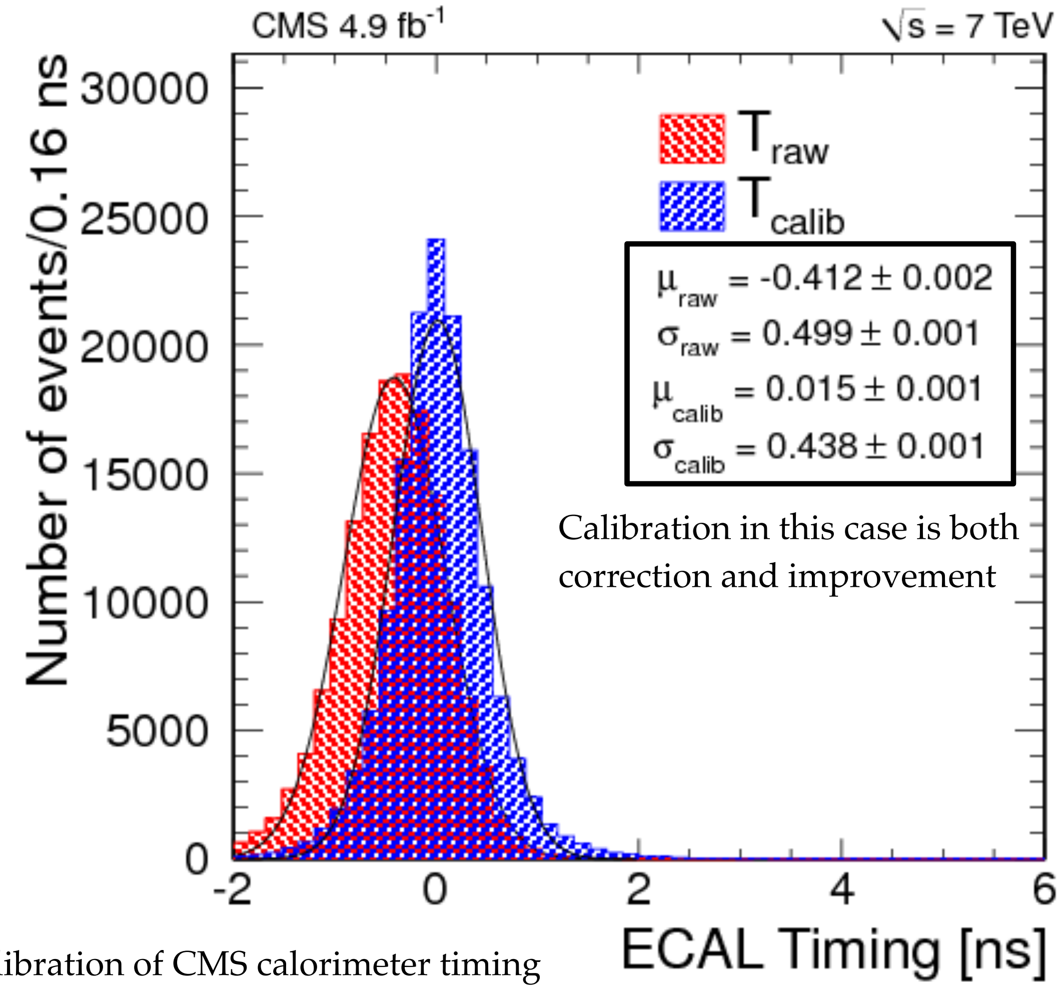


Calibration target of Mars rover "Curiosity"

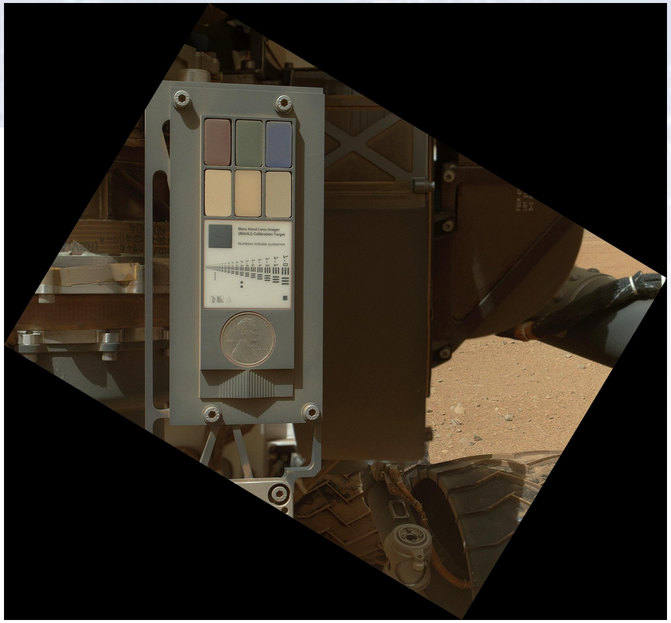


Calibration is many things!

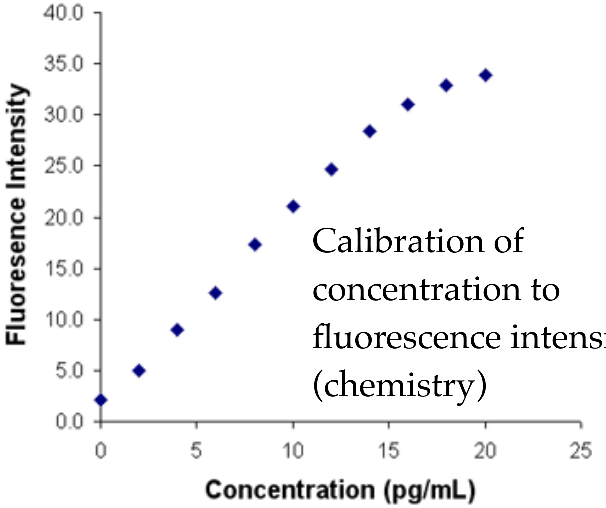
Every field of science involves calibration of some kind.



Calibration of CMS calorimeter timing



Calibration target of Mars rover "Curiosity"



General considerations

Though calibration spans widely, there are a few general considerations:

★ **Using control sample/group:**
- Purpose: To ensure that there is not some (inherent) bias.
- Aim: A good control sample is large and looks "exactly" like signal.
- Example: People without "signal" disease spanning same age/lifestyles.

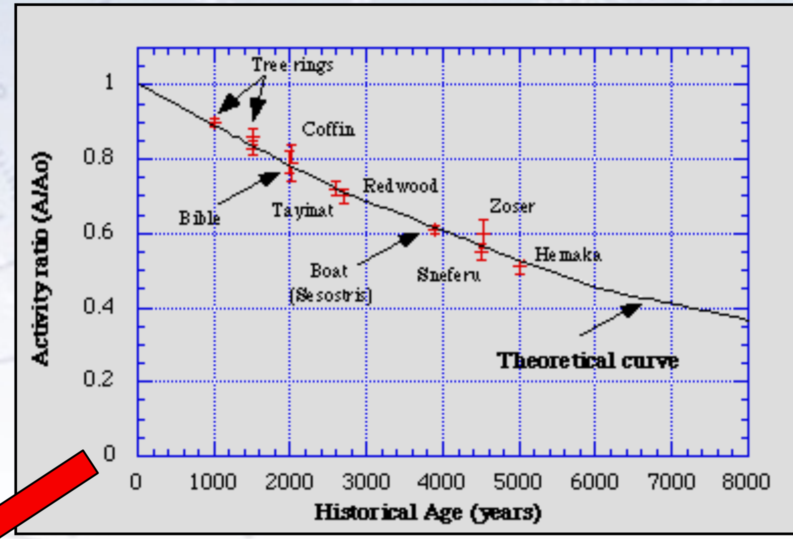
★ **Considering result for already well determined quantity:**
- Purpose: To ensure that there is not some (inherent) bias.
- Aim: A good control measurement is "easy" and well measured.
- Example: Unbiased momentum resolution using particle resonances (Z).

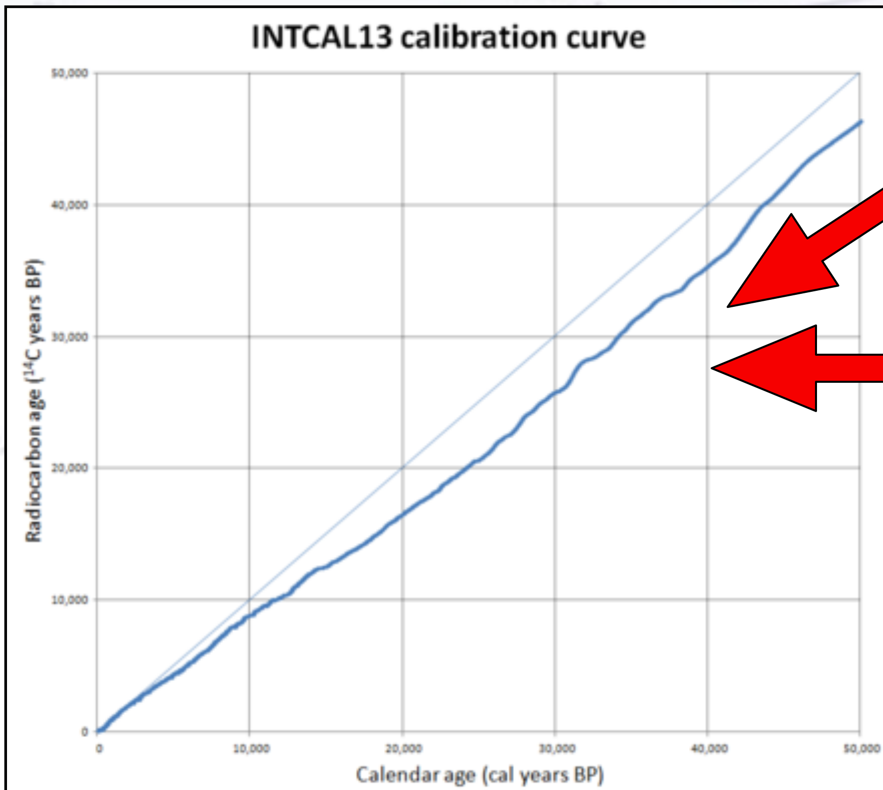
★ **Determining relation to well measurable quantity:**
- Purpose: Infer quantity in question from other sources/measurements.
- Aim: If one can't measure directly, perhaps it can be done indirectly.
- Example: Measuring flow of liquid in pipe using microphone (noise!).

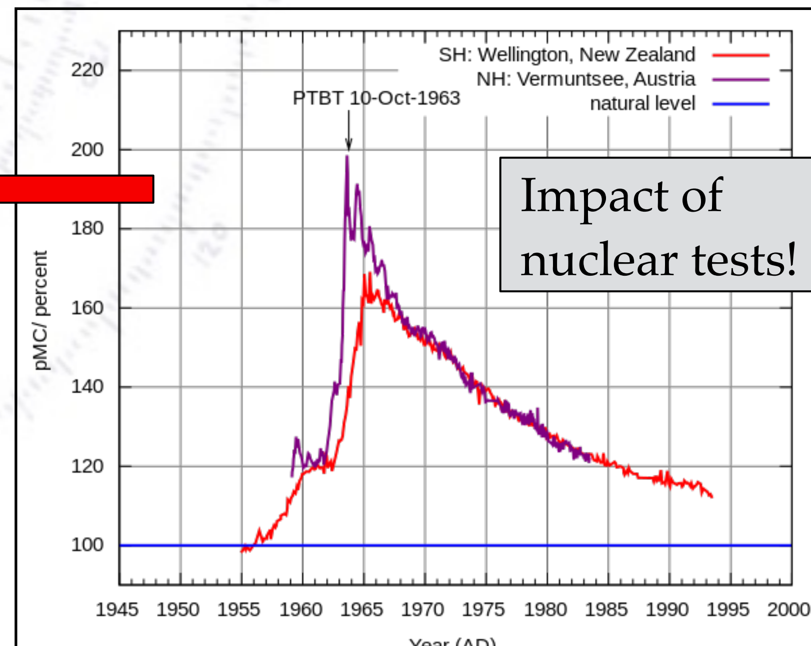
Each field of science have their own "tricks of the trade", and sometimes breakthroughs and Nobel Prizes are made through calibration (length scales in the Universe, search for the ether, accurate carbon 14 dating, etc.).

Example: Carbon 14 dating

Carbon 14 dating used (and uses) samples of known age (from historical sources) to calibrate the scale and uncertainties.
Tree rings have played a central role!

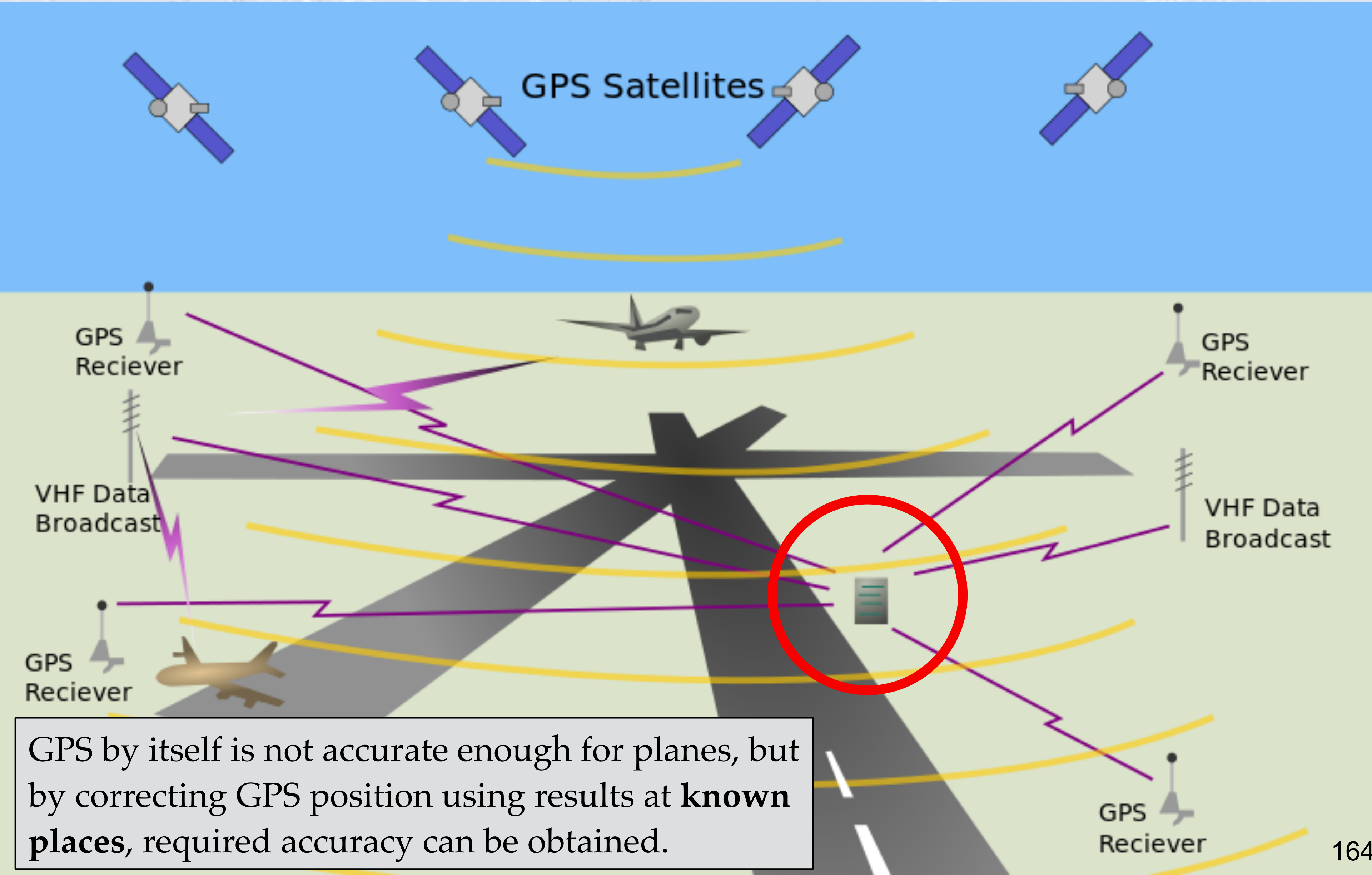






Impact of nuclear tests!

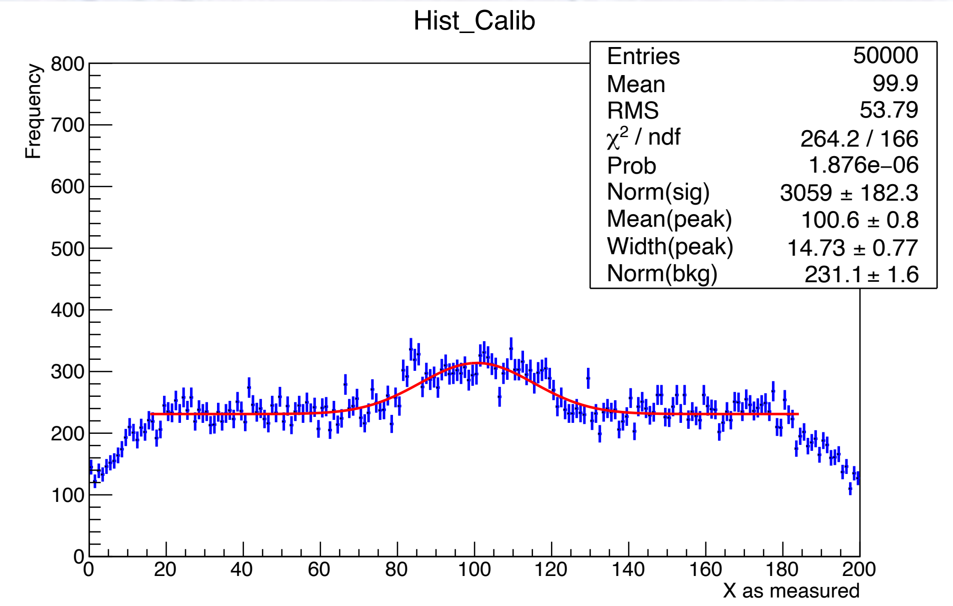
Example: Differential GPS



GPS by itself is not accurate enough for planes, but by correcting GPS position using results at **known places**, required accuracy can be obtained.

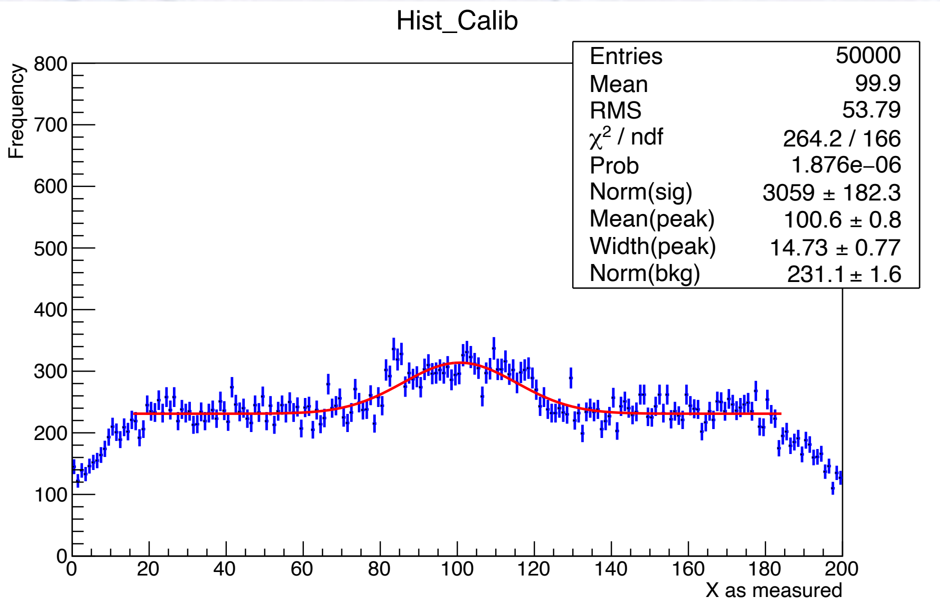
Example calibration

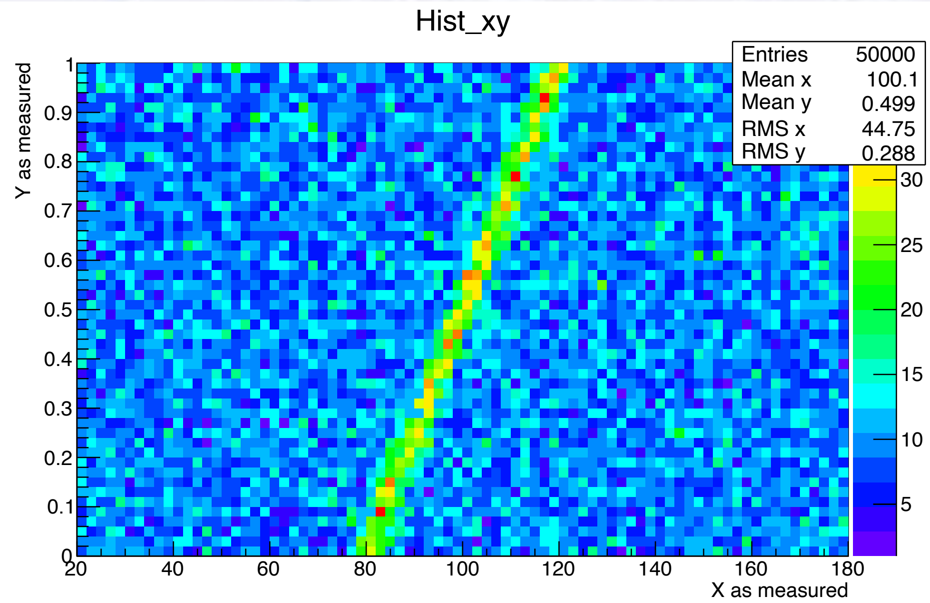
Imagine a variable, X, which has a peak in its spectrum, but which depends on another variable, Y. Variations in Y "smears out" the peak in X, and we would therefore like to calibrate for this.



Example calibration

Imagine a variable, X, which has a peak in its spectrum, but which depends on another variable, Y. Variations in Y "smears out" the peak in X, and we would therefore like to calibrate for this.



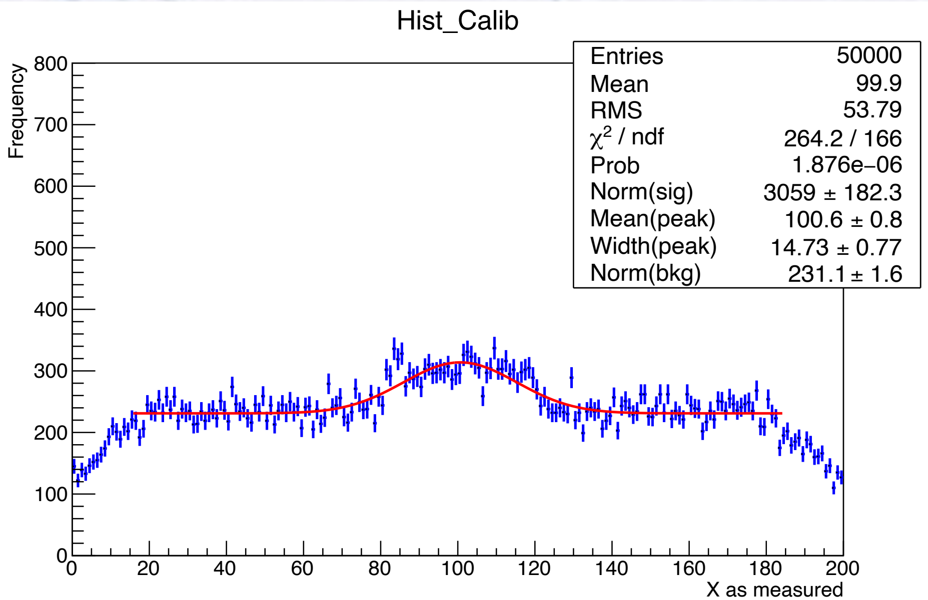


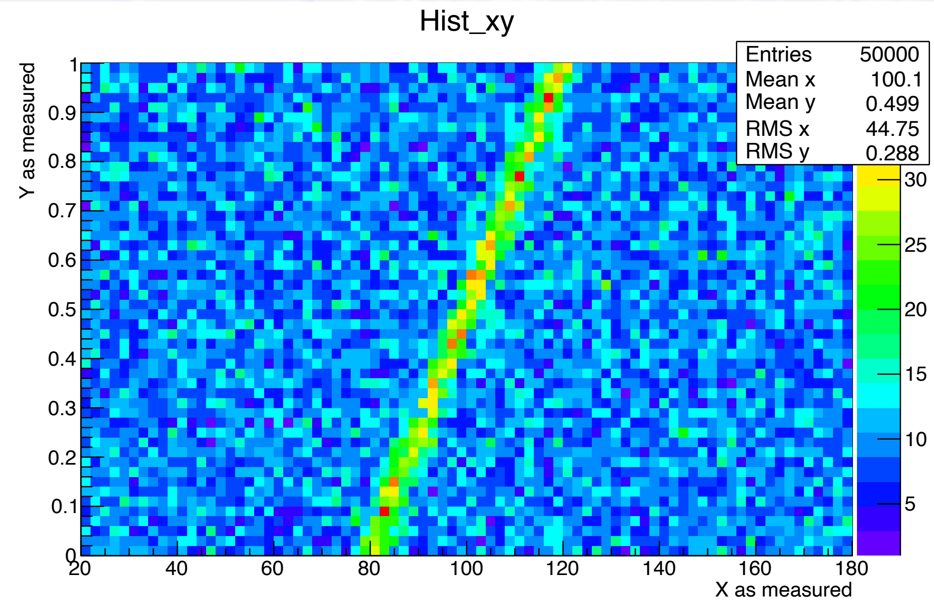
We therefore plot X as a function of Y, and notice a (in this case clear) correlation between Y and X. From this we can deduce how much the peak is shifted as a function of Y, and hence correct for it.

$$X_{calib} = X_{meas} + ???$$

Example calibration

Imagine a variable, X, which has a peak in its spectrum, but which depends on another variable, Y. Variations in Y "smears out" the peak in X, and we would therefore like to calibrate for this.



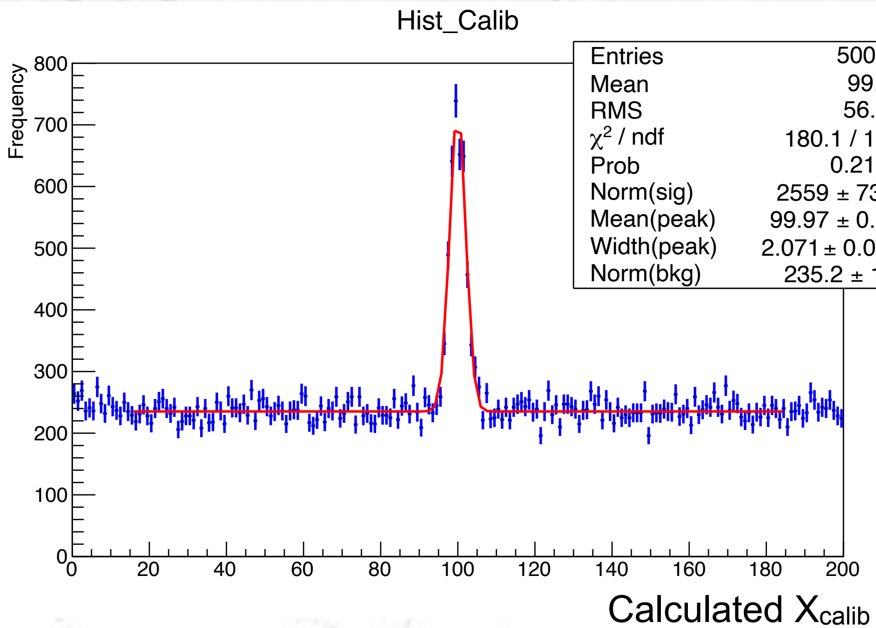


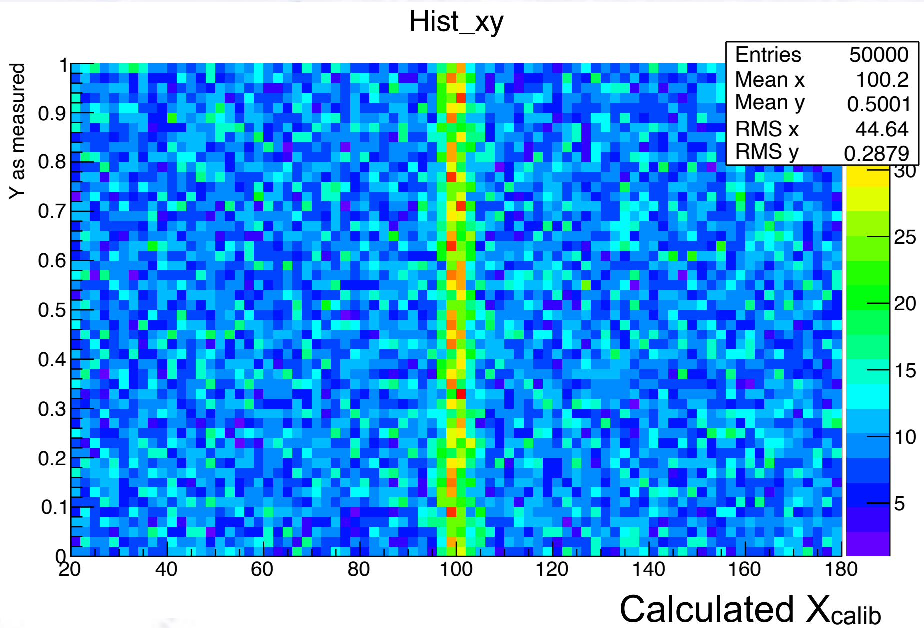
We therefore plot X as a function of Y, and notice a (in this case clear) correlation between Y and X. From this we can deduce how much the peak is shifted as a function of Y, and hence correct for it. A simple inspection yields:

$$\mathbf{X_{calib} = X_{meas} - 40(Y - 0.5)}$$

Example calibration

Applying this yields a new and (much) improved resolution of the peak in X, as would also be expected. At the same time, we can check, that now there is no dependence of the calibrated value of X on Y.





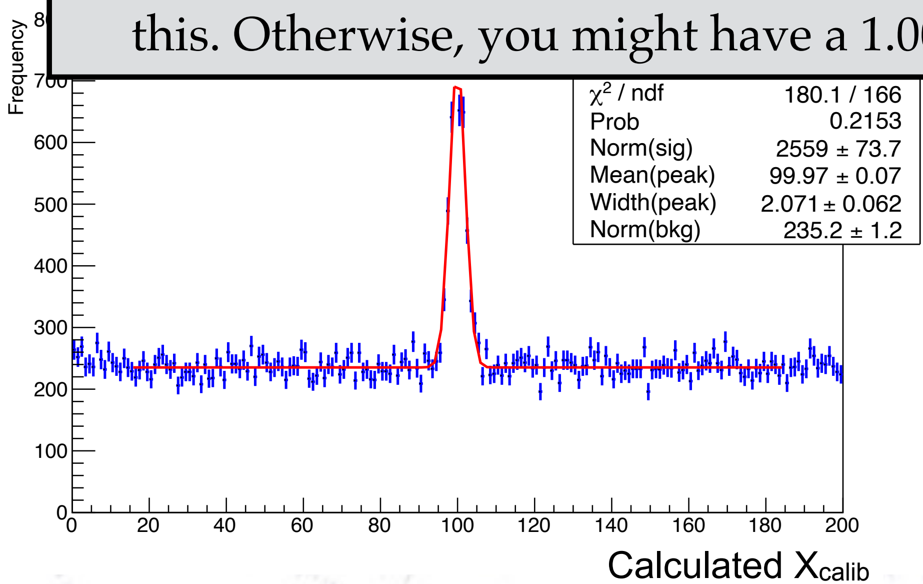
We thus conclude, that the calibration worked, and (of course) describe our calibration in the paper we publish. Note that sometimes, one needs a "control sample" for which the correct value is known through other sources.

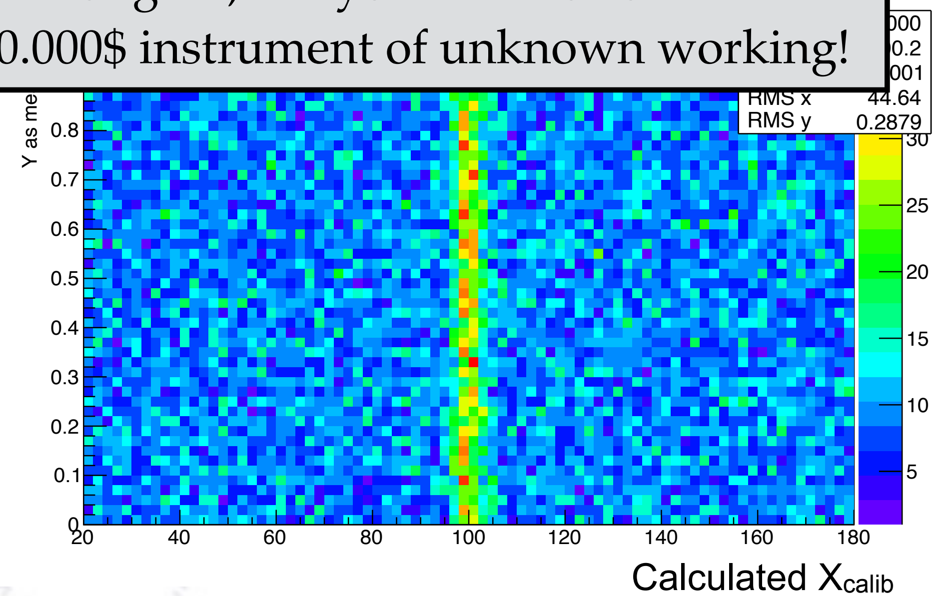
$$X_{calib} = X_{meas} - 40(Y - 0.5)$$

Example calibration

Q: How can we "obtain" a line at say X=100 to be used for calibration?

A: This you have to think AHEAD of time, i.e. when planning the experiment. It might be as simple as sticking a radioactive source down, or shining light on the instrument, or sending particles through it, but you have to consider this. Otherwise, you might have a 1.000.000$ instrument of unknown working!





We thus conclude, that the calibration worked, and (of course) describe our calibration in the paper we publish. Note that sometimes, one needs a "control sample" for which the correct value is known through other sources.

$$X_{calib} = X_{meas} - 40(Y - 0.5)$$

Simpson’s Paradox

(Really: Simpson’s “apparent” Paradox)
(if time allows)

Case: Berkeley admission

In 1973, University of California, Berkeley, were considering which of their applicants got admitted.
As can be seen below, there is seemingly a **bias against women**, as a smaller fraction of women are admitted.
Is that really the case, or is there more to the data than first glance reveals?

Case: Berkeley admission

In 1973, University of California, Berkeley, were considering which of their applicants got admitted.
As can be seen below, there is seemingly a **bias against women**, as a smaller fraction of women are admitted.
Is that really the case, or is there more to the data than first glance reveals?

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

https://homepage.stat.uiowa.edu/~mbognar/1030/Bickel-Berkeley.pdf

Case: Berkeley admission

In 1973, University of California, Berkeley, were considering which of their applicants got admitted.
As can be seen below, there is seemingly a **bias against women**, as a smaller fraction of women are admitted.
Is that really the case, or is there more to the data than first glance reveals?

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

https://homepage.stat.uiowa.edu/~mbognar/1030/Bickel-Berkeley.pdf

Table 1. Decisions on applications to Graduate Division for fall 1973, by sex of applicant—naive aggregation. Expected frequencies are calculated from the marginal totals of the observed frequencies under the assumptions (1 and 2) given in the text. $N = 12{,}763$, $\chi^2 = 110.8$, d.f. $= 1$, $P = 0$ (*18*).

Applicants	Outcome: Observed Admit	Outcome: Observed Deny	Outcome: Expected Admit	Outcome: Expected Deny	Difference Admit	Difference Deny
Men	3738	4704	3460.7	4981.3	277.3	– 277.3
Women	1494	2827	1771.3	2549.7	– 277.3	277.3

Case: Berkeley admission

In 1973, University of California, Berkeley, were considering which of their applicants got admitted.
As can be seen below, there is seemingly a **bias against women**, as a smaller fraction of women are admitted.
Is that really the case, or is there more to the data than first glance reveals?

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

https://homepage.stat.uiowa.edu/~mbognar/1030/Bickel-Berkeley.pdf

Table 1. Decisions on applications to Graduate Division for fall 1973, by sex of applicant—naive aggregation. Expected frequencies are calculated from the marginal totals of the observed frequencies under the assumptions (1 and 2) given in the text. $N = 12{,}763$, $\chi^2 = 110.8$, d.f. $= 1$, $P = 0$ (*18*).

Applicants	Outcome: Observed Admit	Outcome: Observed Deny	Outcome: Expected Admit	Outcome: Expected Deny	Difference Admit	Difference Deny
Men	3738	4704			277.3	− 277.3
Women	1494	2827			− 277.3	277.3

3737 / (3738+4704) = **44.3%**
1494 / (1494+2827) = **34.6%**

Case: Berkeley admission

In 1973, University of California, Berkeley, were considering which of their applicants got admitted.
As can be seen below, there is seemingly a **bias against women**, as a smaller fraction of women ar
Is that rea
data than

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

hel, J. W. O'Connell

30/Bickel-Berkeley.pdf

As already noted, we are aware of the pitfalls ahead in this naive approach, but we intend to stumble into every one of them for didactic reasons.

Table 1. D pplicant—naive aggregation. Expected frequencies are calculated from the marginal totals of the observed frequencies under the assumptions (1 and 2) given in the text. $N = 12{,}763$, $\chi^2 = 110.8$, d.f. = 1, $P = 0$ (*18*).

Applicants	Outcome: Observed Admit	Observed Deny	Expected Admit	Expected Deny	Difference Admit	Difference Deny
Men	3738	4704			277.3	− 277.3
Women	1494	2827			− 277.3	277.3

3737 / (3738+4704) = **44.3%**
1494 / (1494+2827) = **34.6%**

Case: Berkeley admission

Bickel et al. goes on to analyse the data further with several interesting findings:

sex. Our computations, therefore, except where otherwise noted, will be based on the remaining 85. For a start let us identify those of the 85 with bias sufficiently large to occur by chance less than five times in a hundred. There prove to be four such departments. The deficit in the number of women admitted to these four (under the assumptions for calculating expected frequencies as given above) is 26. Looking further, we find six departments biased in the opposite direction, at the same probability levels; these account for a deficit of 64 men.

Out of 85 departments with relevant data, a few seem to show a bias… in both directions, and mostly agains men!!! What!

This seems counter intuitive to what we found to begin with. Where did the bias of 277 women less than expected go?

Case: Berkeley admission

Bickel et al. goes on to analyse the data further with several interesting findings:

sex. Our computations, therefore, except where otherwise noted, will be based on the remaining 85. For a start let us identify those of the 85 with bias sufficiently large to occur by chance less than five times in a hundred. There prove to be four such departments. The deficit in the number of women admitted to these four (under the assumptions for calculating expected frequencies as given above) is 26. Looking further, we find six departments biased in the opposite direction, at the same probability levels; these account for a deficit of 64 men.

Out of 85 departments with relevant data*, a few seem to show a bias… in both directions, and mostly agains men!!! What!

This seems counter intuitive to what we found to begin with. Where did the bias of 277 women less than expected go?

*Here you should ALWAYS ask, what this involves!
In this case, 16 departments either had no women applying, or did not deny any students admission.

Case: Berkeley admission

In order to illustrate the point, Bickel et al. gives a hypothetical (and fun!) case:

Table 2. Admissions data by sex of applicant for two hypothetical departments. For total, $\chi^2 = 5.71$, d.f. $= 1$, $P = 0.19$ (one-tailed).

Applicants	Outcome: Observed Admit	Outcome: Observed Deny	Outcome: Expected Admit	Outcome: Expected Deny	Difference Admit	Difference Deny
			Department of machismatics			
Men	200	200	200	200	0	0
Women	100	100	100	100	0	0
			Department of social warfare			
Men	50	100	50	100	0	0
Women	150	300	150	300	0	0
			Totals			
Men	250	300	229.2	320.8	20.8	− 20.8
Women	250	400	270.8	379.2	− 20.8	20.8

The two (very hypothetical) departments are clearly very fair regarding gender, but still a difference appears between the overall resulting observation and expectation.

Case: Berkeley admission

The “apparent conclusion” (Berkeley discriminates against applications from women) is a result of Simpson’s Paradox (my text):

“Effect for group, which disappears or reverses, when considering subgroups”.

It is effects such as this, which makes statistics difficult, yet at the same time **very important**.

different degree. *The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into.* Moreover this phenomenon is more pronounced in departments with large numbers of applicants. Figure 1

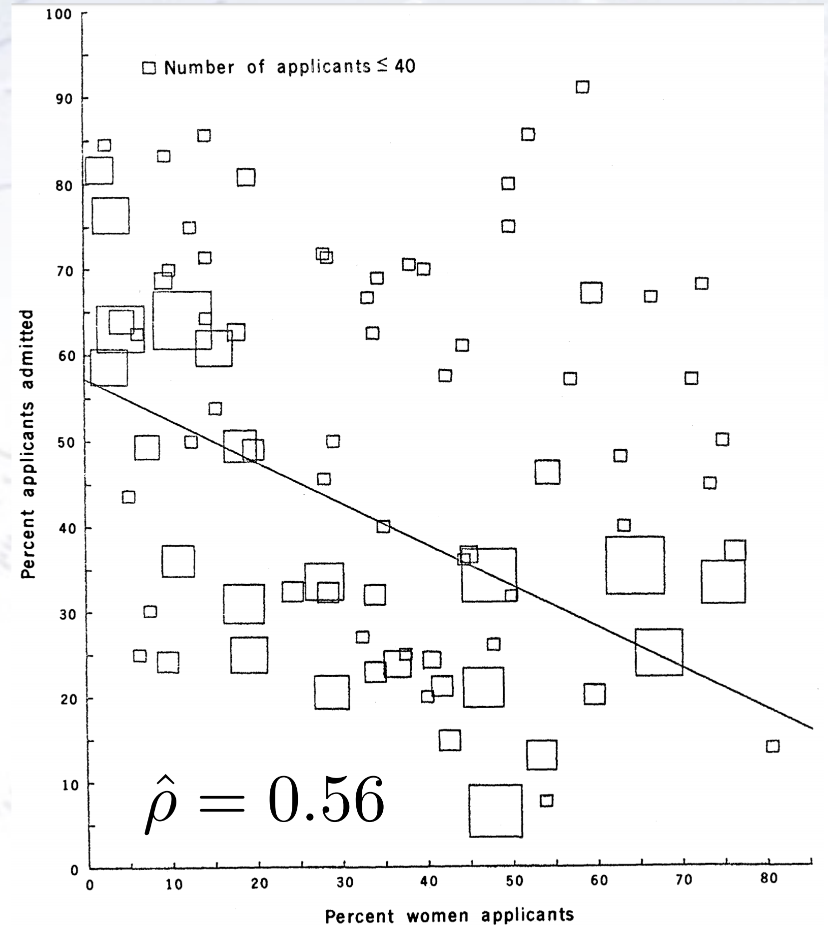


Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

Case: Berkeley admission

The "apparent conclusion" (Berkeley discriminates against applications from women) is a result of Simpson's Paradox (my text):

"Effect for group, which disappears or reverses, when considering subgroups".

It is effects such as this, which makes statistics difficult, yet at the same time **very important**.

different degree. *The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into.* Moreover this phenomenon is more pronounced in departments with large numbers of applicants. Figure 1

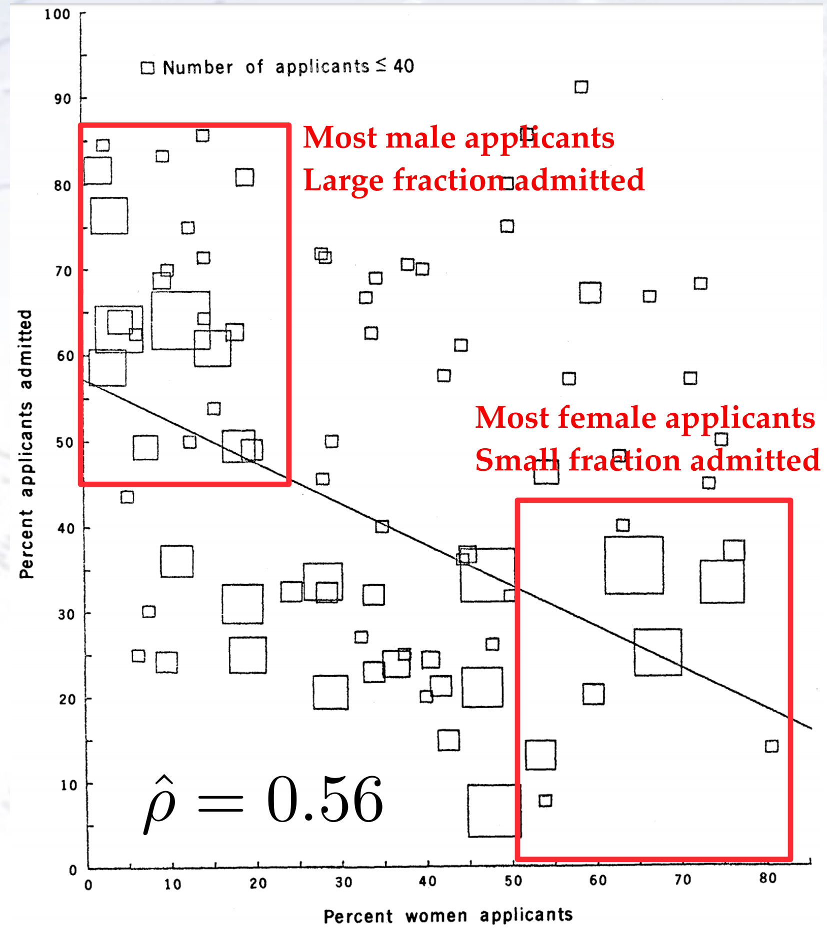


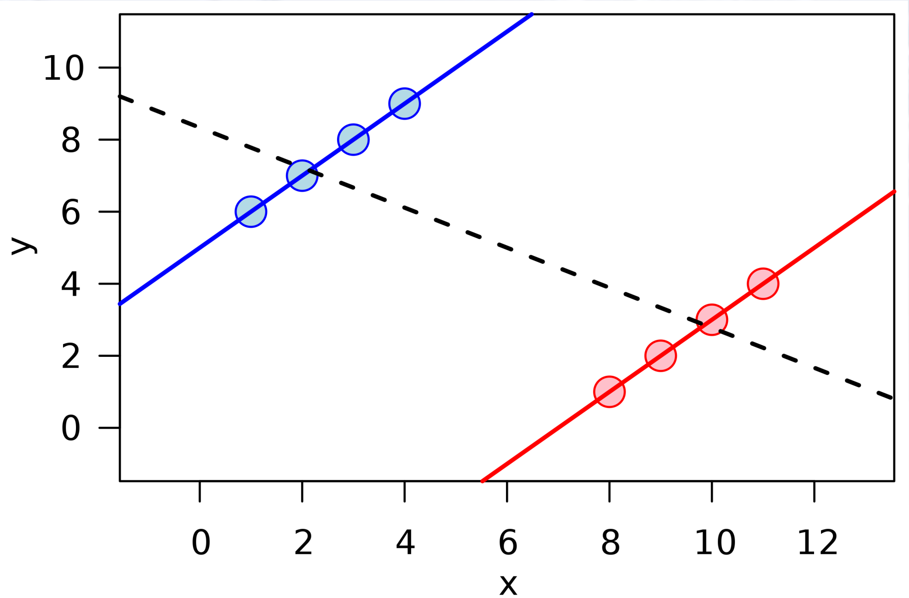
Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

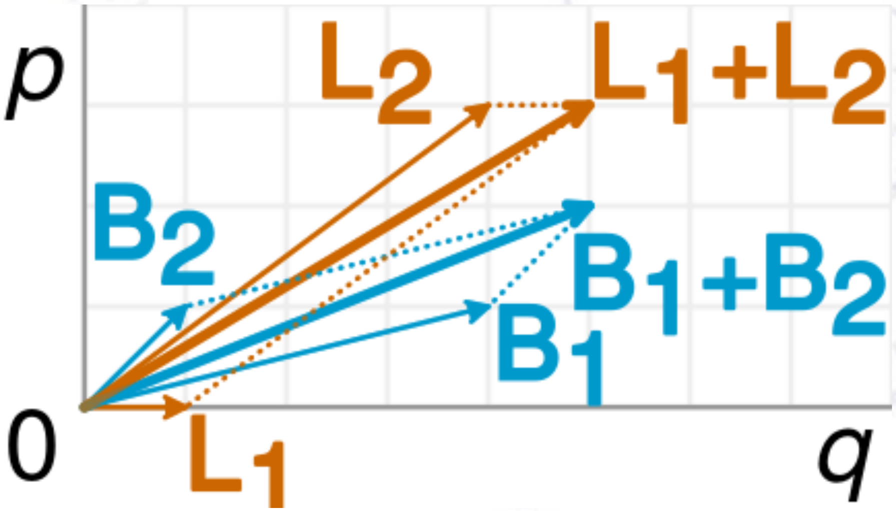
Simpson's Paradox explained

The reason for the **apparent** paradox arise when frequency data is unduly given causal interpretations.

The figure on the right illustrates the "paradox" nicely.

The situation can be illustrated with 2D vectors, as shown below.





A succes rate p/q (successes / attempts) can be represented by vectors with a slope. Higher slope = higher succes rate.

But though B1 is steeper than L1, and B2 is steeper than L2, then B1+B1 is not as steep as L1+L2.

Summary

Summary

1. The Central Limit Theorem is you (new?) friend, as it explains why you should expect Gaussian uncertainties.

2. Estimators are given formulae that you should know in order to obtain (unbiased and efficient) estimates from data.

3. PDFs are in some sense our "model building blocks". Most originate from given processes (that you should know), and should be used accordingly.

4. The ChiSquare is THE way to perform fits, if uncertainties are Gaussian, as it provides a crucial goodness-of-fit measure.

5. Calibration is central part of experimental physics, and requires foresight, insight, and experimental planning.

6. Always consider different types/classes separately, as this augments efficiency, and saves you from Simpson's (apparent) paradox.