UWS UNIVERSITY OF THE WEST of SCOTLAND

Introductory Data Analysis

Lecture 1 - Introduction

SUPA Graduate School 2010
Prof. Andrei Andreyev
Andrei.Andreyev@uws.ac.uk

Nuclear Physics Research Group
School of Engineering and Science

My field: Low-energy Nuclear Physics

- **Measurements of nuclear reactions & decays**
- **Detection of very rare signals (1 per day/week)**
- **Synthesis of new elements and isotopes**
- **Events of purely statistical nature**
- **Measurement of, mostly, three parameters**
  - **Cross sections** (statistical, probability of interaction)
  - **Spectroscopic information** (precise measurement of decay energy –α,γ , or continuous spectra – β decay)
  - **Quantum numbers** (discrete assignment from continuous data)

You

- What year of PhD are you?
- What kind of field are you working in?
- Do you have data yet?
- Did you have at least ‘some’ course on statistics/data analysis in your University studies?

Please, select **one** student from each University to collate all these data and send them to me by e-mail (Andrei.Andreyev@uws.ac.uk)

Course scope

Introduction to dealing with *quantitative* data analysis

Statistical methods/techniques for manipulating data

Bridge to **Advanced Data Analysis SUPA Course**

Course outline (some of topics)

- Introduction to quantitative data analysis
- Probability distributions
- Sampling from distributions
- Uncertainties – statistical & systematic
- Propagation of uncertainties
- Fitting data – extraction of parameters
- Comparison of different measurements
- Comparison to theoretical calculations

Recommended reading

- **R.J. Barlow,** *Statistics – A guide to the use of statistical methods in the physical sciences,* John Wiley and Sons, ISBN 0471922951
- **L. Kirkup**, *Experimental Methods – an Introduction to the Analysis and Presentation of Data*, John Wiley and Sons, ISBN 0471335797
- **L. Kirkup, R. B. Frenkel**, *An Introduction to Uncertainty in Measurement, using the GUM (Guide to the Expression of Uncertainty in Measurement)* Cambridge University Press, 2006, ISBN-13: 9780521605793, ISBN-10: 0521605792
- **T. Greenfield, A. Metcalfe**, *Design And Analyse Your Experiment With Minitab*, Hodder Arnold, 2007, ISBN 9780340807804
- **L. Gonick, W. Smith**, *The Cartoon Guide to Statistics*, Collins, 2000, ISBN 0062731025
- **N. C. Barford**, *Experimental Measurements: Precision, Error and Truth*, John Wiley and Sons, Second Edition, 1985
- **J. R. Taylor**, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, 2nd Edition, University Science Books, 1997

Data Analysis is needed, e.g.:

- To represent/visualise/communicate large quantities of information (e.g., as histograms, graphs, charts ...)

- To extract meaningful information from a set of measurements, e.g.:
  - Mean value, standard deviation
  - Correlations
  - Signal/background
  - Degree of confidence

- To make meaningful comparison between different measurements, and/or different theoretical models

Data Analysis in Research/Physics

- Models are important for physics
- Data imperative to test the validity of models
- Necessary to know to what extent different measurements can reliably test models, and discriminate between different ones

Data Analysis in Research

- You are undertaking research
- You may be the only/first person who has measured a particular quantity
- Measurement could validate/refute a theoretical prediction/model (or discriminate between different models)
- Results could indicate additional/unexpected effects which need a new/extended explanation
- New techniques require validation against old measurements
- Need to communicate your results to other physicists, your supervisor and colleagues

Data sets

Set of N *independent* measurements (e.g. decay energy of some process)

$x_1, x_2, x_3, x_4, \ldots\ldots x_i, \ldots\ldots, x_N$

Some basic properties of those measurements:

Sum $\sum_{i=1}^{N} x_i$

Experimental Mean: $\bar{x}_{exp} = \frac{\sum}{N} = \frac{1}{N}\sum_{i=1}^{n} x_i$

Discrete vs Continuous Data

- The **numerical data** that we will use in this course falls into 1 of 2 categories : **discrete** and **continuous sets**

- A type of data is **discrete** if
  a) there is only **a finite number** of values possible,
AND/OR
  b) there is **a space on the number line** between each 2 possible values.

Example 1 – Discrete data of type a) A 5 question exam is given in a Math class. **The number of correct exam's answers is an example of discrete data.** The number of correct answers will be one of the following : 0, 1, 2, 3, 4, or 5. **There are not an infinite number of values, therefore this data is discrete.** Also, if we were to draw a number line and place each possible value on it, we would see a space between each pair of values.

Discrete Data (continued)

b) There is **a space on the number line** between each two possible values

•**Example 2.** In order to get a driver license, a person must pass a written exam. **How many times it would take this person to pass this exam is also an example of discrete data**. A person could take it once, or twice, or 3 times, or 4 times, or… . So, the possible values are 1, 2, 3, … . **Thus, there are infinitely many possible values** (or the person becomes bankrupt), but if we were to put them on a number line, we would see a space between each pair of values

•**Discrete data usually occurs in a case where there are only a certain number of values, or when we are counting something (using WHOLE numbers).**

Continuous Data

- **Continuous** data makes up the rest of numerical data.
- This is a type of data that is usually associated with **some sort of physical measurement**

Example 3. The height of a tree is an example of continuous data. Is it possible for a tree to be 1 m tall? Sure! How about 1.1 m? Yes! How about 1.11 m? Sure! **The possibilities depends upon the accuracy of our measuring device.**

- One general way to tell if data is continuous is to ask if it is possible for the data to take on **values that are fractions or decimals.** If the answer is yes, this is usually continuous data.

Example 4. The length of time it takes for a light bulb to burn out is an example of continuous data. Could it take 800 hours? How about 800.7? 800.7354? The answer to all 3 is yes.

Frequency Distribution

Set of N *independent* measurements (e.g. velocity of a car, or exam's marks…) $x_1, x_2, x_3, x_4, .. x_i, \ldots\ldots, x_N$

Frequency distribution F(x):

$$F(x) = \frac{\text{number of occurrences of } x}{\text{number of measurements (N)}}$$

•A **frequency distribution** is a tool for organizing data. We use it to:

1) group data into categories, and
2) show the number of observations in each category.

Simple for discrete data, but most data is continuous (eg measuring a voltage, position etc)

Frequency Distribution

Example 5. Here are some test scores from a math class (40 marks in total). (Q: what data set is it?)

65	91	85	76	85	87	79	93
82	75	100	70	88	78	83	59
87	69	89	54	74	89	83	80
94	67	77	92	82	70	94	84
96	98	46	70	90	96	88	72

It's hard to get a feel for this data in this format because it is unorganized!

Frequency Distribution

65	91	85	76	85	87	79	93
82	75	100	70	88	78	83	59
87	69	89	54	74	89	83	80
94	67	77	92	82	70	94	84
96	98	46	70	90	96	88	72

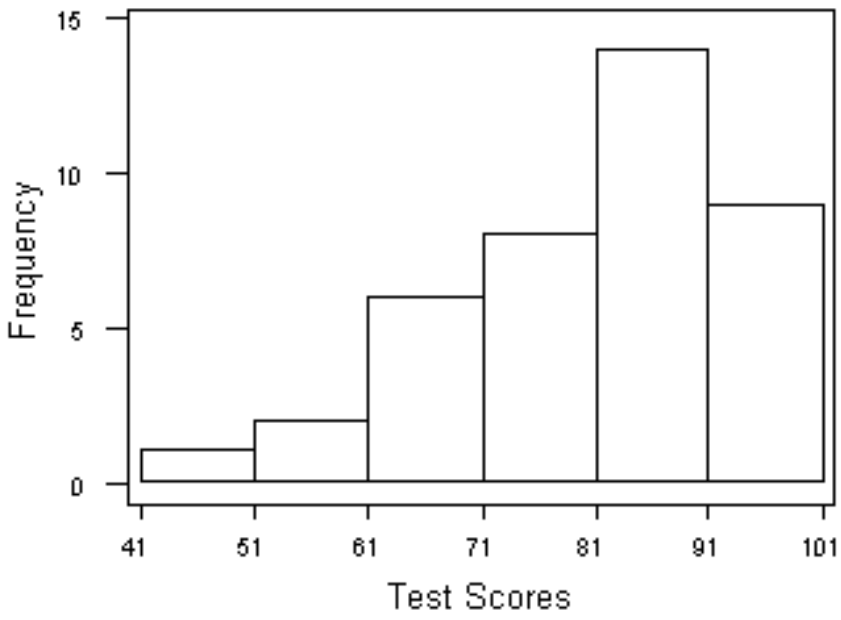
To construct a frequency distribution, one first identifies the **lowest** and **highest values** in the list. The lowest value here is 46, and the highest - 100. A possible set of categories that would work here is 41-50, 51-60, 61-70, 71-80, 81-90, and 91-100.

Mark's range	Frequency
41-50	1
51-60	2
61-70	6
71-80	8
81-90	14
91-100	9

Frequency Distribution and Histograms

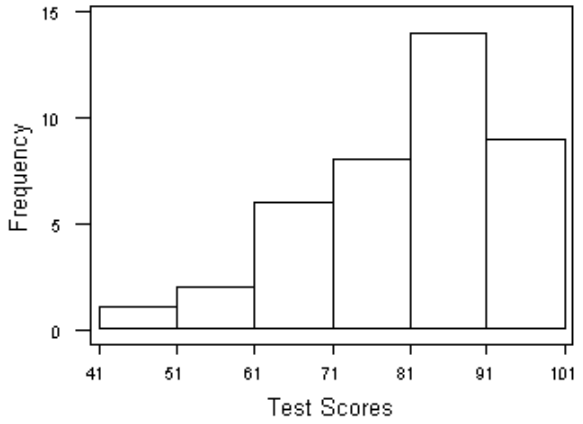
Mark	Frequency
41-50	1
51-60	2
61-70	6
71-80	8
81-90	14
91-100	9

A **histogram** could be thought of **as a graph of a frequency distribution.**



Frequency Distribution

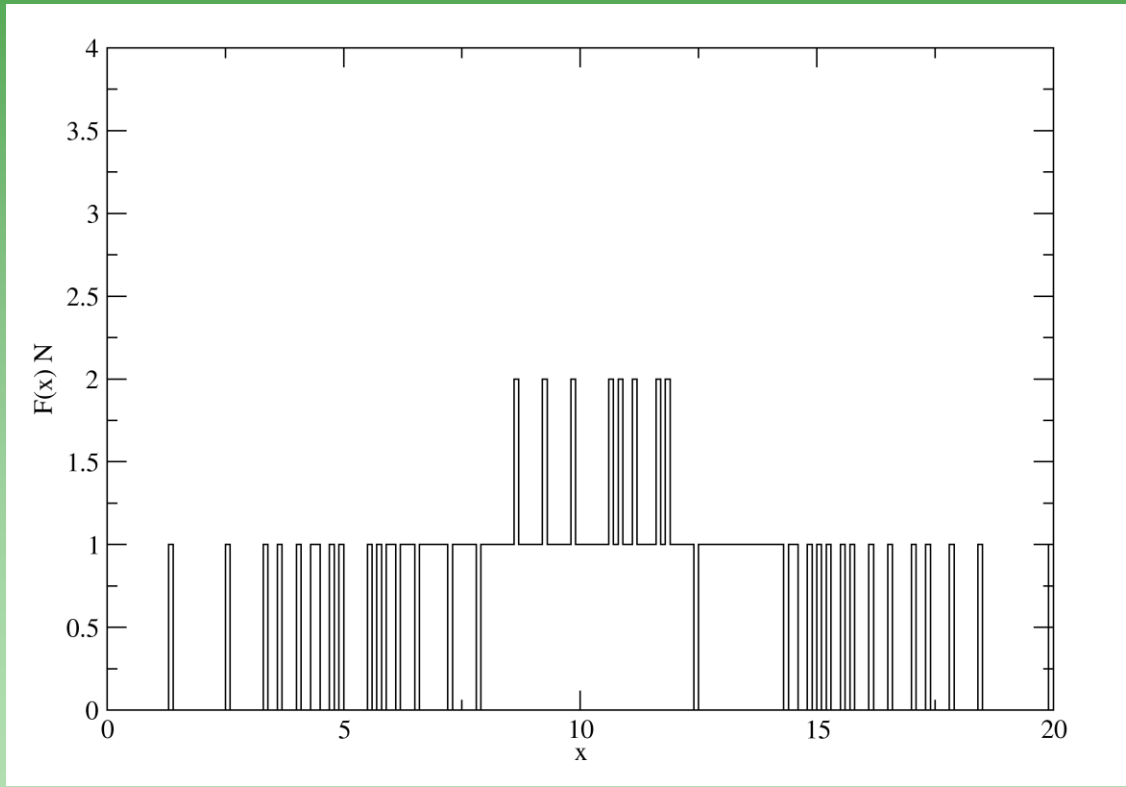
Marks	Frequency	Relative Frequency	Percent Frequency
41-50	1	1/40	2.5%
51-60	2	2/40	5%
61-70	6	6/40	15%
71-80	8	8/40	20%
81-90	14	14/40	35%
91-100	9	9/40	22.5%



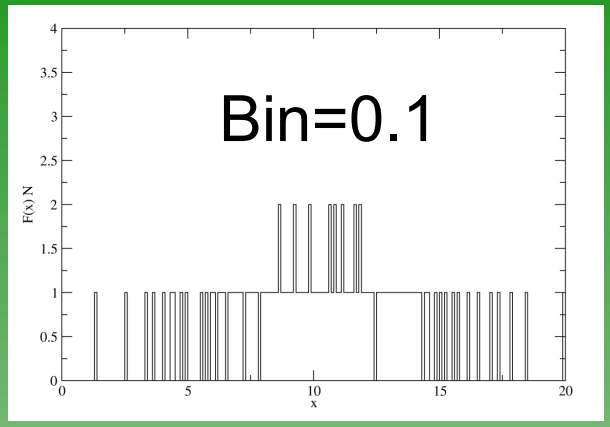
Discretisation or Binning Data

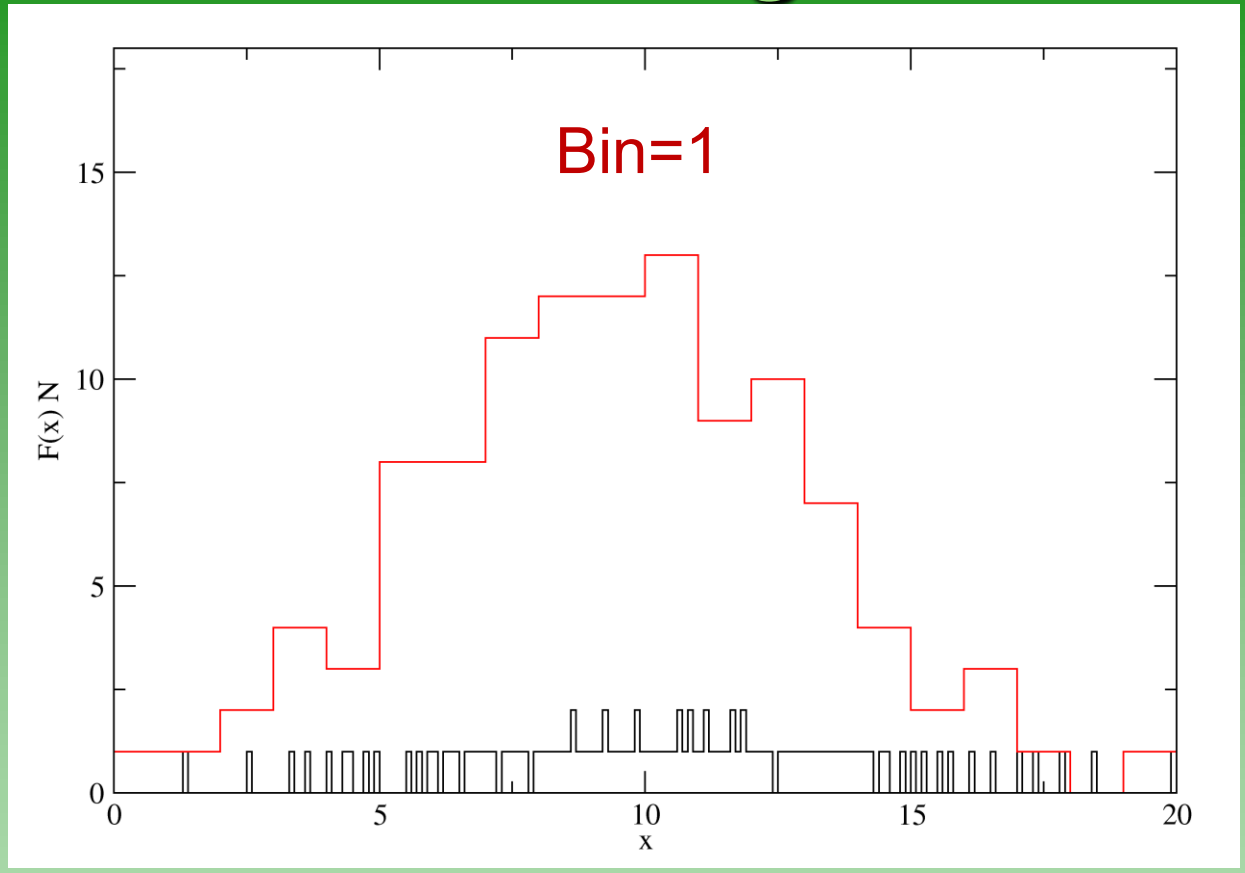
Discretise (or 'bin') data means: to divide x into bins, and count the number of events that fall into each bin)

- Try 'fine' binning first
- The size of the 'bin' here is very 'fine': 0.1 (50 measured values between 10 and 15)
- Max. value: 2



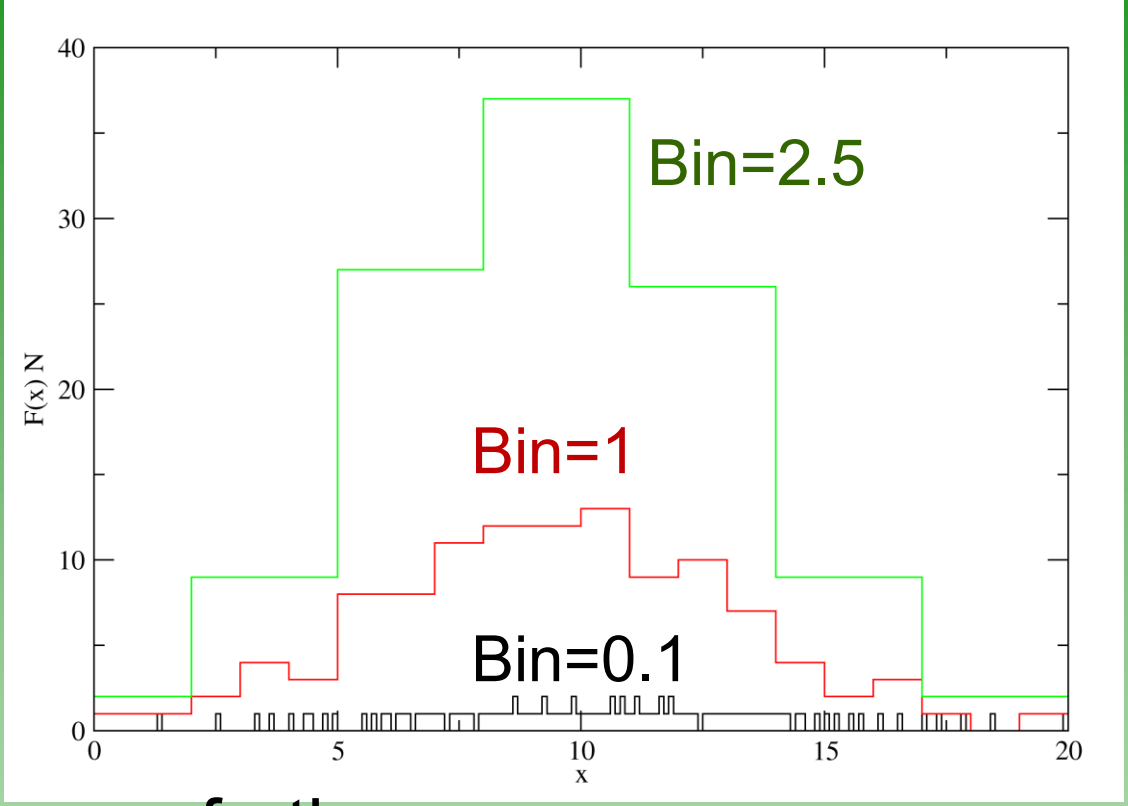
Discretisation or Binning Data





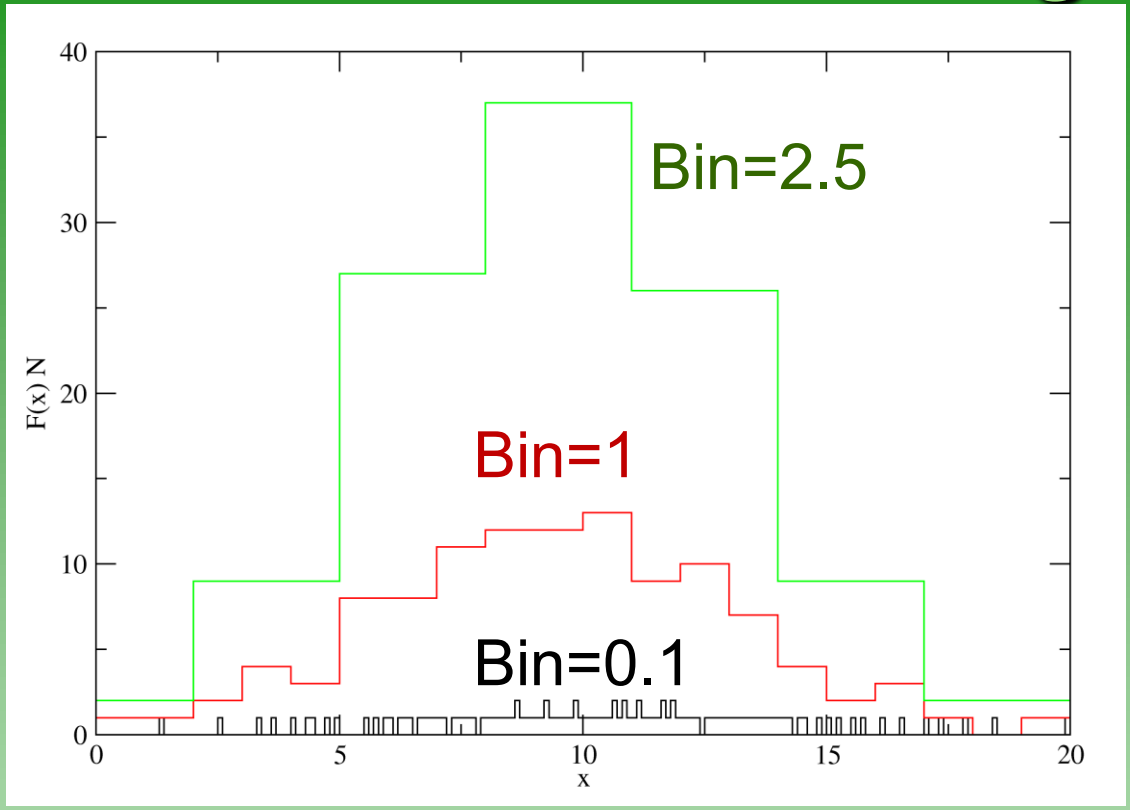
- ‘Re-bin’:
- Make 5 values between 10 and 15: bin is equal to 1
- ‘Max’ value is 12 (looks better.. at least to me)
- Of course, number of data is the same!

Discretisation or Binning Data



- ‘Re-bin’ even further:
- Make 2 values between 10 and 15: bin is equal to 2.5
- ‘Max’ value is 35 – good or bad?
- But, lost ‘structure - everything is symmetric now!

Discretisation or Binning Data

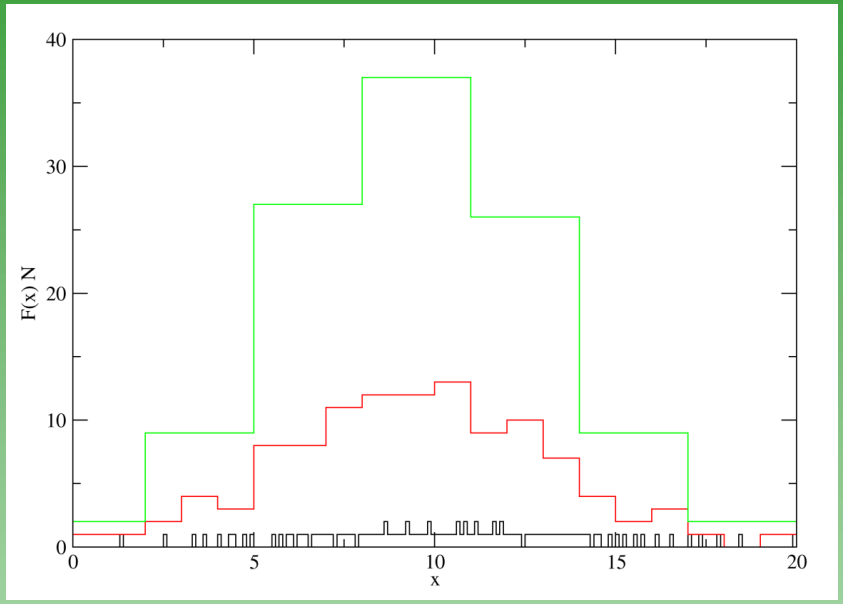


- Very fine binning is not the goal; nor the very large bin
- Instead, **want to match the bin size to the statistics**
- (***You'll see later that you also need to match it to the resolution of your detection system***)

Frequency Distribution: 'Inverse task'

Set of N *independent* measurements

$x_1, x_2, x_3, x_4, \ldots\ldots x_i, \ldots\ldots, x_N$



Possible to extract the same properties from the frequency distribution as from the raw data!

Sum $\sum_{x=0}^{\infty} N\, F(x) = N$

$\bar{x}_{exp} = \sum_{x=0}^{\infty} x\, F(x)$

Sample Variance

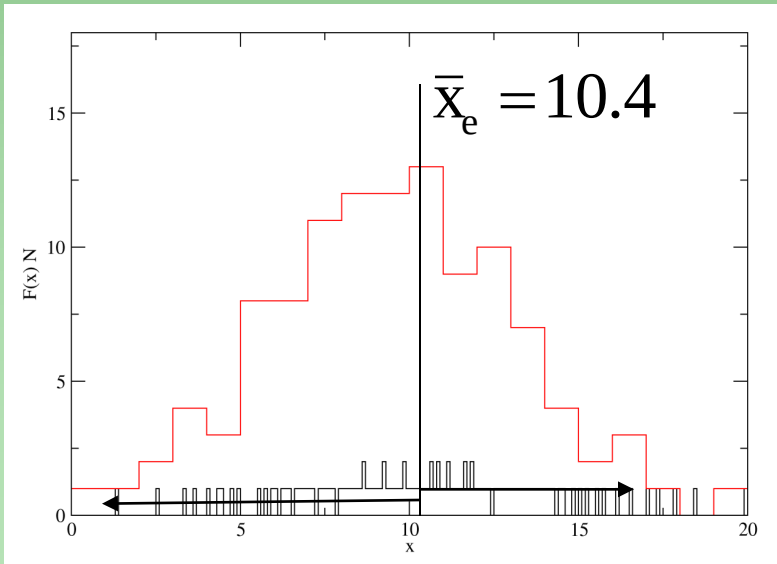
Set of N *independent* measurements $x_1, x_2, x_3, x_4, .. x_i, ., x_N$

$$\bar{x}_{exp} = \frac{\sum}{N} = \frac{1}{N}\sum_{i=1}^{n} x_i$$

One further parameter that we can extract from our distribution **– the sample variance**

Firstly, define the **residuals d_i** $\quad d_i \equiv x_i - \bar{x}_e$

Where, by definition of the mean: $\quad \sum_{i=1}^{N} d_i = 0$



Sample Variance

So, the variance of the sample is:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} d_i^2$$

Can also calculate this directly from the frequency distribution

$$\sigma^2 = \sum_{x=0}^{\infty} (x-\bar{x})^2 F(x)$$

Sample and Parent Distribution

One has been measuring properties of a **sample distribution**, with mean and variance

$$\bar{x} \qquad \sigma^2 = \sum_{x=0}^{\infty} (x - \bar{x})^2 F(x)$$

This has **some relation to the parent distribution**, from which we are sampling, which has mean and variance

$$\mu = \langle x \rangle \qquad \sigma^2_{\text{parent}} = \langle x^2 \rangle - \langle x \rangle^2$$

As $N \to \infty$ $\quad \bar{x} \to \langle x \rangle$ $\quad \sigma \to \sigma_{\text{parent}}$

It is often the parent distribution that we are trying to understand

Uncertainties

2

Uncertainties

2 ± 1

Uncertainties

2.0 ± 0.1

Uncertainties

2 ± 1

What does this number mean?

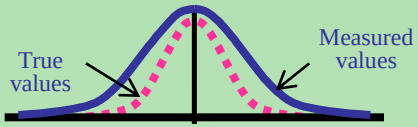
Typically, this is indicative of σ – the **standard deviation** of the distribution of possible values that we have measured

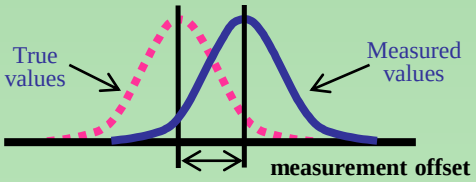
Uncertainty

What is uncertainty?

- Alternative expression for experimental error (but misleading – *error* is not synonymous with *mistake*, or *incorrectness*)
- It is NOT a description of how far your measurement is from the real value
- Instead, it tells how well you know the value
- If you repeat the measurement, how close are subsequent measurements likely to be?
- ***All*** measurements should include an uncertainty value

Precision vs accuracy?





UWS UNIVERSITY OF THE WEST of SCOTLAND

Introductory Data Analysis

Lecture 2

Statistical Uncertainty and Probability Distributions

SUPA Graduate School 2010

Prof. Andrei Andreyev

Andrei.Andreyev@uws.ac.uk

Nuclear Physics Research Group

School of Engineering and Science

Lecture 2 outline

- Precision vs Accuracy
- Uncertainties – statistical & systematic
- Properties of statistical uncertainties
- Properties of systematic uncertainties
- Presenting data with uncertainties

- Probability Distributions

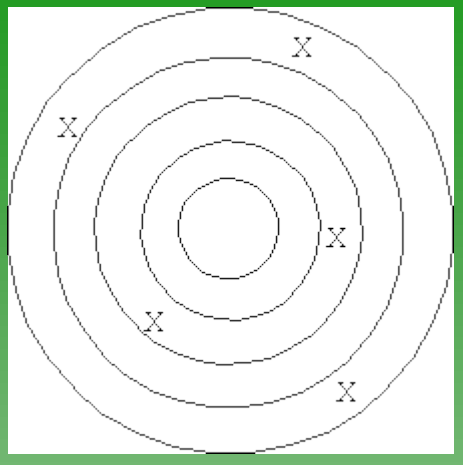
Uncertainties

2 ± 1

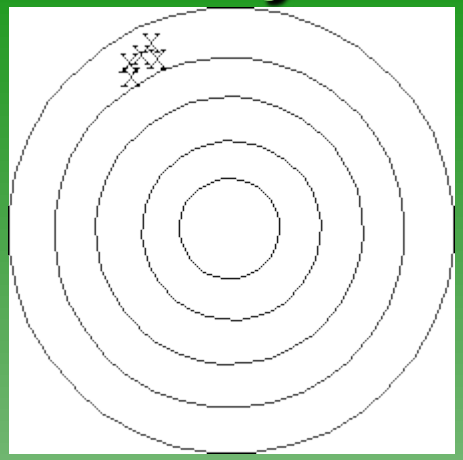
What does this number mean?

Typically, this is indicative of σ – the **standard deviation** of the distribution of possible values that we have measured

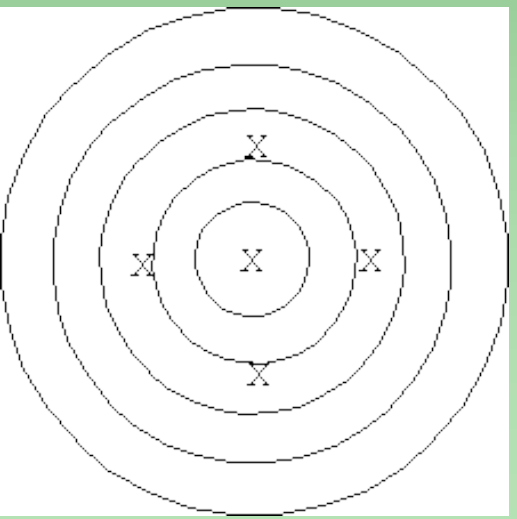
Precision vs Accuracy



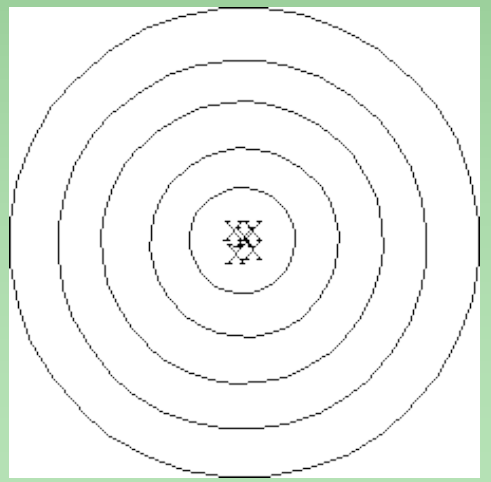
Radmom-like pattern, **neither precise nor accurate**. The darts are not clustered together and are not near the bull's eye.



This is a **precise pattern**, but **not accurate**. The darts are clustered together but did not hit the intended mark.



This is an **accurate pattern**, but **not precise**. The darts are not clustered, but their 'average' position is the center of the bull's eye.

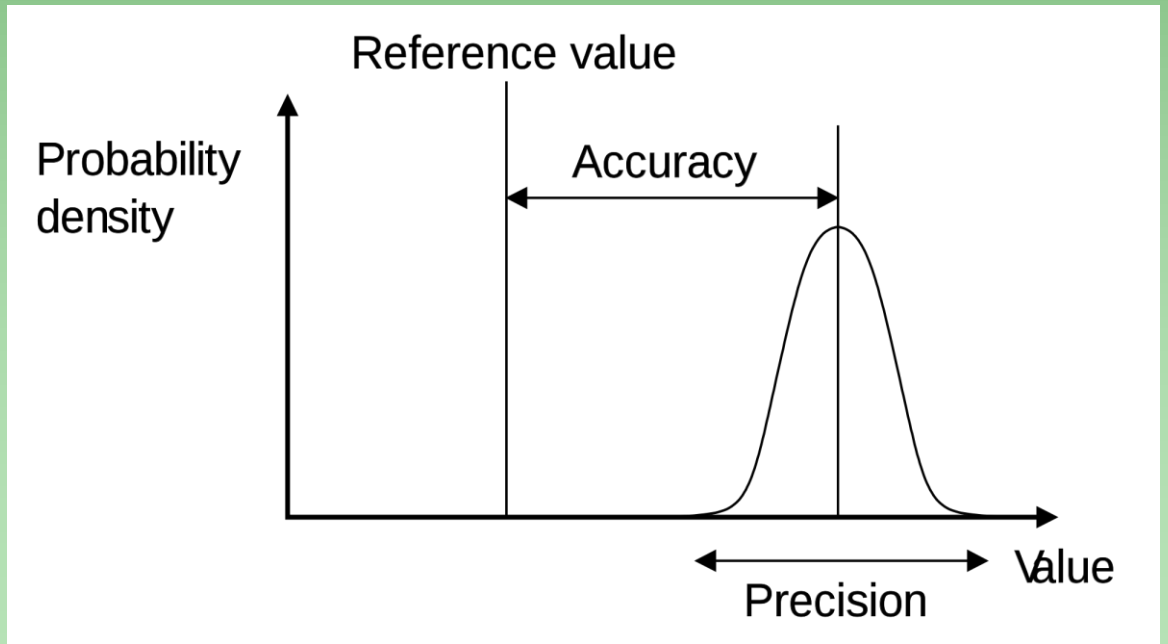


This pattern is **both precise and accurate**

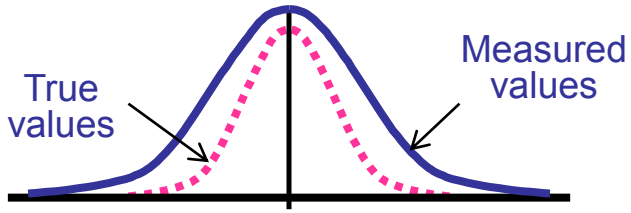
Precision vs Accuracy

•In the fields of science, engineering, industry and statistics, the **accuracy** of a measurement system is the **degree of closeness** of measurements of a quantity to its actual (true) value.

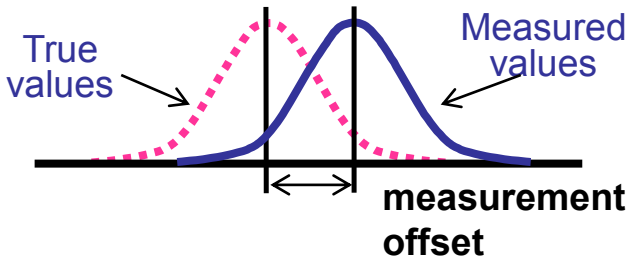
•The **precision** of a measurement system, also called **reproducibility or repeatability**, is the degree to which repeated measurements under unchanged conditions show the same results.



accurate, not precise



precise, not accurate



Why to determine uncertainty?

- Central to data analysis
- To understand what you have measured
- To report what you have measured
- To make comparisons with other measurements
- To compare with (various) theoretical predictions/models
- Draw conclusions

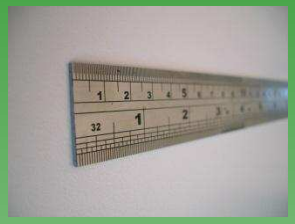
Contributions to Uncertainty

- **Intrinsic resolution of instruments**

http://en.wikipedia.org/wiki/Measuring_instrument



"atomic" watch



ruler

calliper

micrometer

- **Calibration of instruments** (linearity, offsets, efficiency, …)
- **Variations in experimental conditions** (temperature, pressure, humidity, light, rate-dependent dead time and/or efficiency…)
- **Backgrounds** (thermal background, cosmic rays, undesired contributions from competing mechanisms, vibrations…)
- **Genuine physical uncertainties** (eg radioactive decay)

Statistical and Systematic Uncertainty

Two types of uncertainty

Statistical uncertainties

*Uncertainties resulting from **random variations** in the experimental system or measured value itself – variations randomly in "both" direction (above or below the 'true' value).*

Usually, relatively easy to estimate.

Systematic uncertainties

*Uncertainties that are **introduced via the method of measurement itself** – usually in one (unknown) direction (eg. an offset of all measured values by a constant shift).*

*Also – due to the **theoretical model used to deduce/analyse data!***

(All in all, there could be several types of systematic uncertainties, 'shifted' in different 'directions')

Properties of statistical uncertainty

•Random or statistical uncertainties arise from random fluctuations in a measurement

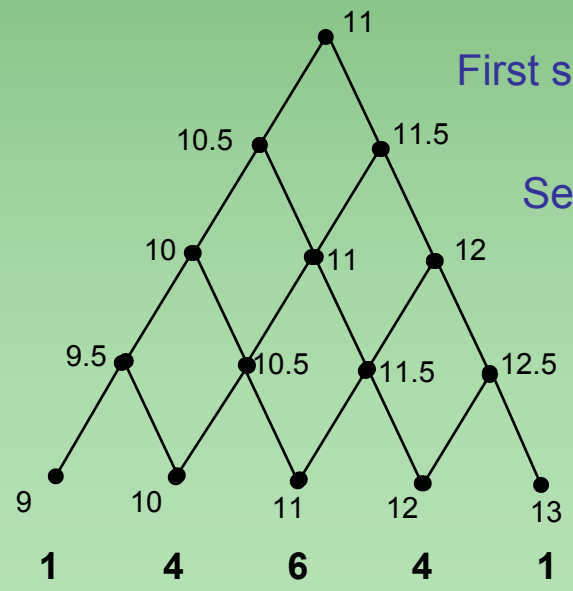
•These random fluctuations can be **due to the measuring device itself**. Eg., electronic noise and/or air currents lead to a rapid but small fluctuation in motion detector readings. These fluctuations occur, even when the motion detector is measuring the distance to a stationary object.

•Random fluctuations can also be a **characteristic of the quantity being measured**. Eg., if we use a meter stick to measure the landing positions of tennis balls from a tennis-ball launcher, we see significant random variations which do not arise from the limitations of the meter stick. Instead, we suspect that the launch velocity given to the balls by the launcher is subject to small random variations.

Properties of statistical uncertainty

•Random variations are **well described by statistical methods** (so we have tools to get a handle on the value we're trying to measure, and our uncertainties on it)

•Example: we measure a value (in our case - 11, to which 4 different uncertaintly sources contribute, each with an uncertainty of 0.5 relative to the 'main' value)



First source of uncertainty

Second source of uncertainty

Third source of uncertainty

Fourth source of uncertainty

Probability is proportional to the number of paths to each value

Properties of statistical uncertainty

•**Truly random fluctuations average to zero**, and so the way to remove them **is to average a large number of measurements**:

Experimental Mean/Average:

$$\bar{x}_{exp} = \frac{\sum}{N} = \frac{1}{N}\sum_{i=1}^{n} x_i$$

•The average/mean value approaches the ``true value" as the **number of measurements in the average approaches infinity**.

•Random fluctuations are described by **the normal distribution**, or Gaussian distribution. The uncertainty in the ”best value" of a large collection of normally distributed measurements can be calculated using the ***sample*** standard deviation

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Importance of systematic uncertainty

- A particular error is considered to be random if it affects only one data point (better to say, each point is affected, but in a 'random/different' manner
- It is, however, systematic if it affects **two or more data points** in a **correlated or anticorrelated** manner

- The overwhelming importance of the systematic component of an error may be realized from equation

$$TE = RE + SE,$$

where *TE represents the total error in the mean and RE* and *SE represent respectively the random and the systematic* components of the error.

- As RE approaches 0 after many measurements, the TE asymptotically approaches SE!

Properties of systematic uncertainty

- Systematic errors could arise due **to a defect in the equipment** or **methods used to make/analyse** the measurements.
- For example, a motion sensor can **be poorly calibrated** so that it gives distance readings which are only 90% of the true values. **It has a systematic uncertainty (a constant shift of 10%)** that is much greater in magnitude than the statistical uncertainty in its readings

- **Systematic errors are often difficult to detect**, because they do not show up as fluctuations in the results of repeated measurements.
- **It is important to think about possible sources of systematic errors** and to try to correct them or rule them out, for example by:
  - checking calibrations
  - comparing results with accepted values/etalons
  - comparing results obtained via independent means/methods

Properties of systematic uncertainty

•One particular type of systematic uncertainties is a **'drift' of some value**

Some examples:

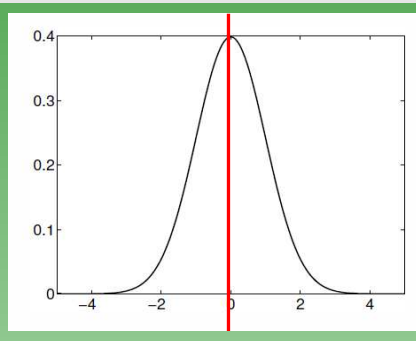
• Temperature-dependent drift of the measurement device or of the experimental setup (e.g. the magnetic field of an accelerator)

• Rate-dependent efficiency of the data-acquisition system (dead-time..)

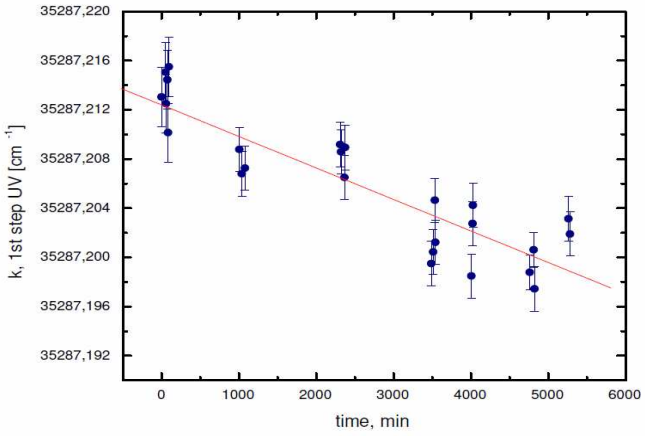
•This type of systematic uncertainty can usually be corrected/accounted for by using proper calibrations before the real experiment starts

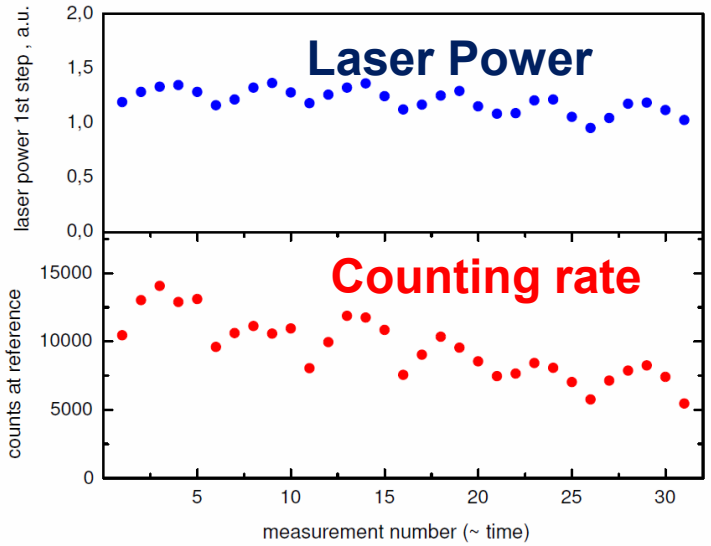
Properties of systematic uncertainty

Example: **Temperature(time)-dependent drift** of the **measurement device** (in this case, wave-length/frequency meter of a laser system) and of the laser power (CERN experiment to measure radii of exotic nuclei)



Measurements at the peak of a resonance, thus ideally – need a **constant laser frequency at the peak value**





Ratio Rate/Power = Const!

Figure 6.3: Variation of the resonance wavenumber for ^{208}Pb as a function of time, in minutes, as a result of the drift of the wavemeter. The errorbars are statistical and depend on the step size in the frequency scan (0.02,0.01 or 0.005 cm^{-1}, before frequency doubling)

3.6: (Upper plot) Read-out of the laser power in the first excitation step, mea- regular time intervals during a scan. (Lower plot) α-intensity at the reference r, for the consecutive reference measurements.

Properties of systematic uncertainty

Example: Measurement of the mass of a nucleus in a high-precision Penning Trap.

Here, **the resonance frequency,** which is used to deduce the mass of a nucleus, depends (slightly) **on the number of nuclei in the trap!**

C. Weber et al. / Nuclear Physics A 803 (2008) 1–29

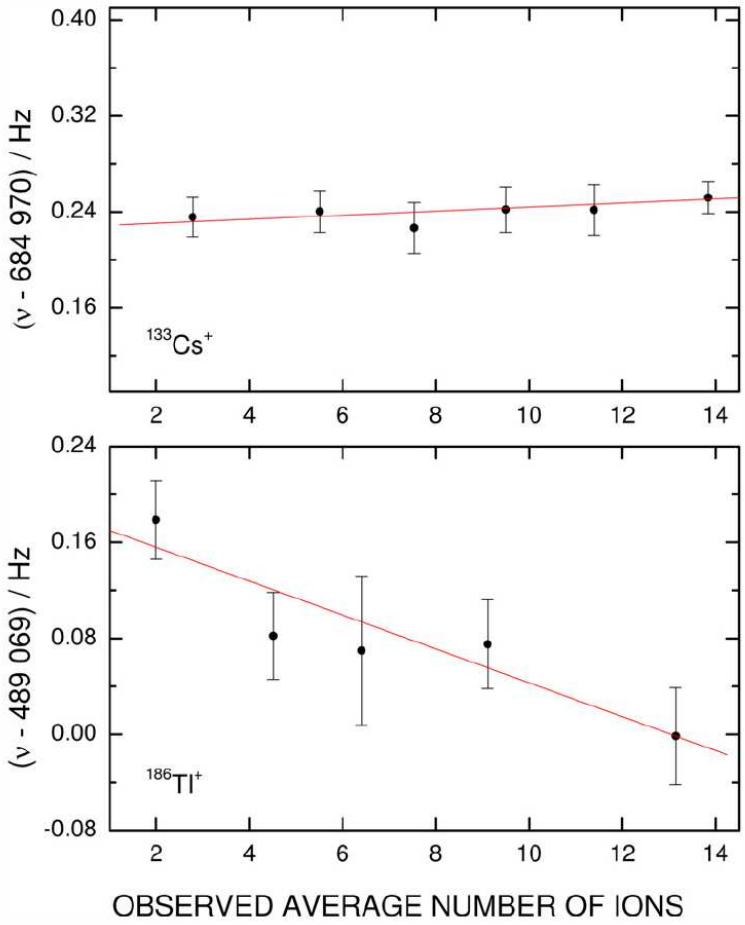
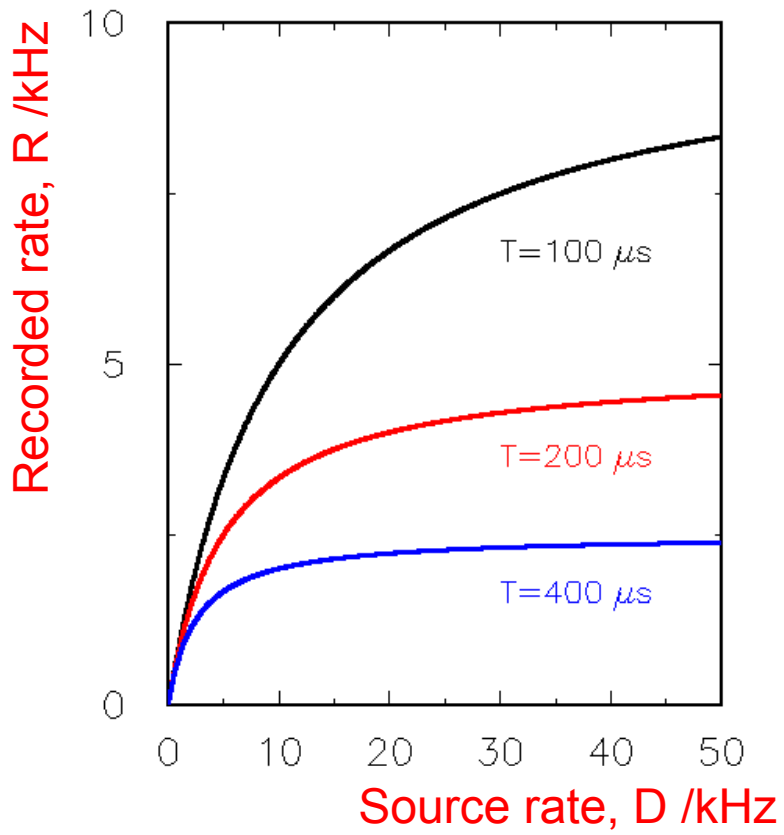


Fig. 3. Dependance of the frequency of the cyclotron resonance on the number of detected ions for ^{133}Cs (top) and ^{186}Tl (bottom). Both measurements were performed with an excitation time $T_{obs} = 1.5$ s. The center frequency ν is plotted versus the count-rate class containing an average number of $N_{av} = 320$ ions as detected by MCP 5 (see Fig. 1). The straight line is a linear regression weighted by the individual error bars. The frequency ν_c, extrapolated to zero ions and its respective error is used as the final result.

Properties of systematic uncertainty

Example: Rate-dependence of a non-paralyzable data acquisition (DAQ) system



The amount of recorded data of a triggered DAQ depends on the rate of triggers delivered to the system and on the time it takes to processs one event.

In the case of non-paralyzable detectors and electronics the ratio between the number of delivered events **D** and recorded events **R** can be calculated for random time distributions of the events for a given dead time **T** as:

R = D / (1 + D*T)

Properties of systematic uncertainty

•Another (very common and important) type of **systematic uncertainties** is due to the **theoretical models used** to extract or interpret measured experimental values

•Eg. We measured some continuous distribution, now we have to extract eg., it's mean value (or, eg. half-life for a radioactive decay)

•**Here, we have to make some assumption on the 'expected' type of the distribution we believe is most appropriate to the measured value**. – Is it Gaussian? Poisson, Lorenzian, exponential...?

•Furthermore, **is there any background**? And if 'yes' – what the **most appropriate fitting function** should be in this case?

Properties of systematic uncertainty

Example: **Half-life determination of a nucleus by fitting its decay curve with a set of functions:**
$N(t)=A_o e^{-\lambda 0*t}$+Second function ($A_1 e^{-\lambda 1*t}$, Linear? Constant?)

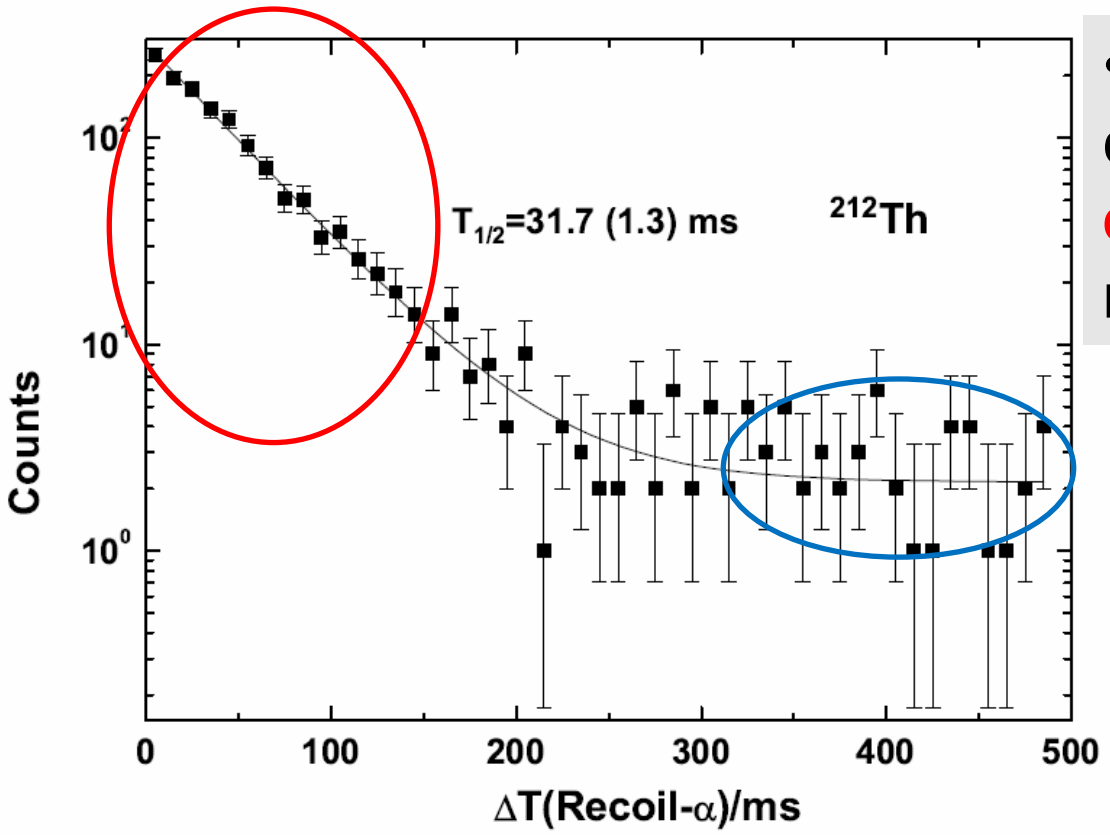


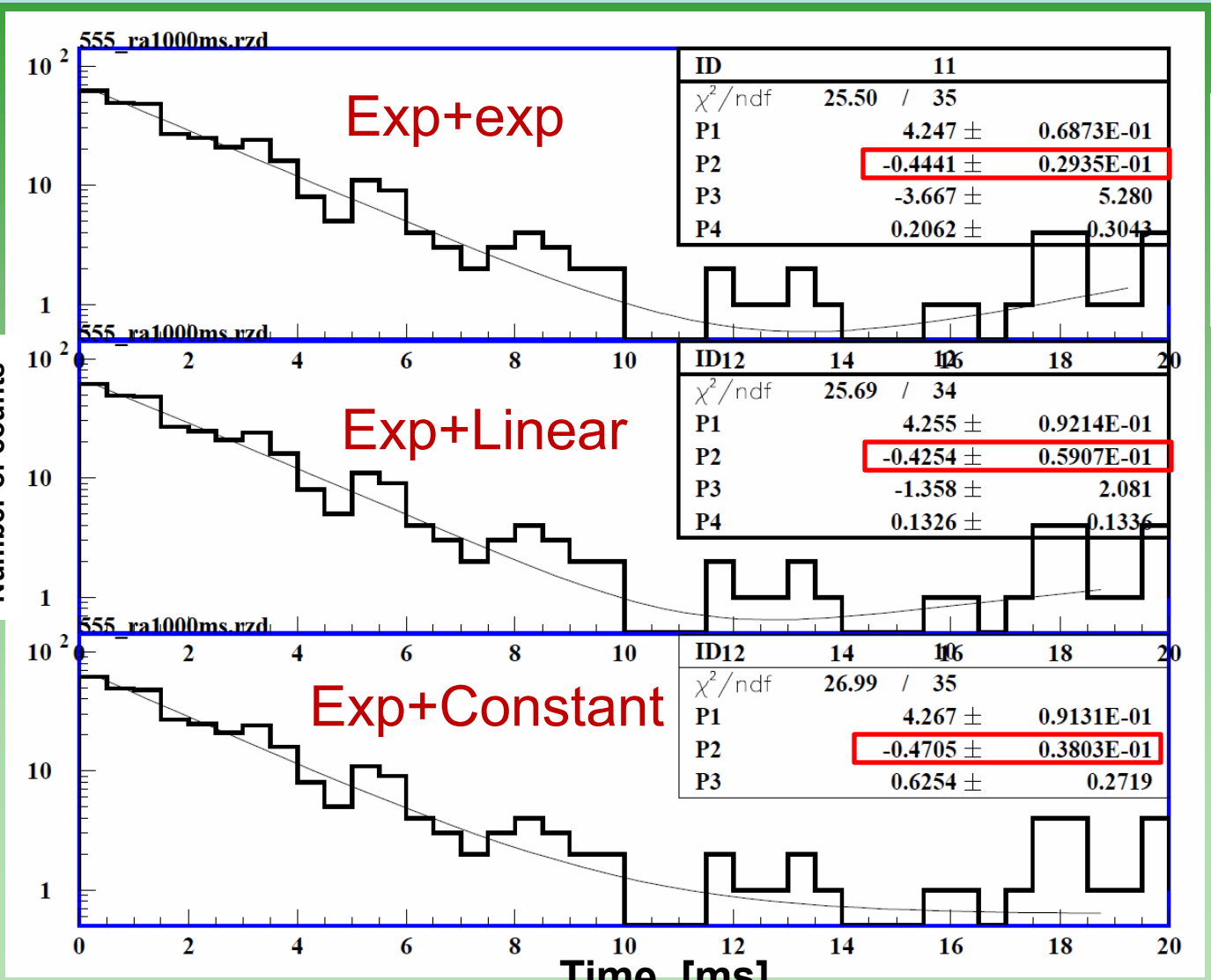
Fig. 3: ER-α time distribution of the ^{212}Th α decays from fig 1a. The continuous solid line shows the result of an exponentia decay fit with a constant background.

- Main part of the decay curve is of course **exponential** (exponential radioactive decay $A_o e^{-\lambda 0*t}$)

- But, what about the '**background**' part of the decay curve? Is it also exponential $A_1 e^{-\lambda 1*t}$ or is it linear?
- And, if 'linear' – is it a **constant** or is it a '**sloping**' (~C*t) line?

Properties of systematic uncertainty

Example: **Half-life determination of a nucleus by fitting its decay curve (the same data!) with 3 variants of functions**



$T_{1/2}=\ln 2/P2$

$T_{1/2}=1.56$ ms
smallest fit uncertainty

$T_{1/2}=1.63$ ms

$T_{1/2}=1.47$ ms

Reporting Results with Uncertainties

•Experimental Results are always reported with uncertainties, both statistical and systematic (if known)

•The uncertainty is assumed to be in **the last reported digit of the result (very important to remember this!)**

- $x \pm \sigma_{stat}$ OR $x(\sigma_{stat})$

  e.g. 10 ± 1 cm or 10(1) cm

  10.1 ± 0.1 cm or 10.1(1) cm - look for precision!

  Not: 10.1(0.1) cm !

  10.11 ± 1.20 cm or 10.11(120) cm

  Not 10.11(1.20) cm

  10.110 ± 1.20 cm or 10.11 ± 0.111 cm - **BOTH**

WRONG - under/over-precision

- $x \pm \sigma_{stat} \pm \sigma_{syst}$ OR $x(\sigma_{stat})(\sigma_{syst})$

  e.g. $10 \pm 1_{stat} \pm 2_{syst}$ cm or 10(1)(2) cm

Reporting Results with Uncertainties

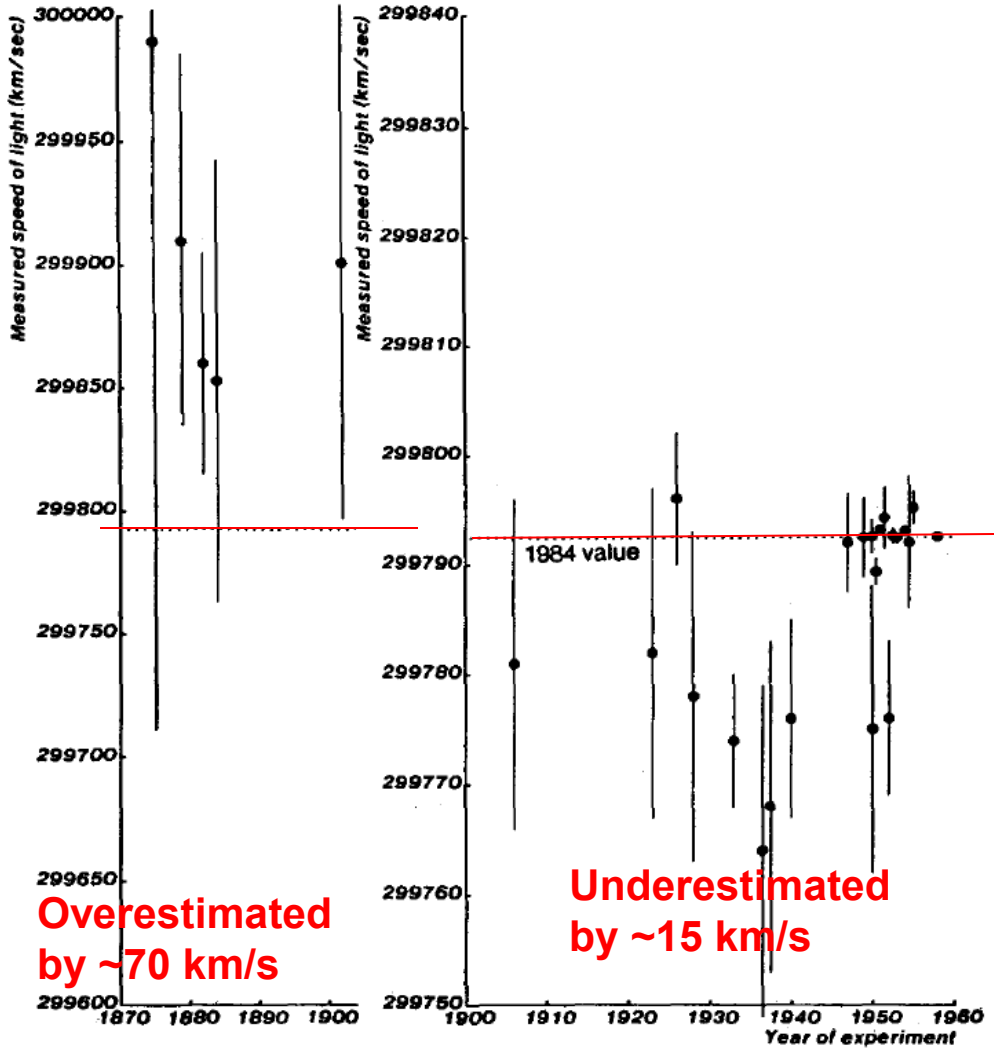
Example: Mass of top quark

http://www.science20.com/quantum_diaries_survivor/who_underestimates_their_systematic_uncertainties

$170.7^{+\ 4.2}_{-\ 3.9} \pm 3.5$	[14,15] AALTONEN	08C	CDF	dilepton, $\sigma_{t\bar{t}}$ constrained
$177.1 \pm 4.9 \pm 4.7$	[16,17] AALTONEN	07	CDF	6 jets with ≥ 1 b vtx
$172.3^{+10.8}_{-\ 9.6} \pm 10.8$	[18] AALTONEN	07B	CDF	≥ 4 jets (b-tag)
$173.7 \pm 4.4^{+\ 2.1}_{-\ 2.0}$	[17,19] ABAZOV	07F	D0	lepton + jets
$170.3^{+\ 4.1}_{-\ 4.5} \pm 1.2 \pm 1.8$	[4,20] ABAZOV	06U	D0	lepton + jets (b-tag)
$173.2^{+\ 2.6}_{-\ 2.4} \pm 3.2$	[21,22] ABULENCIA	06D	CDF	lepton + jets
$173.5^{+\ 3.7}_{-\ 3.6} \pm 1.3$	[15,21] ABULENCIA	06D	CDF	lepton + jets
$165.2 \pm 6.1 \pm 3.4$	[4,23] ABULENCIA	06G	CDF	dilepton
$170.1 \pm 6.0 \pm 4.1$	[15,24] ABULENCIA	06V	CDF	dilepton
$178.5 \pm 13.7 \pm 7.7$	[25,26] ABAZOV	05	D0	6 or more jets

Measuring Velocity of Light

1875-1958



1929-1973

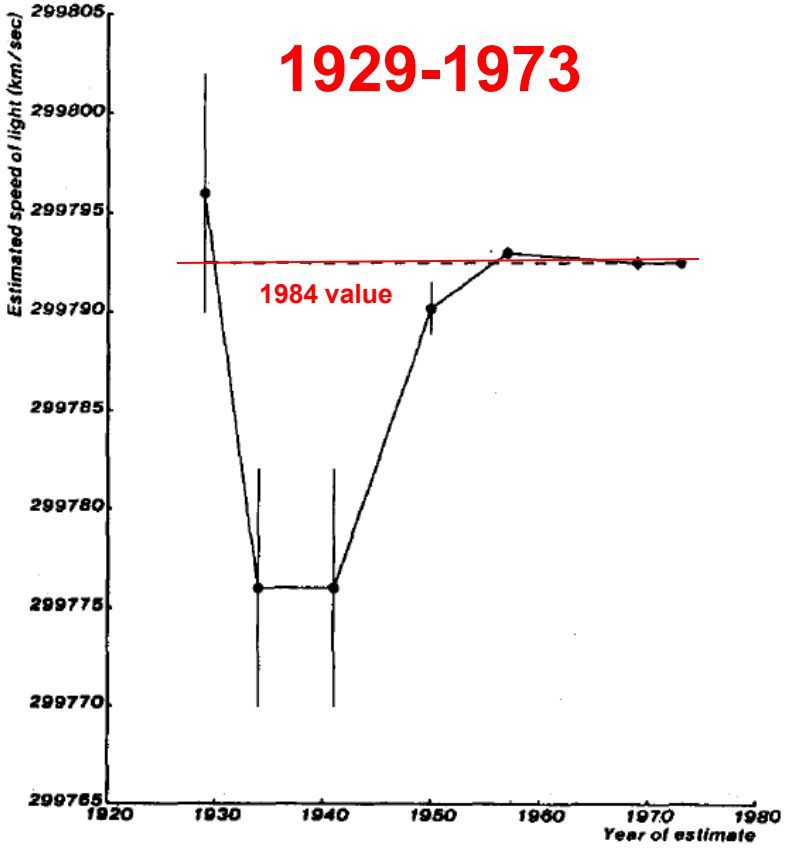
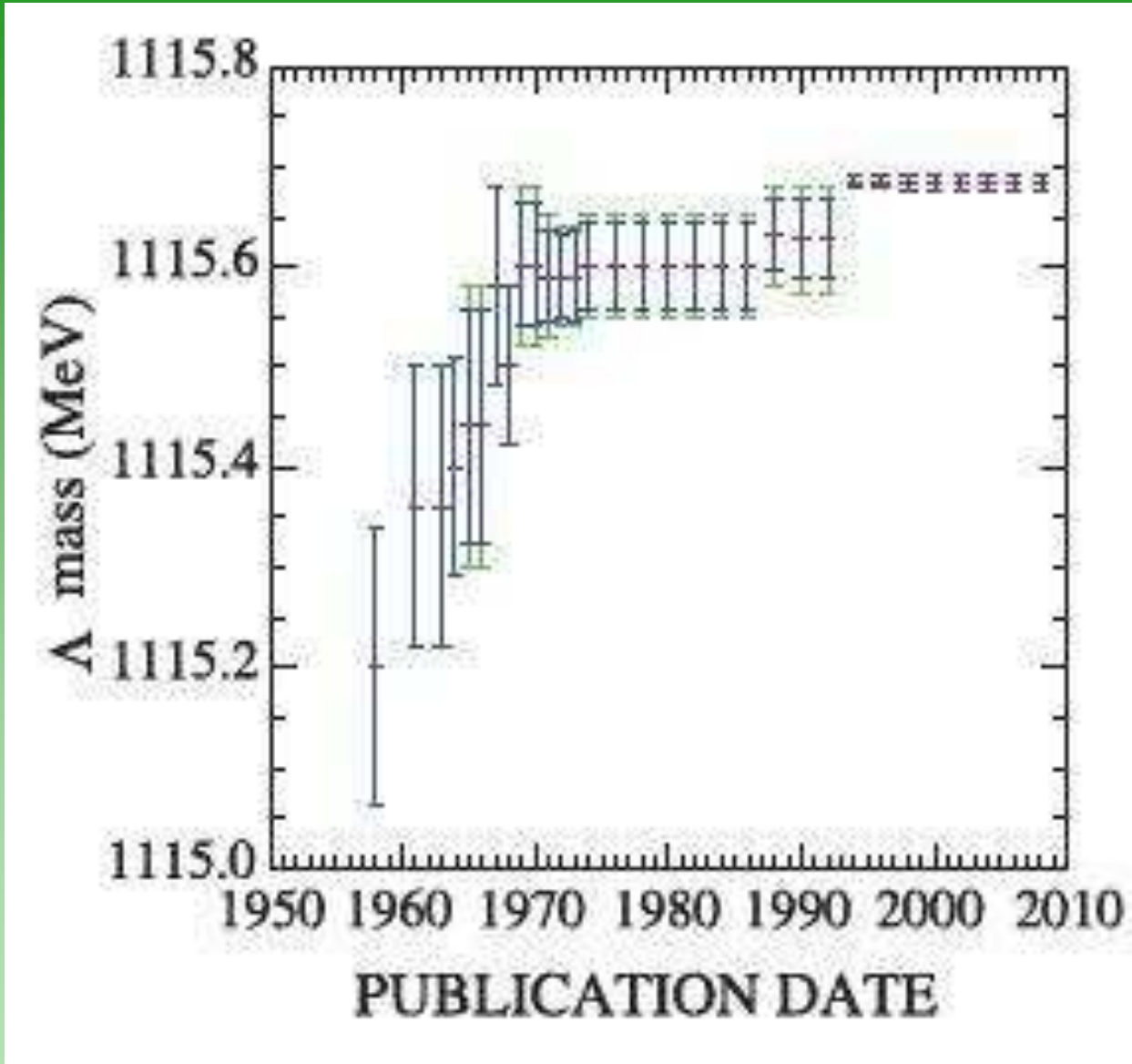


Fig. 2. Recommended values for the velocity of light; 1929–1973.

Measuring Mass of Lambda



Probability Distributions

- **Binomial Distribution**
- **Poisson Distribution**
- **Gaussian Distribution**
- **Gaussian distribution as a limit of the Binomial/ Poisson distributions**
- **Simple Counting Experiment**
- **Central Limit Theorem**

The Binomial Distribution

- Extremely general statistical model
- Applicable to describe the **frequency distribution** of a set of **many repeated measurements** of **constant likelihood "p"** (e.g. probability to get one of the values when throwing a dice is p=1/6)
- Discrete distribution of a binary process (**True** or **False**; **Yes** or **No**)
- Unwieldy when the numbers involved become large (can simplify in certain limits)
- It is frequently used to model number of successes in a sample of size ***n*** from a population of size ***N***. For *N* much larger than *n*, the binomial distribution is a good approximation.

The Binomial Distribution

Assume that we make a number of trials n

Each trial has a constant likelihood of success p

(therefore, the chance of failure is $1-p$)

Then the predicted probability of x 'successes' is

$$P(x) = \frac{n!}{(n-x)!\,x!}\,p^{x}(1-p)^{n-x}$$

for integer values of n and x

Mean $\bar{x} = pn$

Variance: $\sigma^2 = np(1-p)$

The Binomial Distribution

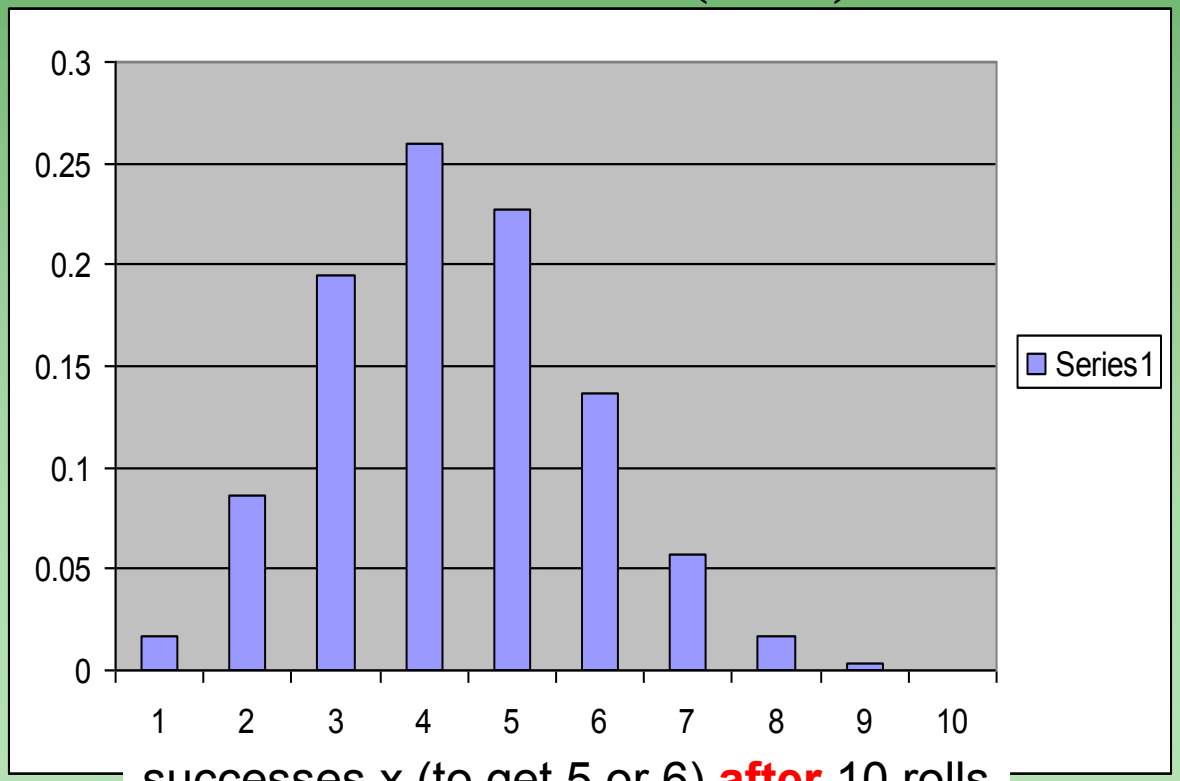
- Assume we role a die
- We call a "success" if we get a 5 or 6 – $p = {}^1/_3$
- We make 10 rolls (10 *trials*) $n = 10$

$$P(x) = \frac{n!}{(n-x)!\,x!}\,p^x(1-p)^{n-x}$$

$$\sum_{x=0}^{n} P(x) = 1$$

$$\bar{x} = \sum_{x=0}^{n} xP(x)$$

$$\bar{x} = pn = 3.3333$$



The Binomial Distribution

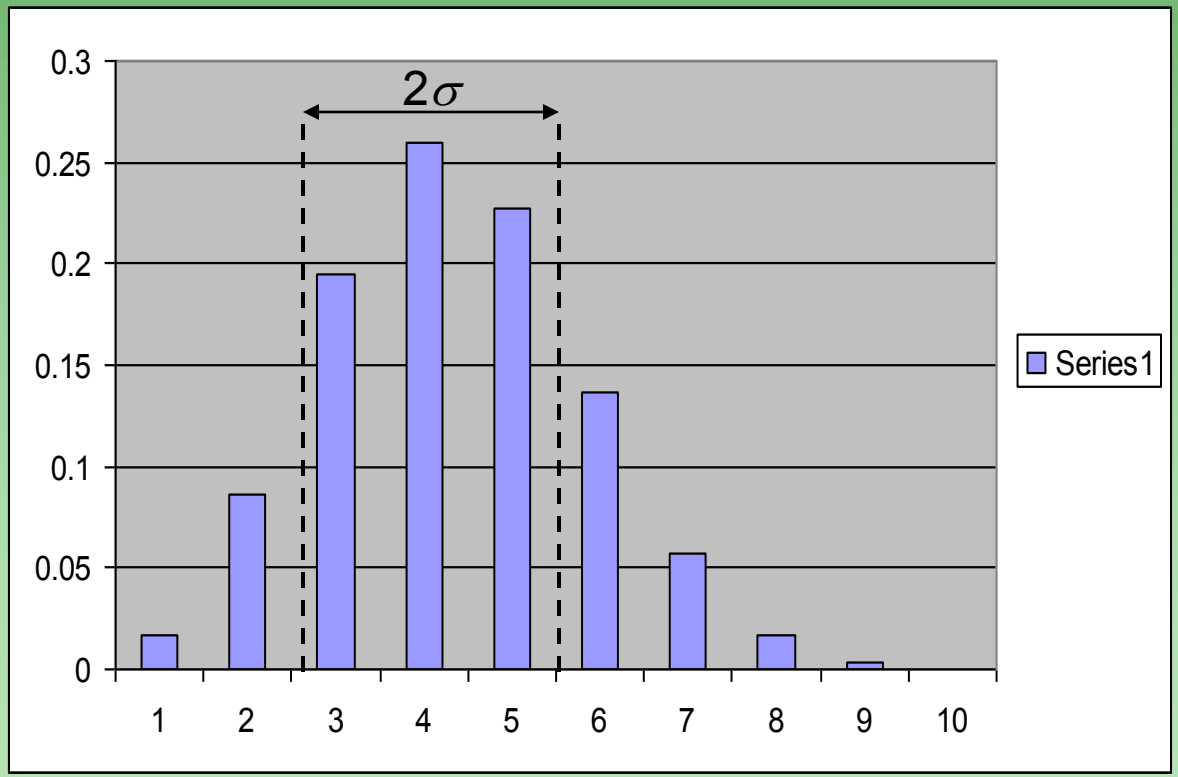
Width is given by

$$\sigma^2 = \sum_{x=0}^{n} (x - \bar{x})^2 P(x) \qquad \sigma = 1.49$$

Or alternatively

$$\sigma^2 = np(1-p)$$

The width σ gives us an idea of the typical deviation from the true mean of any single measurement



The Binomial Distribution

For large number of trials, numbers soon blow up to be unrealistically manageable

$$P(x) = \frac{n!}{(n-x)!\,x!}\, p^{x}(1-p)^{n-x}$$

$10! = 3.6 \times 10^{6}$

$20! = 2.8 \times 10^{18}$

$50! = 3 \times 10^{64}$

$100! = 9 \times 10^{157}$

Imagine if we were dealing with a radioactive sample (very large number of nuclei!) – impossible to handle with Binominal distribution!

The Poisson Distribution

In the limit where the **number of trials is large** and **the success probability is small**, the binomial distribution reduces to the Poisson form

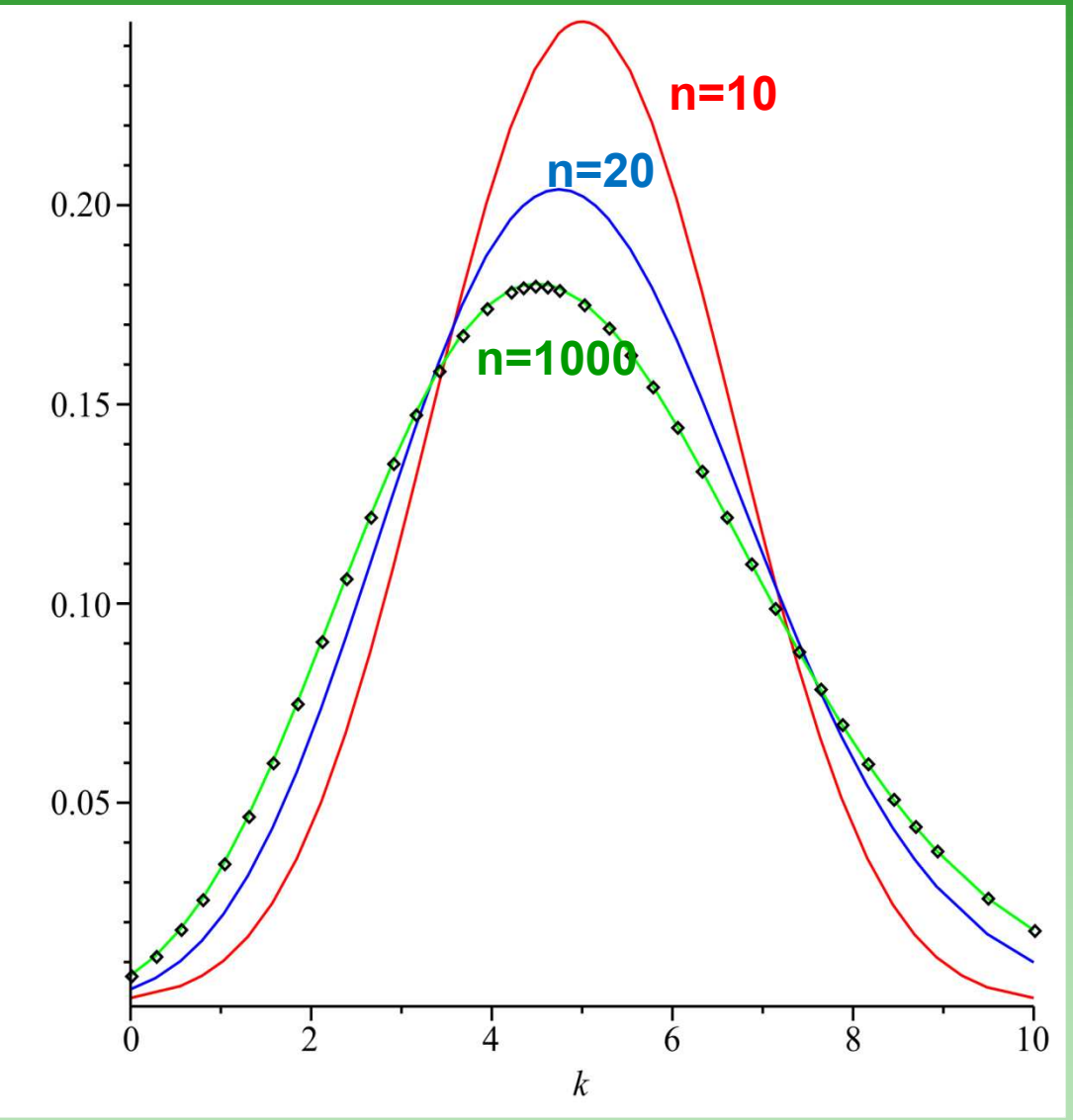
$$P(x) = \frac{(pn)^x e^{-pn}}{x!}$$

Which, as $\bar{x} = pn$, is equivalent to

$$P(x) = \frac{\bar{x}^x e^{-\bar{x}}}{x!}$$

Again, for integer values of n and x

The Binominal and Poisson Distribution



Comparison of the **Poisson distribution (black dots)** and the binomial distribution with **n=10 (red line), n=20 (blue line), n=1000 (green line).** All distributions have a mean of 5. The x-axis shows the number of events *k*. Notice that as n gets larger, the Poisson distribution becomes an increasingly better approximation for the binomial distribution with the same mean.

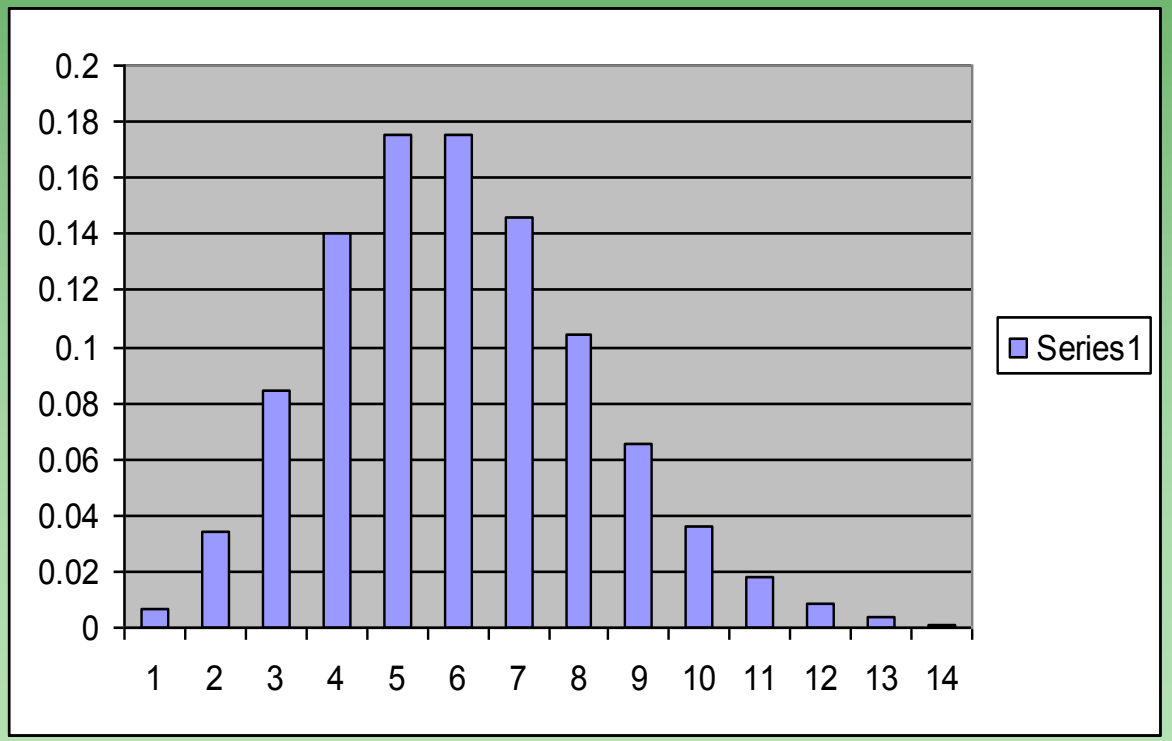
The Poisson Distribution

Assume we now have a process with n = 1000 trials and our success probability is p = 0.005

$$\sum_{x=0}^{n} P(x) = 1$$

$$\bar{x} = \sum_{x=0}^{n} xP(x)$$

$$\bar{x} = pn = 5$$



The Poisson Distribution

However, now

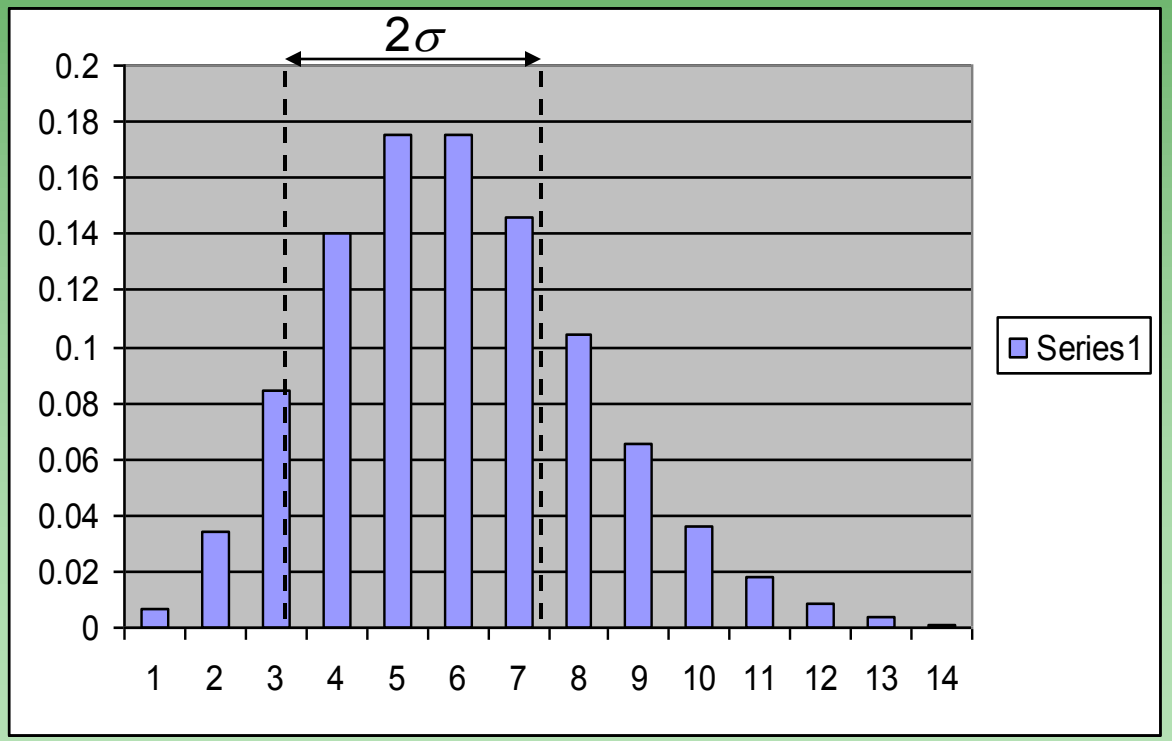
$$\sigma^2 = \sum_{x=0}^{n} (x - \bar{x})^2 P(x) = pn$$

And, as

$$\bar{x} = pn$$

$$\sigma = \sqrt{\bar{x}}$$

Very important result!



UWS UNIVERSITY OF THE WEST of SCOTLAND

Introductory Data Analysis

Lecture 3

Statistical Uncertainty and Probability Distributions

SUPA Graduate School 2010

Prof. Andrei Andreyev

Andrei.Andreyev@uws.ac.uk

Nuclear Physics Research Group

School of Engineering and Science

Lecture 3 outline

- Properties of systematic uncertainties
- Presenting data with uncertainties
- Probability Distributions
    - Binominal
    - Poisson
    - Gaussian(Normal)

Properties of systematic uncertainty

•Another (very common and important) type of **systematic uncertainties** is due to the **theoretical models used** to extract or interpret measured experimental values

•Eg. We measured some continuous distribution, now we have to extract eg., it's mean value (or, eg. half-life for a radioactive decay)

•**Here, we have to make some assumption on the 'expected' type of the distribution we believe is most appropriate to the measured value**. – Is it Gaussian? Poisson, Lorenzian, exponential...?

•Furthermore, **is there any background**? And if 'yes' – what the **most appropriate fitting function** should be in this case?

Properties of systematic uncertainty

Example: **Half-life determination of a nucleus by fitting its decay curve with a set of functions:**
$N(t)=A_o e^{-\lambda 0*t}$+Second function ($A_1 e^{-\lambda 1*t}$, Linear? Constant?)

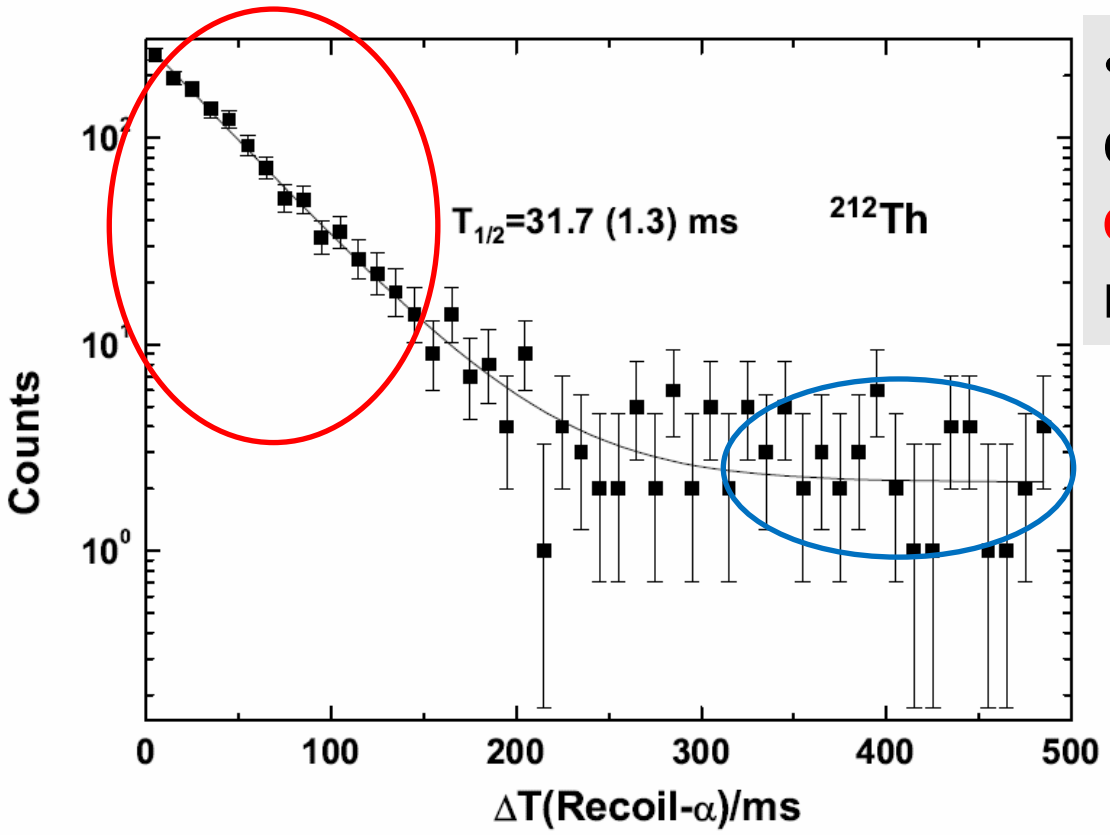


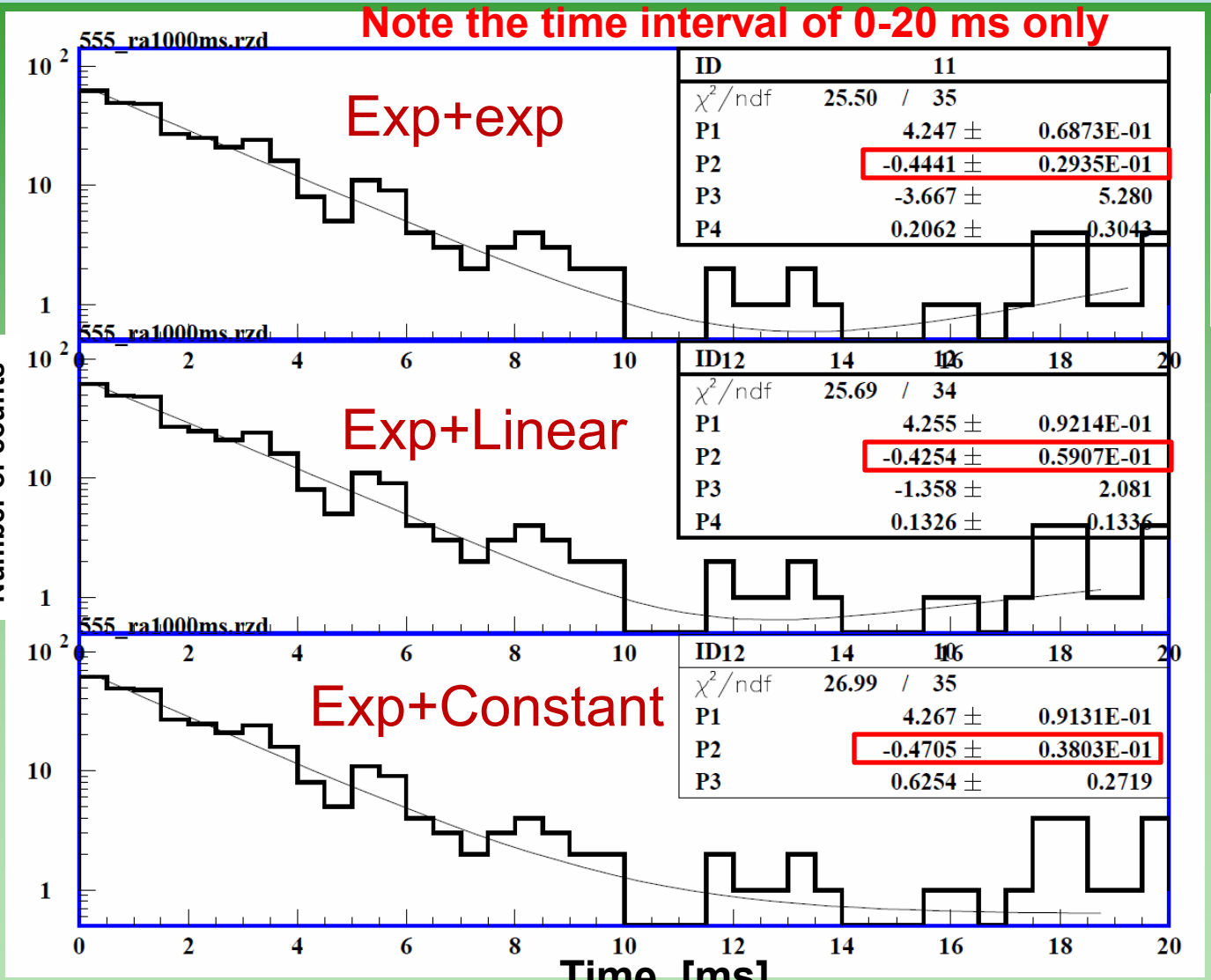
Fig. 3: ER-α time distribution of the ^{212}Th α decays from fig 1a. The continuous solid line shows the result of an exponentia decay fit with a constant background.

- Main part of the decay curve is of course **exponential** (exponential radioactive decay $A_o e^{-\lambda 0*t}$)

- But, what about the '**background**' part of the decay curve? Is it also exponential $A_1 e^{-\lambda 1*t}$ or is it linear?
- And, if 'linear' – is it a **constant** or is it a '**sloping**' (~C*t) line?

Properties of systematic uncertainty

Example: **Half-life determination of a nucleus by fitting its decay curve (the same data!) with 3 variants of functions**



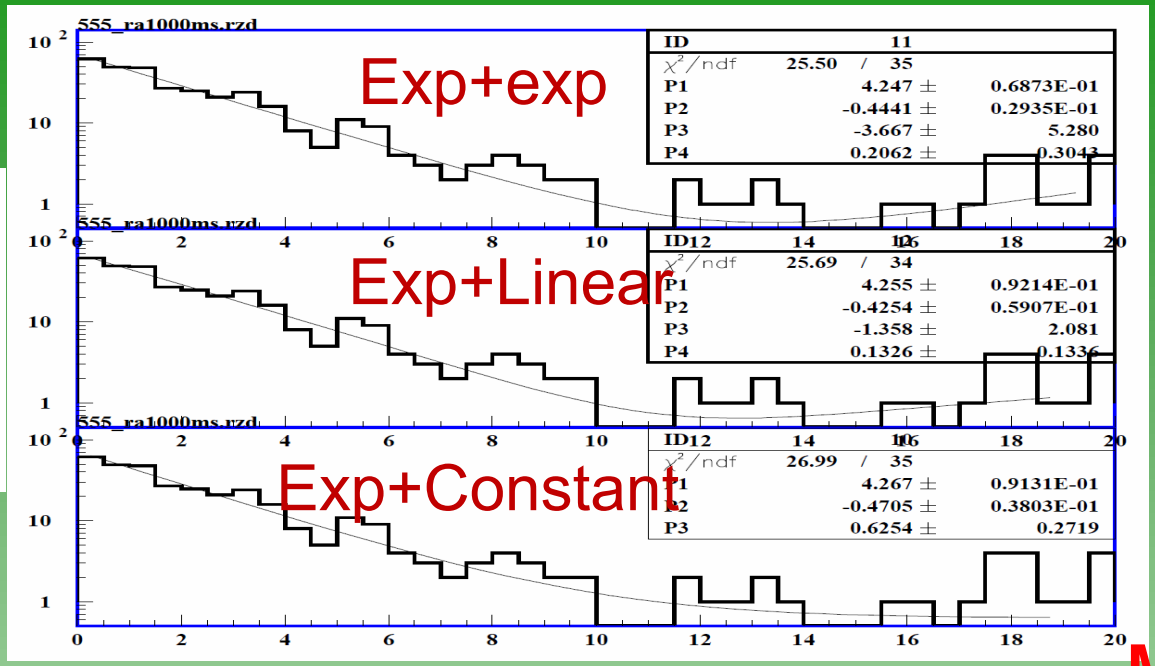
$T_{1/2}=\ln 2/P2$

$T_{1/2}=1.56$ ms
smallest fit uncertainty

$T_{1/2}=1.63$ ms

$T_{1/2}=1.47$ ms

Properties of systematic uncertainty



$T_{1/2}$=1.56 ms

$T_{1/2}$=1.63 ms

$T_{1/2}$=1.47 ms

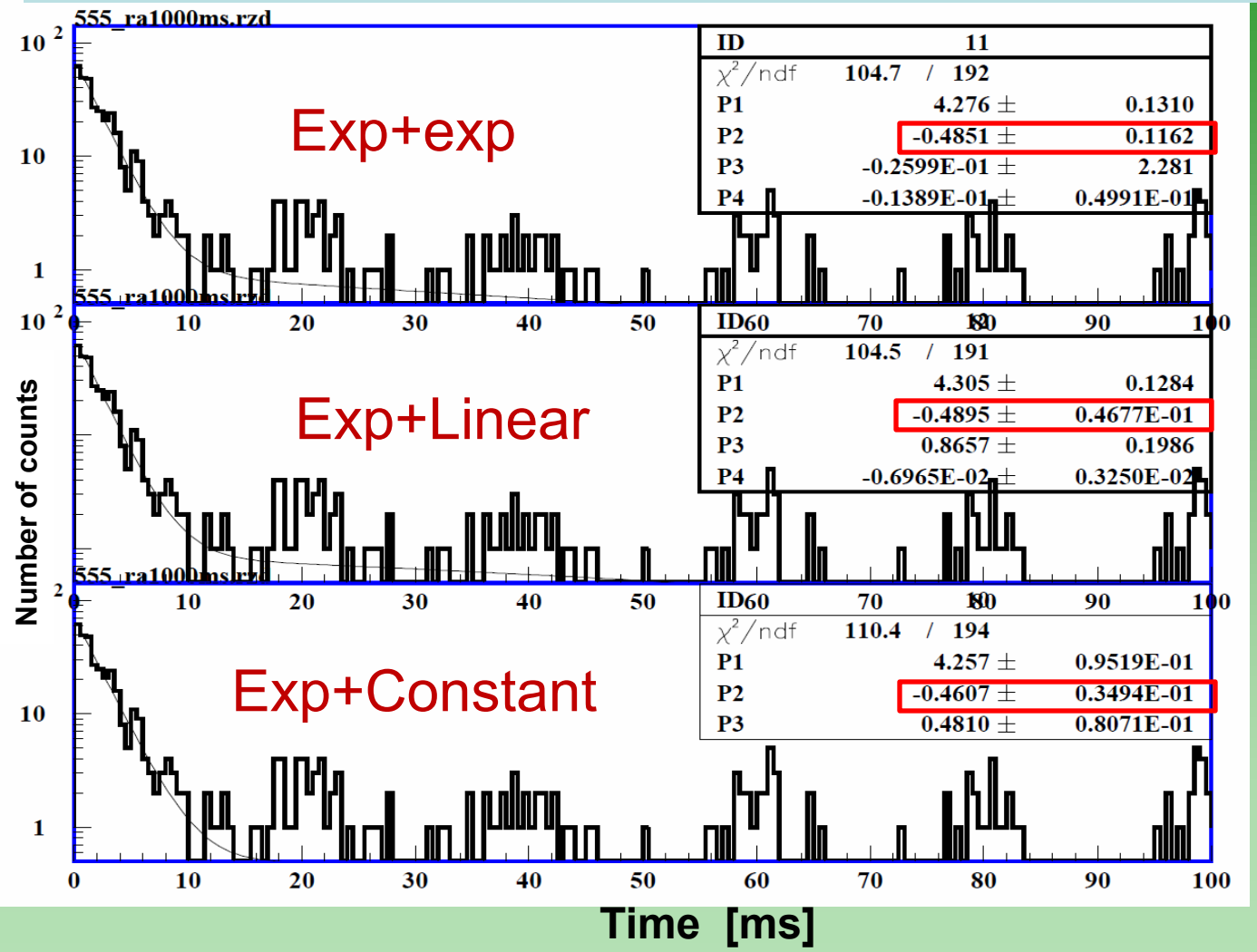
Max Difference 0.16 ms

Why different $T_{1/2}$ values?

- Most probably, because, in this case, the time **interval of 0-20 ms is too short to determine the background's behaviour with the necessary precision**
- **Need a longer interval for fitting!?**

Try a longer time interval (0-100 ms)

Now, try the “same” data, but **in the larger range** (up to 100 ms) – this should make the ‘background’ determination more ‘stable’



$T_{1/2}=\ln 2/P2$

$T_{1/2}=1.43$ ms

$T_{1/2}=1.42$ ms

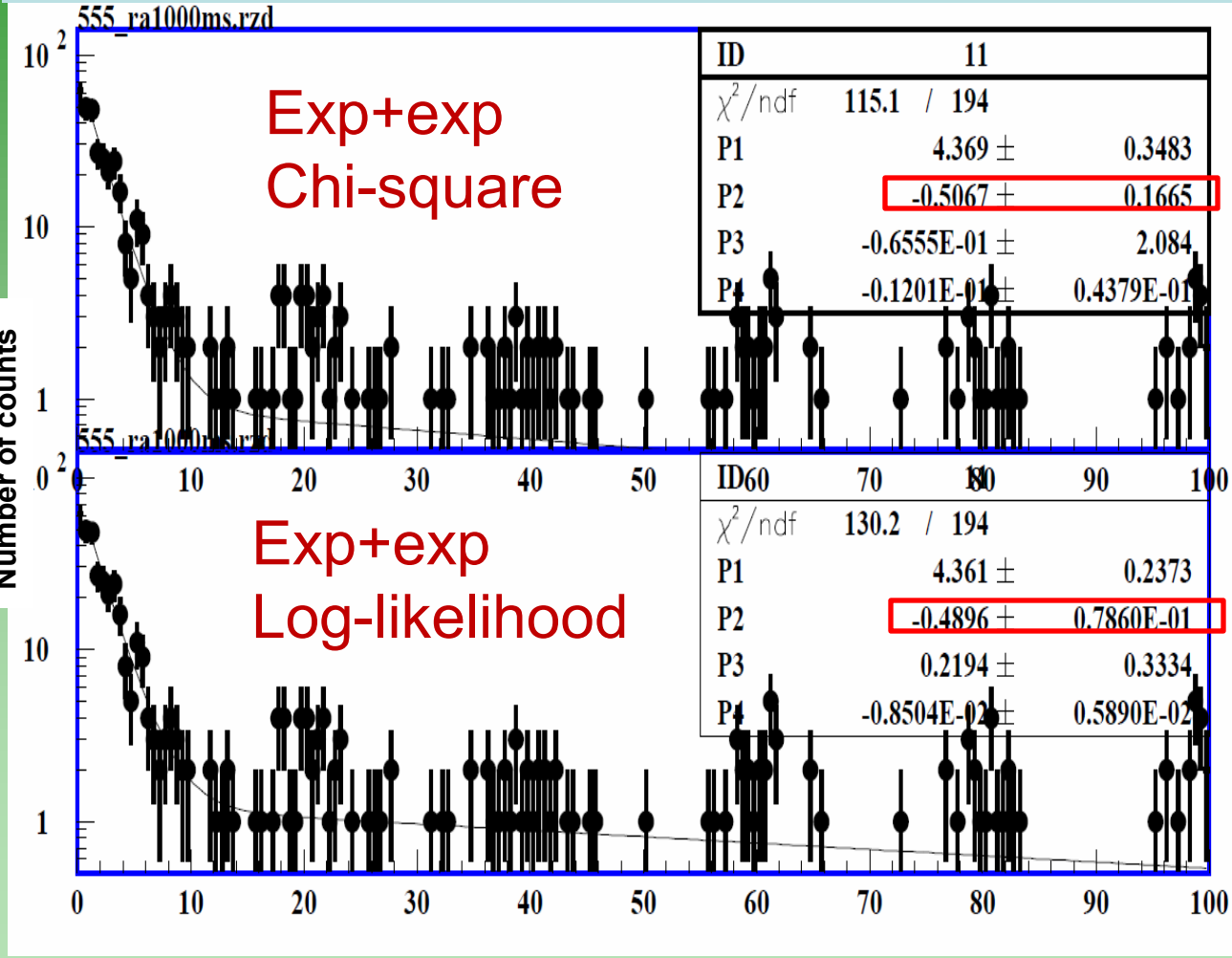
$T_{1/2}=1.50$ ms

Max Difference now is 0.08 ms!

It seems, the results are now much closer to each other!

Chi-square vs Log. Likelihood

Now, try the "same" data, in the range up to 100 ms, fitting with 2 exponents, but **using either Logariphmic Likelihood or Chi-square methods** (to be studied later)



$T_{1/2}$=ln2/P2

$T_{1/2}$=1.37 ms

$T_{1/2}$=1.42 ms

Difference now is 0.05 ms

Quite usual situation in the analysis

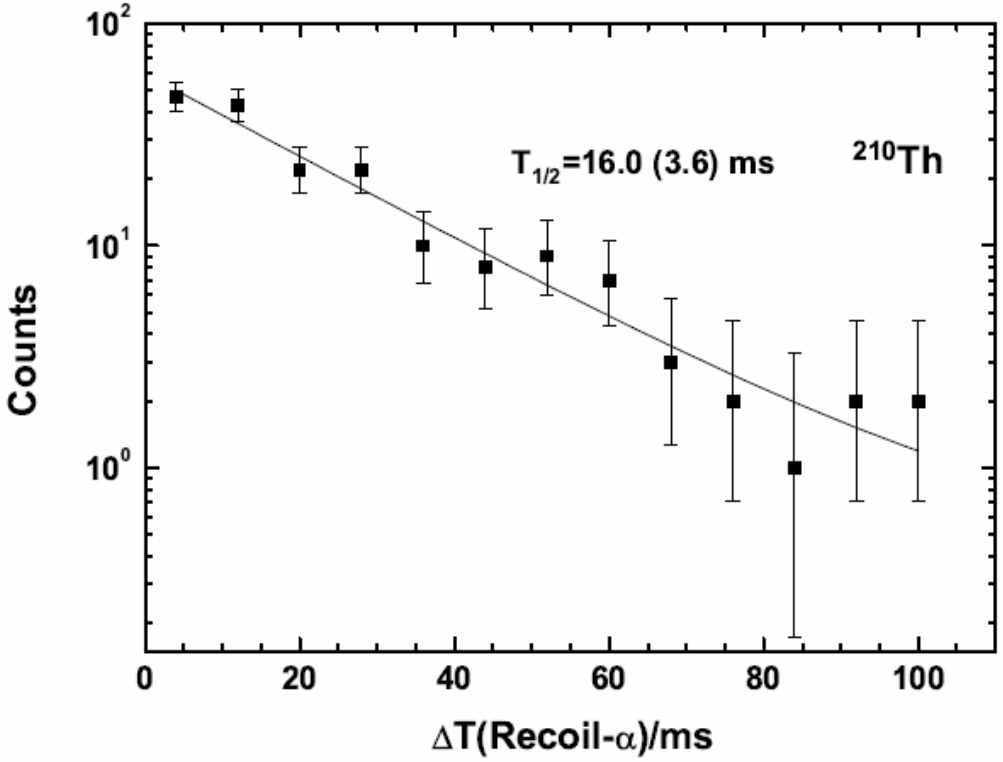


Fig. 2: ER-α time distribution of the ^{210}Th α decays from fig.1b. The continuous solid line shows the result of an exponential decay fit.

- **Low statistics**
- **Too short range for fitting**
- **....**

- In the Figure on the left –**the range is not enough** to say something certain on **how the background looks like**
- **We made a simple exponential fit only.**

Conclusions on the whole $T_{1/2}$ example

•Another (very common and important) type of **systematic uncertainties** is due to the **theoretical models used** to extract or interpret measured experimental values

•YES, indeed – **depending on our choice of the model fitting parameters** such as, the range for fitting, fitting functions, and/or the method of fitting – '**somewhat' different $T_{1/2}$ values were obtained!**

•Though the example was from Nuclear Physics, but it illustrates the general approach and problems arising in most of fitting procedures.

•Watch out carefully for model-related systematic uncertainties!

Reporting Results with Uncertainties

•Experimental Results are always reported with uncertainties, both statistical and systematic (if known)

•The uncertainty is assumed to be in **the last reported digit of the result (very important to remember this!)**

- $x \pm \sigma_{stat}$ OR $x(\sigma_{stat})$

  e.g. 10 ± 1 cm or 10(1) cm

  10.1 ± 0.1 cm or 10.1(1) cm - look for precision!

  Not: 10.1(0.1) cm !

  10.11 ± 1.20 cm or 10.11(120) cm

  Not 10.11(1.20) cm

  10.110 ± 1.20 cm or 10.11 ± 0.111 cm - **BOTH**

WRONG - under/over-precision

- $x \pm \sigma_{stat} \pm \sigma_{syst}$ OR $x(\sigma_{stat})(\sigma_{syst})$

  e.g. $10 \pm 1_{stat} \pm 2_{syst}$ cm or 10(1)(2) cm

Reporting Results with Uncertainties

Example: Mass of top quark

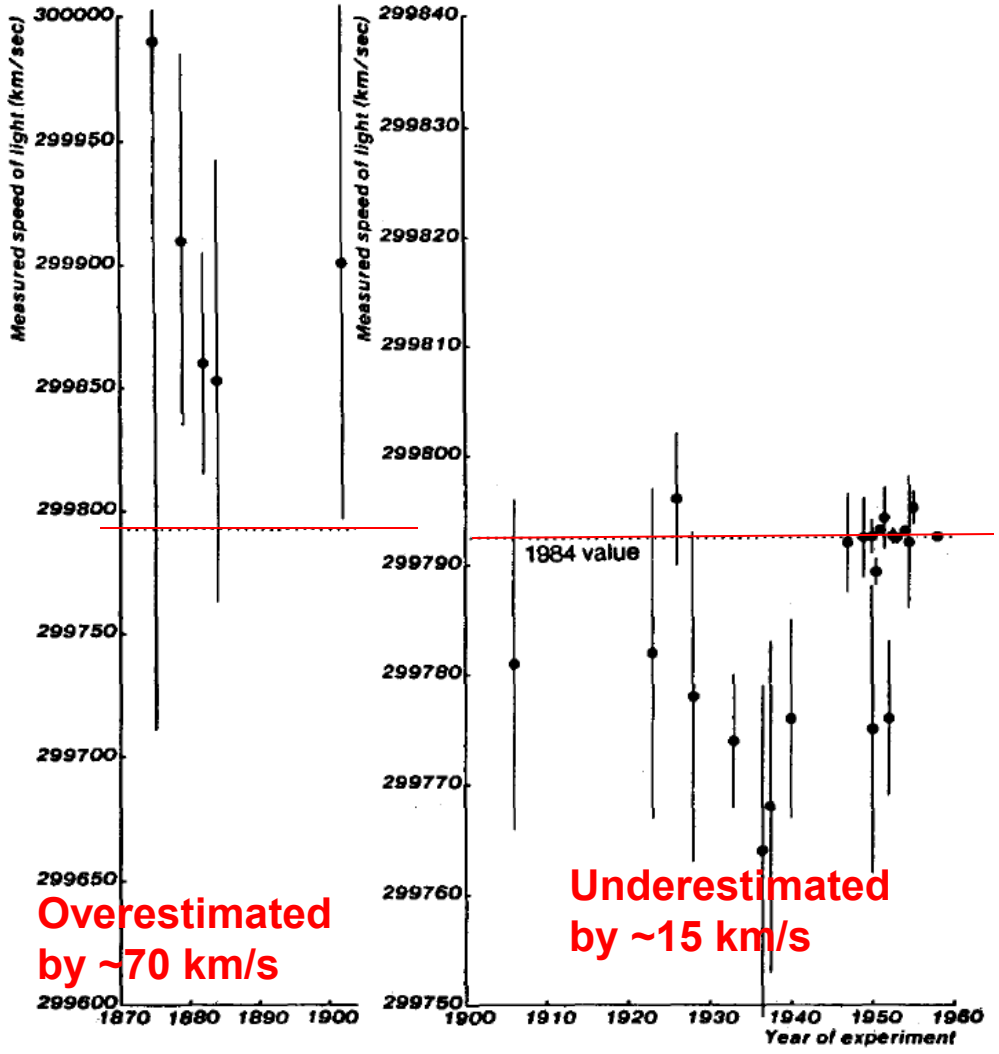
http://www.science20.com/quantum_diaries_survivor/who_underestimates_their_systematic_uncertainties

$170.7^{+4.2}_{-3.9} \pm 3.5$		[14,15]	AALTONEN	08C	CDF	dilepton, $\sigma_{t\bar{t}}$ constrained
$177.1 \pm 4.9 \pm 4.7$		[16,17]	AALTONEN	07	CDF	6 jets with ≥ 1 b vtx
$172.3^{+10.8}_{-9.6} \pm 10.8$		[18]	AALTONEN	07B	CDF	≥ 4 jets (b-tag)
$173.7 \pm 4.4^{+2.1}_{-2.0}$		[17,19]	ABAZOV	07F	D0	lepton + jets
$170.3^{+4.1}_{-4.5} \pm 1.2 \pm 1.8$		[4,20]	ABAZOV	06U	D0	lepton + jets (b-tag)
$173.2^{+2.6}_{-2.4} \pm 3.2$		[21,22]	ABULENCIA	06D	CDF	lepton + jets
$173.5^{+3.7}_{-3.6} \pm 1.3$		[15,21]	ABULENCIA	06D	CDF	lepton + jets
$165.2 \pm 6.1 \pm 3.4$		[4,23]	ABULENCIA	06G	CDF	dilepton
$170.1 \pm 6.0 \pm 4.1$		[15,24]	ABULENCIA	06V	CDF	dilepton
$178.5 \pm 13.7 \pm 7.7$		[25,26]	ABAZOV	05	D0	6 or more jets

stat syst

Bias?: Measuring Velocity of Light

1875-1958



1929-1973

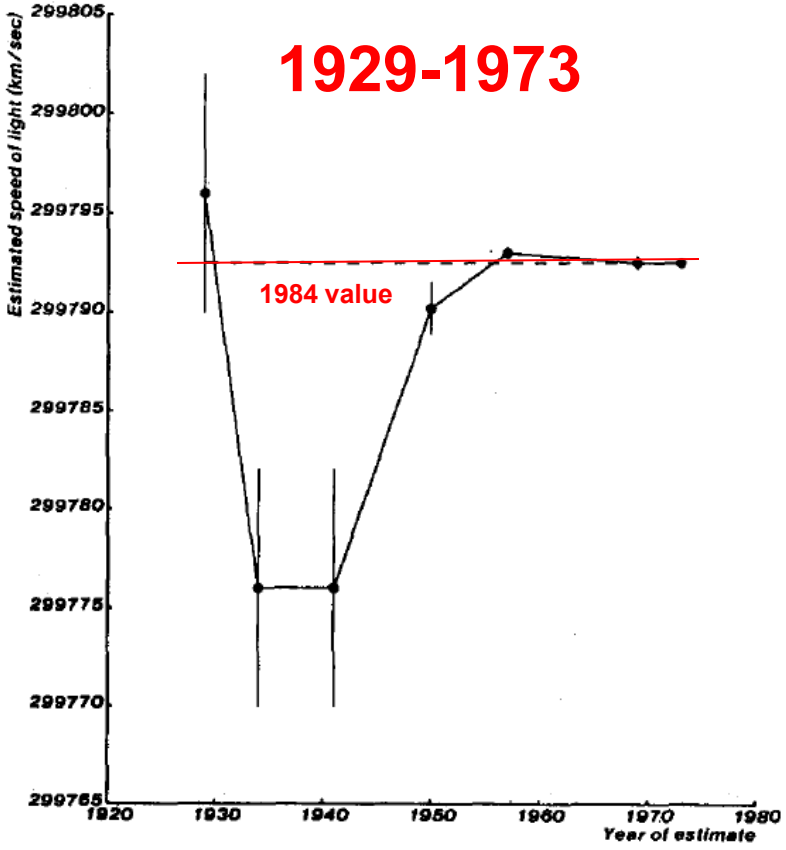
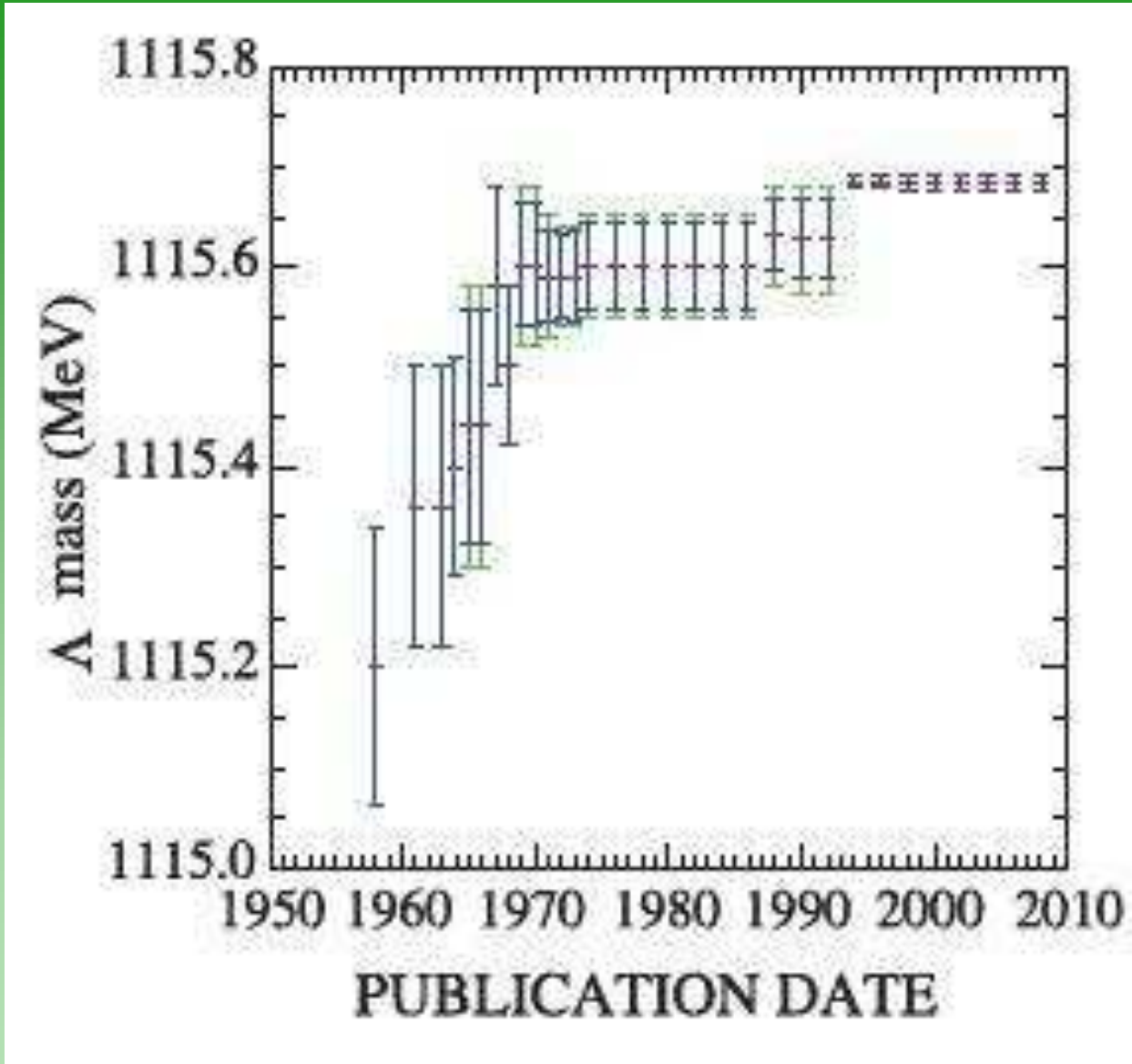


Fig. 2. Recommended values for the velocity of light; 1929–1973.

Bias: Measuring Mass of Lambda



Probability Distributions

- **Binomial Distribution**
- **Poisson Distribution**
- **Gaussian Distribution**
- **Gaussian distribution as a limit of the Binomial/ Poisson distributions**
- **Simple Counting Experiment**
- **Central Limit Theorem**

The Binomial Distribution

- Extremely general statistical model
- Applicable to describe the **frequency distribution** of a set of **many repeated measurements** of **constant likelihood "p"** (e.g. probability to get one of the values when throwing a dice is p=1/6)
- Discrete distribution of a binary process (**True** or **False**; **Yes** or **No**)
- Unwieldy when the numbers involved become large (can simplify in certain limits)
- It is frequently used to model number of successes in a sample of size ***n*** from a population of size ***N***. For *N* much larger than *n*, the binomial distribution is a good approximation.

The Binomial Distribution

Assume that we make a number of trials n

Each trial has a constant likelihood of success p

(therefore, the chance of failure is $1-p$)

Then the predicted probability of x 'successes' is

$$P(x) = \frac{n!}{(n-x)!\,x!}\,p^{x}(1-p)^{n-x}$$

for integer values of n and x

Mean $\bar{x} = pn$

Variance: $\sigma^2 = np(1-p)$

The Binomial Distribution

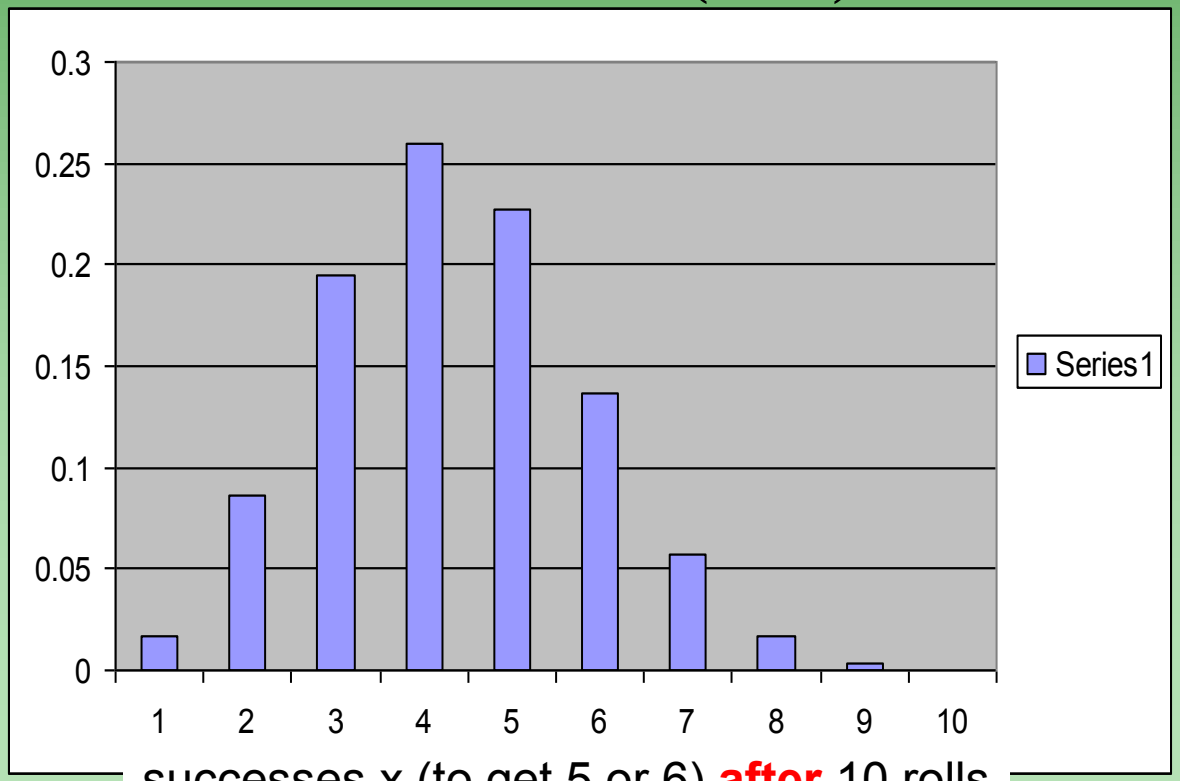
- Assume we role a die
- We call a "success" if we get a 5 or 6 – $p = {}^1/_3$
- We make 10 rolls (10 *trials*) $n = 10$

$$P(x) = \frac{n!}{(n-x)!\,x!}\,p^x(1-p)^{n-x}$$

$$\sum_{x=0}^{n} P(x) = 1$$

$$\bar{x} = \sum_{x=0}^{n} xP(x)$$

$$\bar{x} = pn = 3.3333$$



The Binomial Distribution

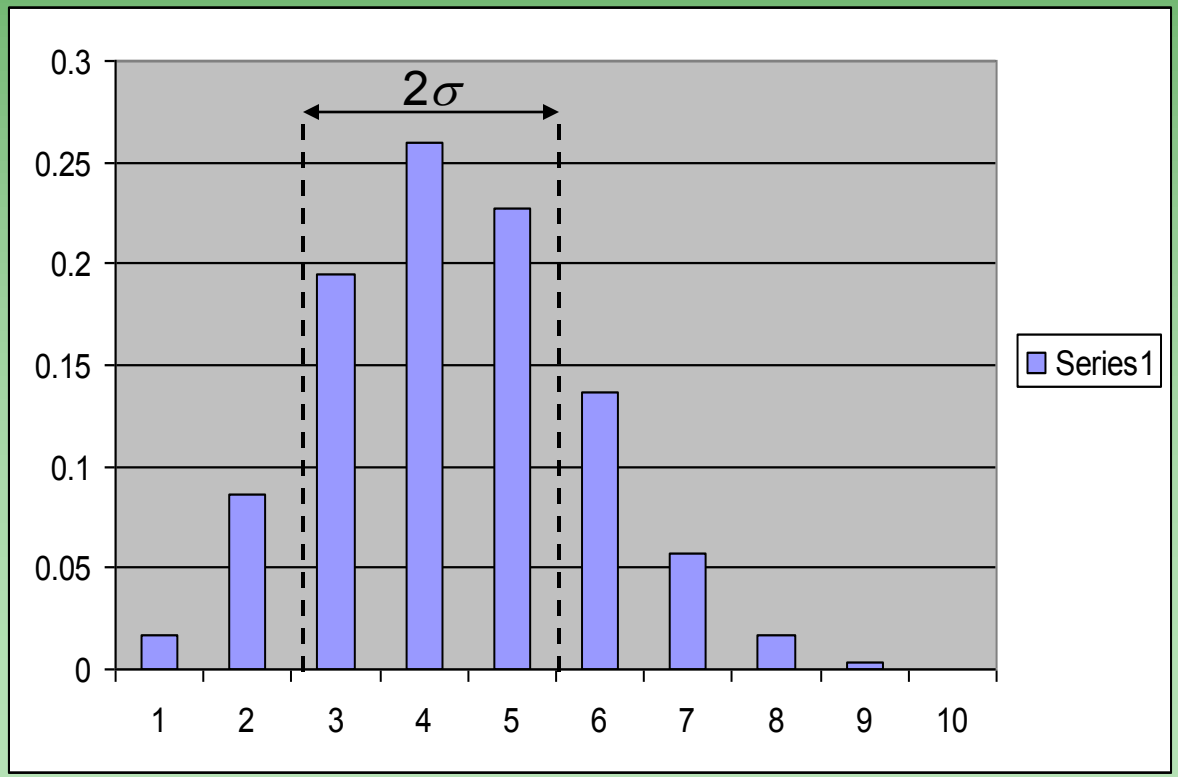
Width is given by

$$\sigma^2 = \sum_{x=0}^{n} (x - \bar{x})^2 P(x) \qquad \sigma = 1.49$$

Or alternatively

$$\sigma^2 = np(1-p)$$

The width σ gives us an idea of the typical deviation from the true mean of any single measurement



The Binomial Distribution

For large number of trials, numbers soon blow up to be unrealistically manageable

$$P(x) = \frac{n!}{(n-x)!\,x!}\, p^{x}(1-p)^{n-x}$$

$10! = 3.6 \times 10^{6}$

$20! = 2.8 \times 10^{18}$

$50! = 3 \times 10^{64}$

$100! = 9 \times 10^{157}$

Imagine if we were dealing with a radioactive sample (very large number of nuclei!) – impossible to handle with Binominal distribution!

The Poisson Distribution

In the limit where the **number of trials is large** and **the success probability is small**, the binomial distribution reduces to the Poisson form

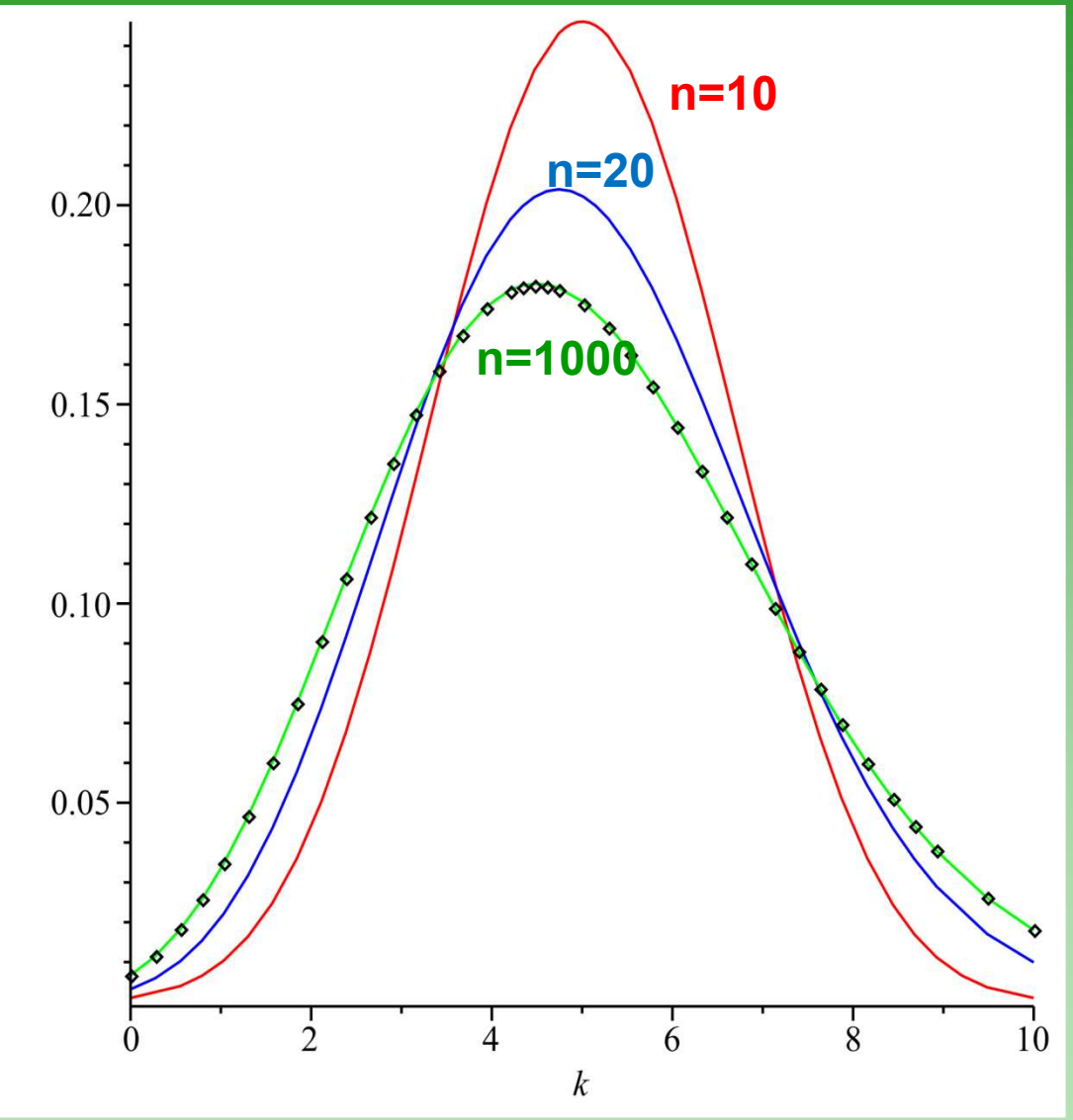
$$P(x) = \frac{(pn)^x e^{-pn}}{x!}$$

Which, as $\bar{x} = pn$, is equivalent to

$$P(x) = \frac{\bar{x}^x e^{-\bar{x}}}{x!}$$

Again, for integer values of n and x

The Binominal and Poisson Distribution



Comparison of the **Poisson distribution (black dots)** and the binomial distribution with **n=10 (red line), n=20 (blue line), n=1000 (green line).** All distributions have a mean of 5. The x-axis shows the number of events *k*. Notice that as n gets larger, the Poisson distribution becomes an increasingly better approximation for the binomial distribution with the same mean.

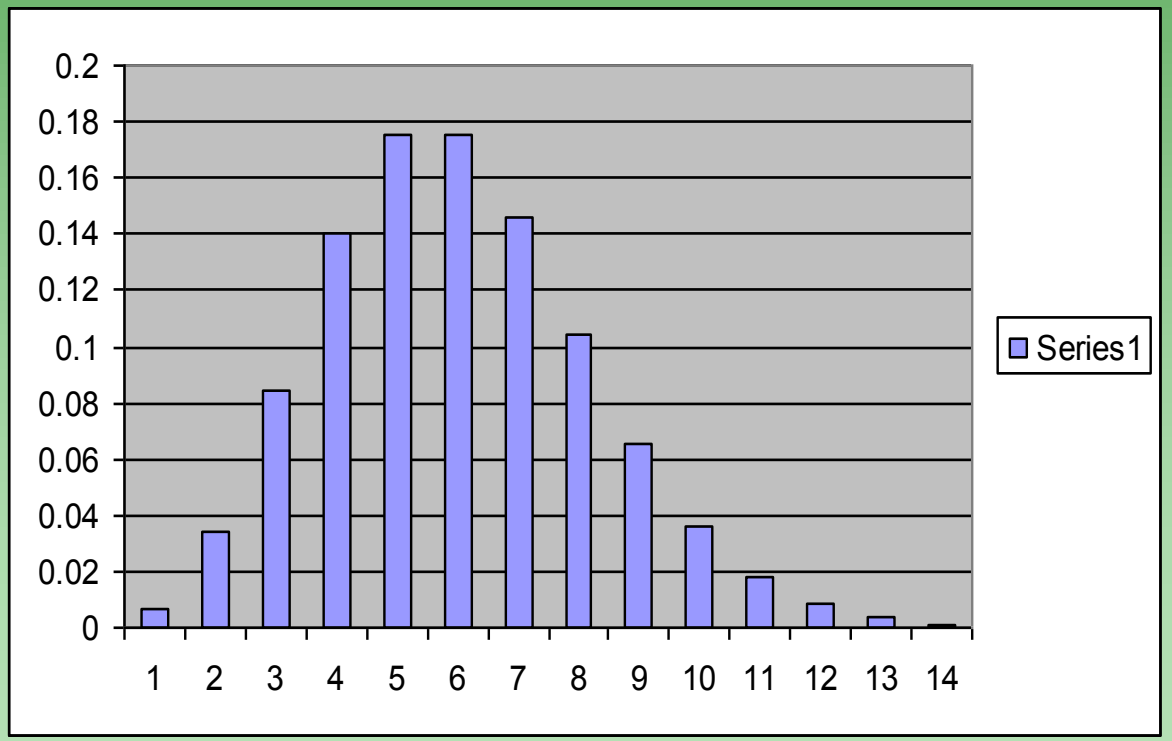
The Poisson Distribution

Assume we now have a process with n = 1000 trials and our success probability is p = 0.005

$$\sum_{x=0}^{n} P(x) = 1$$

$$\bar{x} = \sum_{x=0}^{n} xP(x)$$

$$\bar{x} = pn = 5$$



The Poisson Distribution

However, now

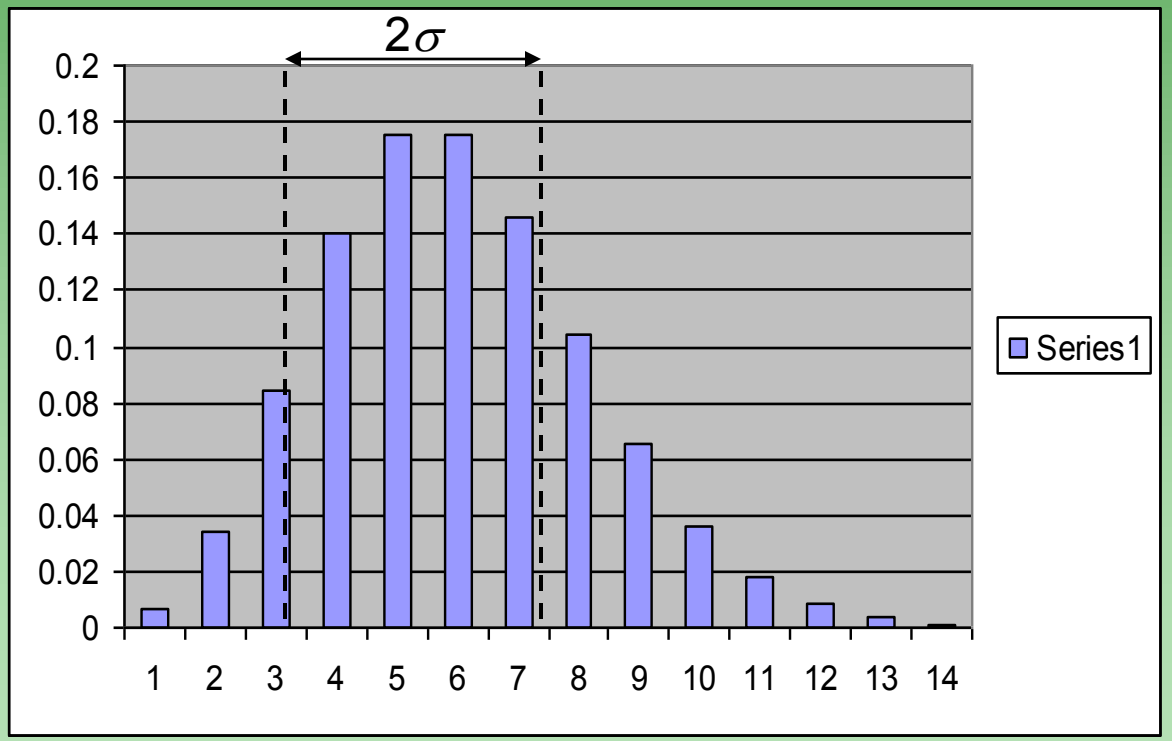
$$\sigma^2 = \sum_{x=0}^{n} (x - \bar{x})^2 P(x) = pn$$

And, as

$$\bar{x} = pn$$

$$\sigma = \sqrt{\bar{x}}$$

Very important result!



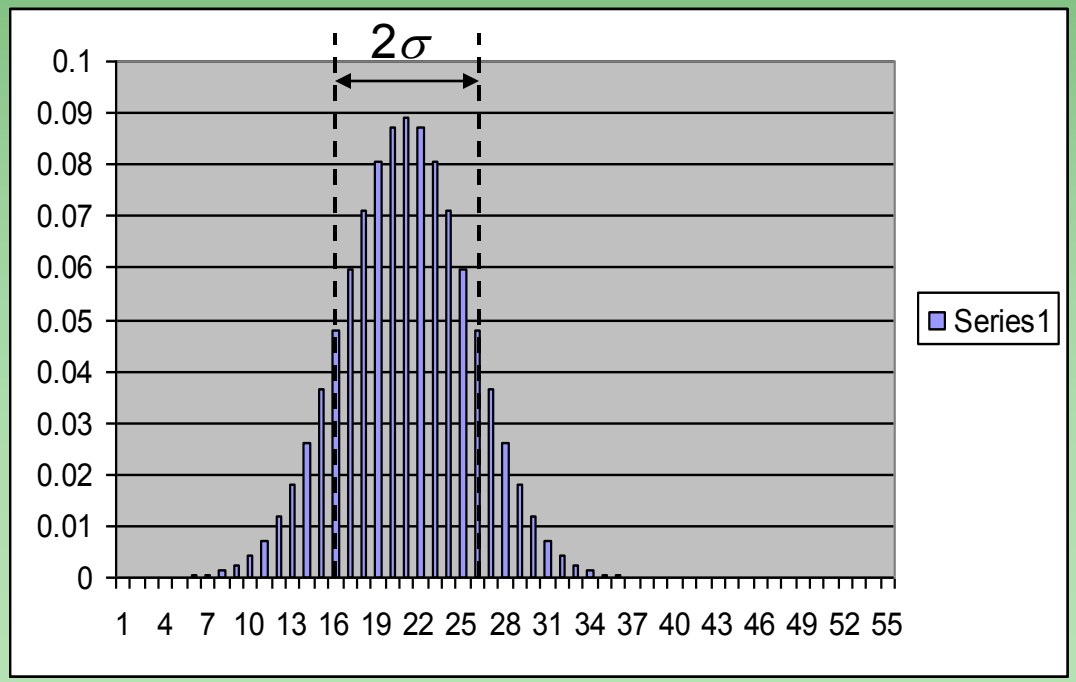
The Gaussian Distribution

In the limit of very large n, the Poisson distribution simplifies to the Gaussian/Normal distribution:

$$P(x) = \frac{1}{\sqrt{2\pi\bar{x}}} e^{\frac{-(x-\bar{x})^2}{2\bar{x}}}$$

$$\sum_{x=0}^{n} P(x) = 1$$

$$\sigma = \sqrt{\bar{x}}$$



The Gaussian Distribution

However, we now have an expression which we can treat as continuous:

$$P(x) = \frac{1}{\sqrt{2\pi\bar{x}}} e^{\frac{-(x-\bar{x})^2}{2\bar{x}}}$$

And that we can generalise

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

μ = centroid

σ = standard deviation

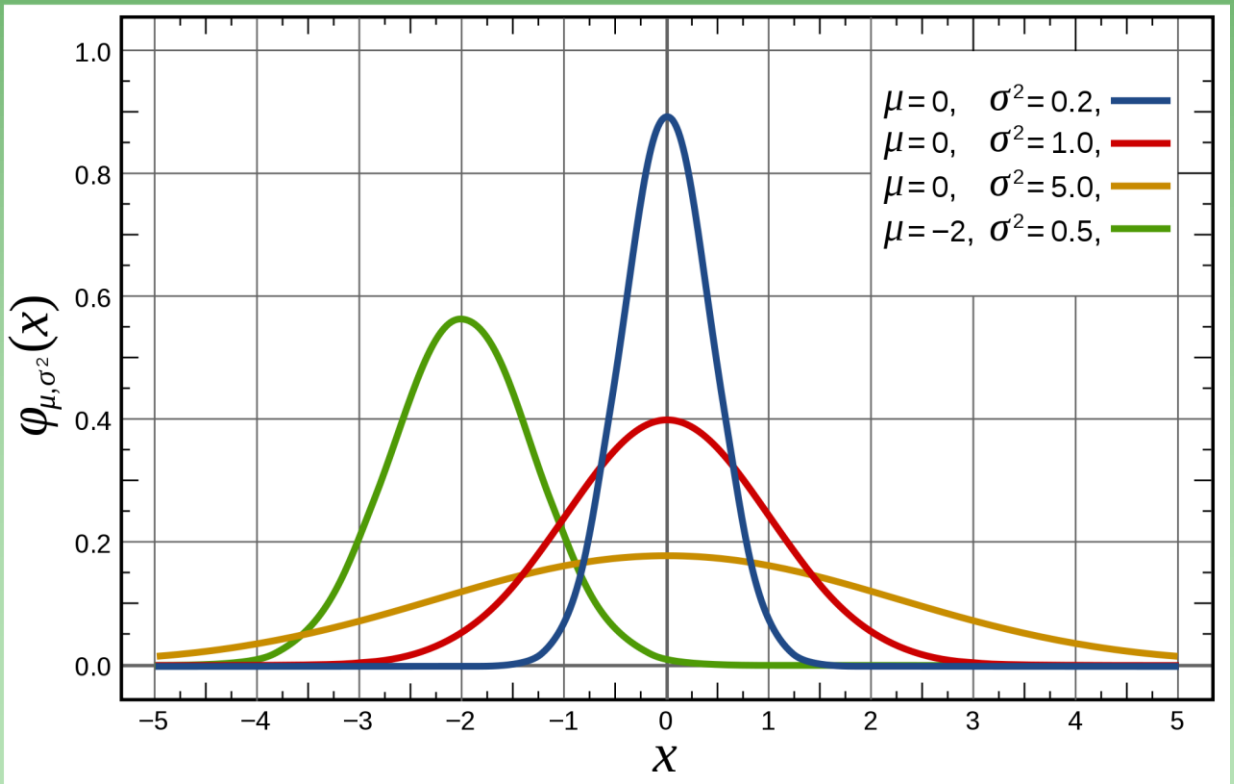
$$\int_{-\infty}^{\infty} P(x;\mu,\sigma)dx = 1$$

The Gaussian Distribution

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

μ = centroid

σ = standard deviation



The Gaussian Distribution

- Full parameterisation of width and centroid
- Extremely versatile probability function
- Suitable of for describing more complex systems, with sums of many distributions
- Important for describing distributions resulting from most sources of statistical fluctuations

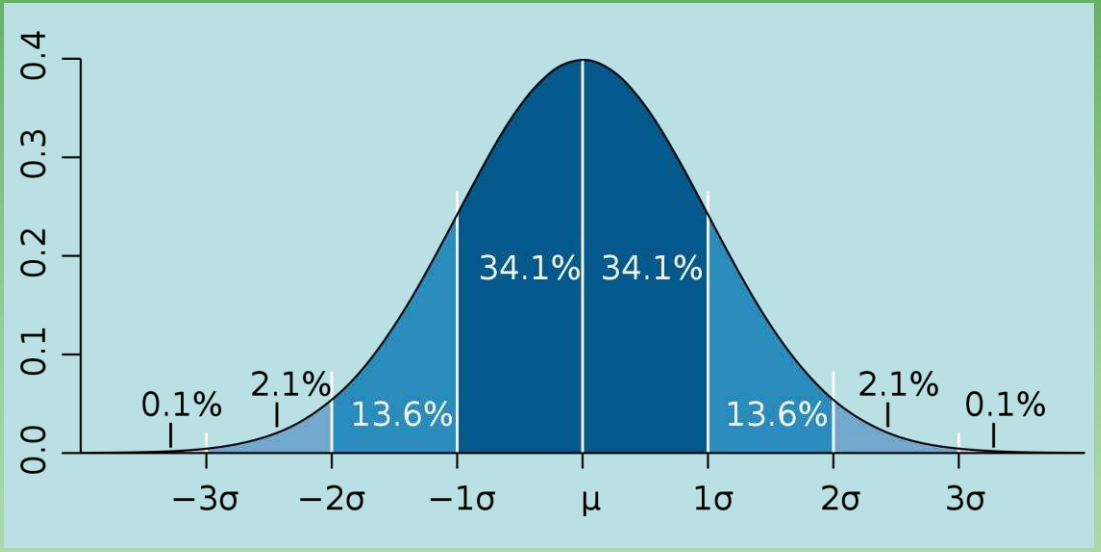
The Gaussian Distribution

Integrals within limits:

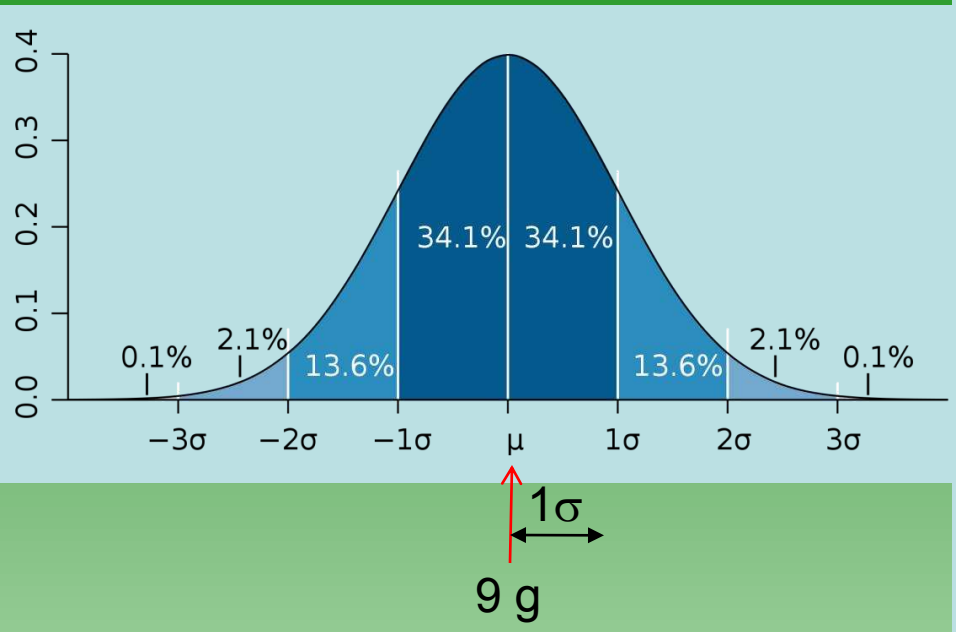
$$\int_{\mu-\sigma}^{\mu+\sigma} P(x;\mu,\sigma)dx = 0.6827$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} P(x;\mu,\sigma)dx = 0.9545$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} P(x;\mu,\sigma)dx = 0.9973$$



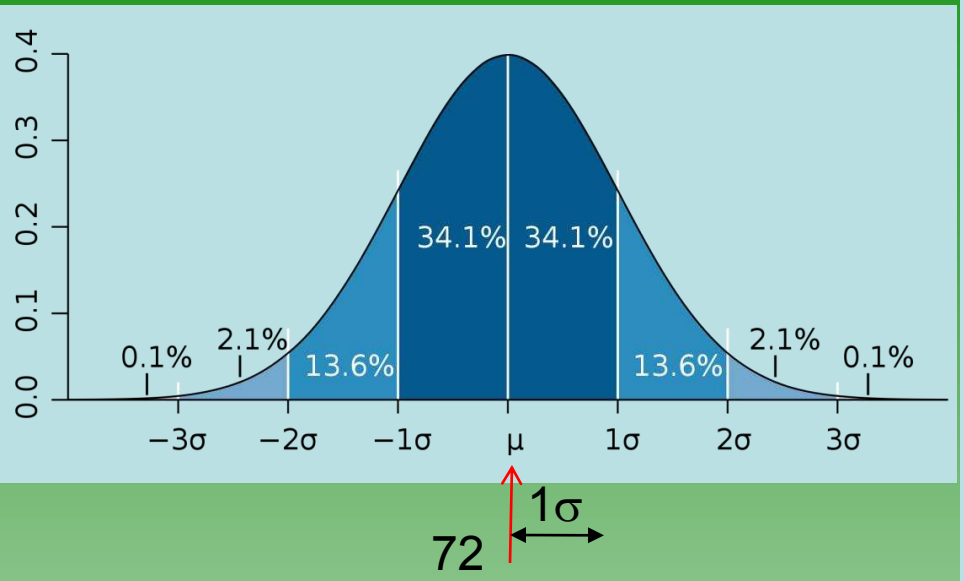
Examples: The Gaussian Distribution



The amount of mustard dispensed from a machine at *Mac Donalds* is ***normally distributed*** **with a mean of 9 g and a standard deviation of 1 g.** If the machine is used 500 times, *approximately* how many times will it be expected to dispense 10 or more grams of mustard?

•The mean is 9 g and the standard deviation is 1g. If one standard deviation is added to the mean, the result is 10 g. Therefore, **dispensing 10 or more grams falls into the category above one standard deviation to the right of the mean**. Reading from the bell curve chart above, 15.8% of data falls at or above 1 standard deviation.15.8% x 500 ~ 80 times to dispense 10 or more grams of mustard.

Examples: The Gaussian Distribution



There are 184 students in the lecture class. The scores on the midterm exam are *normally distributed* with a mean of 72 and a standard deviation of 8. How many students in the class can be expected to receive a score between 80 and 88?

Mean plus one standard deviation = 80, mean plus two standard deviation = 88

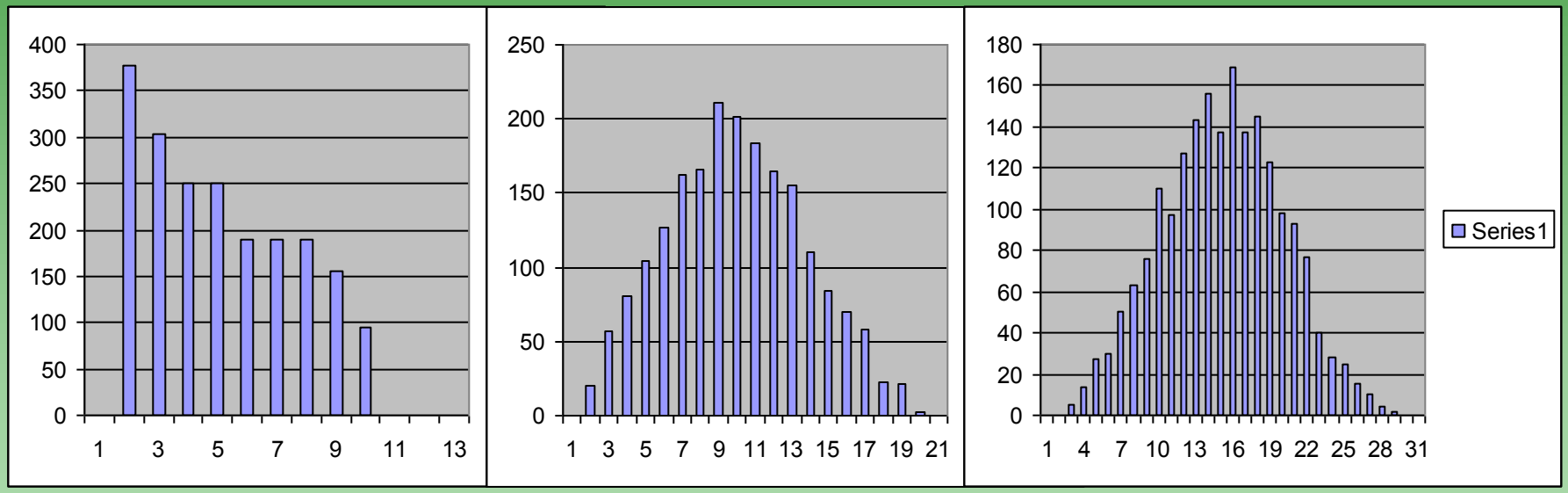
We need the range between 1 and 2 standard deviations.

As read from the above chart, the amount of the distribution between 1 and 2 standard deviations above the mean is 13.6%.

13.6% of 184 students ~ 25 students

Central Limit Theorem

True for ***any*** original distribution form or combination thereof (providing there are enough, and they are uncorrelated)



Skewed distribution

One skewed + one flat distribution

One skewed + two flat distributions

Gaussian extremely versatile for describing generic statistical uncertainties

Central Limit Theorem

Measurements have typically many sources of variation (uncertainty/error)

If these variations are independent (uncorrelated), the Central Limit Theorem states, if:

$$X = x_1 + x_2 + x_3 \ldots + x_N$$

where each distribution has a mean μ_i and variance σ_i then:

$$\langle X \rangle = \sum \mu_i$$

$$V(X) = \sum \sigma_i^2$$

and tends to Gaussian form as N becomes large.

Functions of dependent variables

Suppose that the data are expressed in two variables x and y

To what extent are the two dependent on each other?

$$\mathrm{cov}(x, y) = \frac{1}{N}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$\mathrm{cov}(x, y) = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

If large x tends to occur with large y, then $\mathrm{cov}(x, y) = +\mathrm{ve}$

If large x tends to occur with small y, then $\mathrm{cov}(x, y) = -\mathrm{ve}$

If x and y are uncorrelated, then $\mathrm{cov}(x, y) = 0$

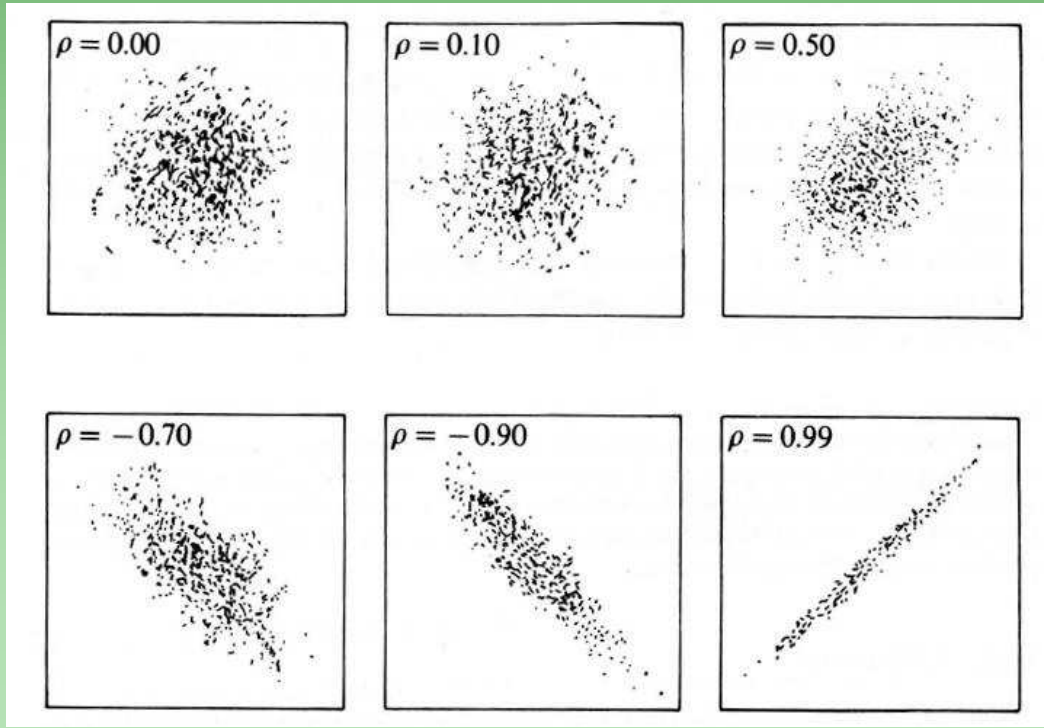
However, the magnitude of the covariance is dependent on the data set, and has dimensions

Functions of dependent variables

Sometimes, a more useful parameter is the correlation coefficient ρ

$$\rho = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} \qquad \rho = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\sigma_x \sigma_y}$$

ρ is between -1 and +1



Roger Barlow

Statistics

John Wiley & Sons

Functions of dependent variables

For a system of three variables, x, y and z we can define the covariance between each pair of variables

$$V_{xy} = \mathrm{cov}(x, y) = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

$$V_{xz} = \mathrm{cov}(x, y) = \langle xz \rangle - \langle x \rangle \langle z \rangle \qquad \ldots\text{etc}$$

These are elements of a *covariance matrix* $\mathbf{V}$ – a symmetric matrix, with the diagonal elements being the variances of each parameter

Equivalently, the elements of a *correlation matrix* can be defined as

$$\rho_{xy} = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\sigma_x \sigma_y} \qquad \rho_{xz} = \frac{\langle xz \rangle - \langle x \rangle \langle z \rangle}{\sigma_x \sigma_z}$$

UWS UNIVERSITY OF THE WEST of SCOTLAND

Introductory Data Analysis

Lecture 4

Gaussian Distribution, Central Limit Theorem, Propagation of uncertainties, Estimator

SUPA Graduate School 2010

Prof. Andrei Andreyev

Andrei.Andreyev@uws.ac.uk

Nuclear Physics Research Group

School of Engineering and Science

Lecture 4 outline

- Gaussian Distribution
- Central Limit theorem
- Propagation of uncertainties
- Estimator
- Maximum Likelihood method

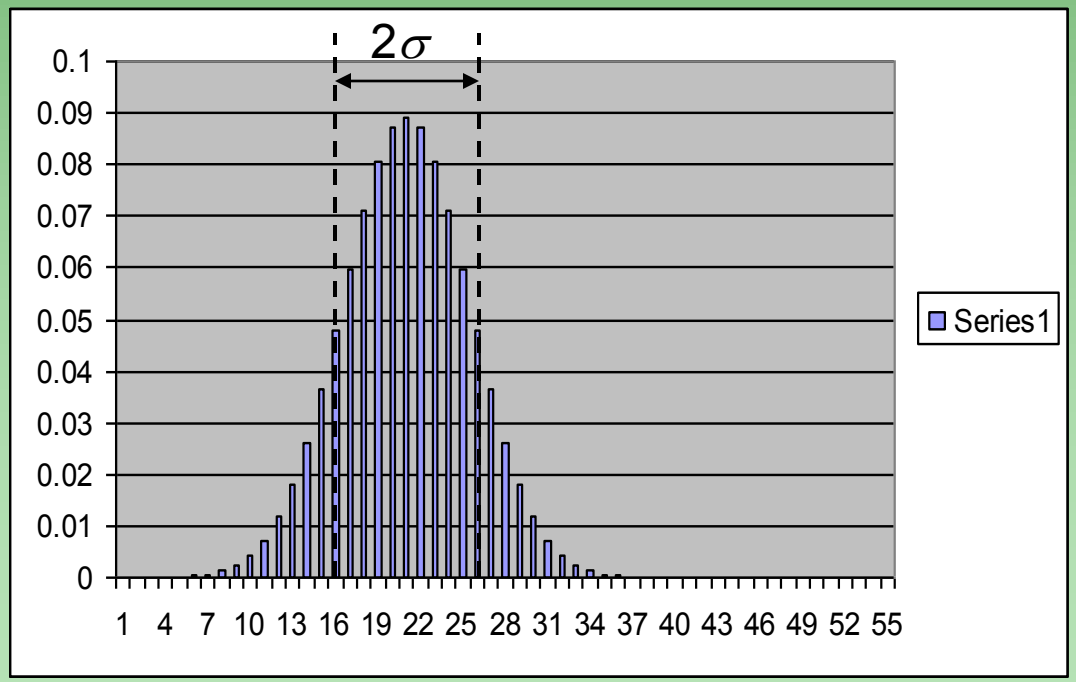
The Gaussian Distribution

In the limit of very large n, the Poisson distribution simplifies to the Gaussian/Normal distribution:

$$P(x) = \frac{1}{\sqrt{2\pi\bar{x}}} e^{\frac{-(x-\bar{x})^2}{2\bar{x}}}$$

$$\sum_{x=0}^{n} P(x) = 1$$

$$\sigma = \sqrt{\bar{x}}$$



The Gaussian Distribution

However, we now have an expression which we can treat as continuous:

$$P(x) = \frac{1}{\sqrt{2\pi\bar{x}}} e^{\frac{-(x-\bar{x})^2}{2\bar{x}}}$$

And that we can generalise

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

μ = centroid

σ = standard deviation

$$\int_{-\infty}^{\infty} P(x;\mu,\sigma)dx = 1$$

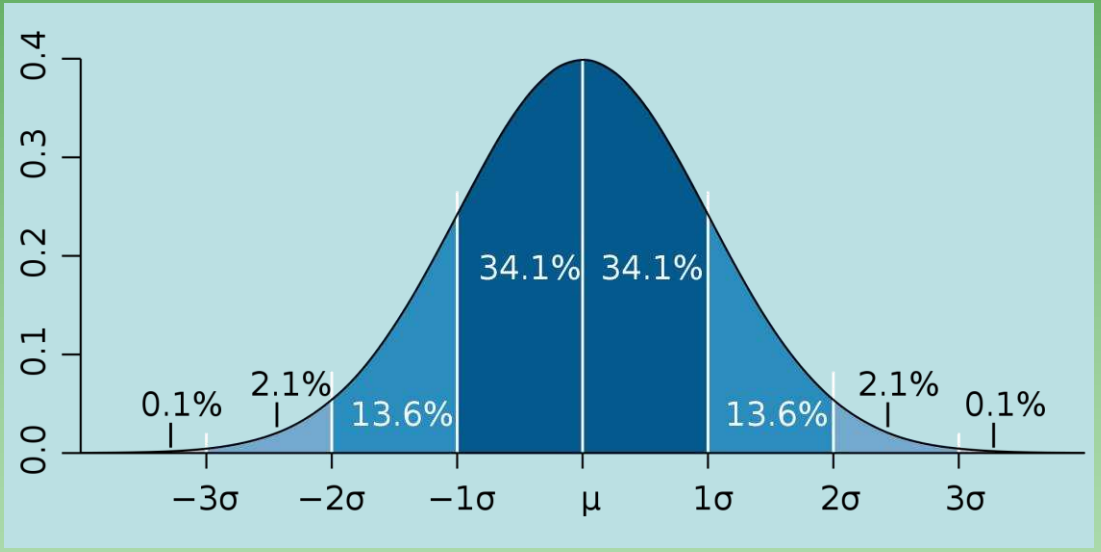
The Gaussian Distribution

Integrals within limits:

$$\int_{\mu-\sigma}^{\mu+\sigma} P(x;\mu,\sigma)dx = 0.6827$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} P(x;\mu,\sigma)dx = 0.9545$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} P(x;\mu,\sigma)dx = 0.9973$$



Central Limit Theorem

Measurements have typically many sources of variation (uncertainty/error)

If these variations are independent (**uncorrelated**), the **Central Limit Theorem** states, if:

$$X = x_1 + x_2 + x_3 \ldots + x_N$$

where each distribution x_i has a mean μ_i and variance σ_i then:

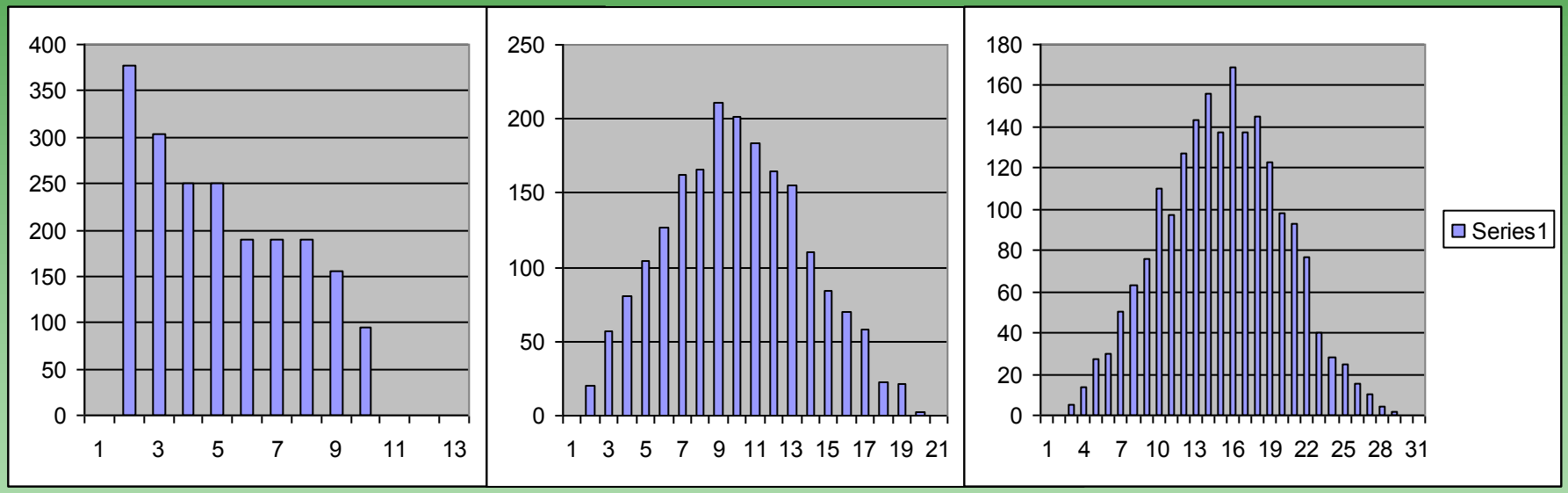
$$\langle X \rangle = \sum \mu_i$$

$$V(X) = \sum \sigma_i^2$$

and tends to Gaussian form as N becomes large.

Central Limit Theorem

True for ***any*** original distribution form or combination thereof (providing there are enough, and they are uncorrelated!)



Skewed distribution

One skewed + one flat distribution

One skewed + two flat distributions

Gaussian extremely versatile for describing generic statistical uncertainties

Functions of dependent variables

Suppose that the data are expressed in two variables x and y

To what extent are the two dependent on each other?

$$\mathrm{cov}(x, y) = \frac{1}{N}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$\mathrm{cov}(x, y) = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

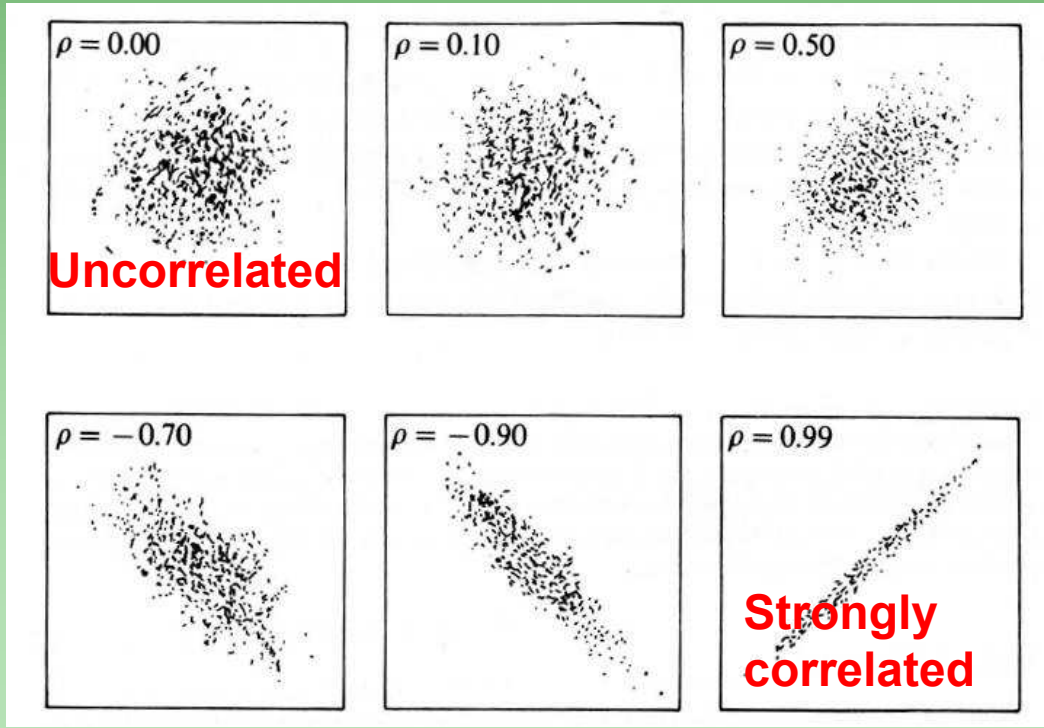
If x and y are uncorrelated, then $\mathrm{cov}(x, y) = 0$

Functions of dependent variables

Sometimes, a more useful parameter is the correlation coefficient ρ

$$\rho = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} \qquad \rho = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\sigma_x \sigma_y}$$

ρ is between -1 and +1



Roger Barlow

Statistics

John Wiley & Sons

Functions of dependent variables

For a system of three variables, x, y and z we can define the covariance between each pair of variables

$$V_{xy} = \mathrm{cov}(x, y) = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

$$V_{xz} = \mathrm{cov}(x, y) = \langle xz \rangle - \langle x \rangle \langle z \rangle \qquad \ldots\text{etc}$$

These are elements of a ***covariance matrix*** $\mathbf{V}$ – a symmetric matrix, with the diagonal elements being the variances of each parameter

Equivalently, the elements of a *correlation matrix* can be defined as

$$\rho_{xy} = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\sigma_x \sigma_y} \qquad \rho_{xz} = \frac{\langle xz \rangle - \langle x \rangle \langle z \rangle}{\sigma_x \sigma_z}$$

Propagation of Uncertainties

Now, consider the more general case where function f is some generic function of x

For ***small*** differences, expand in a Taylor series

(***small*** *means the differential changes little over a few* σ)

$$f(x) = f(x_0) + (x - x_0)\left(\frac{df}{dx}\right)\Big|_{x_0}$$

Going through simple algebra and calculate variance

$$V(f) = \langle f^2 \rangle - \langle f \rangle^2$$

$$V(f) = \left(\frac{df}{dx}\right)^2 V(x)$$

$$\sigma_f = \left|\frac{df}{dx}\right| \sigma_x$$

Specific Uncertainty Rules

Multiplication or division by a constant

$$f = Ax \qquad f = \frac{x}{A}$$

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 \qquad \frac{\partial f}{\partial x} = A \qquad \sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 \qquad \frac{\partial f}{\partial x} = \frac{1}{A}$$

$$\sigma_f^2 = A^2\sigma_x^2 \qquad \sigma_f^2 = \frac{1}{A^2}\sigma_x^2$$

$$\sigma_f = A\sigma_x \qquad \sigma_f = \frac{\sigma_x}{A}$$

Again, an intuitive result – the fractional uncertainty remains constant

Specific Uncertainty Rules

Sums and differences

$$f = x + y \quad \text{or} \quad f = x - y$$

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

$$\frac{\partial f}{\partial x} = 1 \qquad \frac{\partial f}{\partial y} = \pm 1$$

$$\sigma_f^2 = (1)^2 \sigma_x^2 + (\pm 1)^2 \sigma_y^2$$

$$\sigma_f = \sqrt{\sigma_x^2 + \sigma_y^2}$$

Specific Uncertainty Rules

Multiplication and division

$$f = xy \quad \text{or} \quad f = \frac{x}{y} \quad \text{or} \quad f = \frac{y}{x}$$

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 \qquad \frac{\partial f}{\partial x} = y \qquad \frac{\partial f}{\partial y} = x$$

$$\sigma_f^2 = y^2\sigma_x^2 + x^2\sigma_y^2 \qquad \text{dividing by} \qquad f^2 = x^2y^2$$

$$\frac{\sigma_f^2}{f^2} = \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2}$$

$$\frac{\sigma_f}{f} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2}$$

Specific Uncertainty Rules

Mean of multiple series of independent measurements

$$X = x_1 + x_2 + x_3 \ldots + x_N$$

Total variance (width) is

$$\sigma_X^2 = \sigma_{x_i}^2 + \sigma_{x_2}^2 + \sigma_{x_3}^2 + \ldots + \sigma_{x_N}^2 \quad \text{and as} \quad \sigma_{x_i} = \sqrt{x_i}$$

$$\sigma_X^2 = x_1 + x_2 + x_3 + \ldots + x_N = X \quad \text{giving} \quad \sigma_X = \sqrt{X}$$

and as $\bar{x} = \dfrac{X}{N}$

$$\sigma_{\bar{x}} = \frac{\sigma_X}{N} = \frac{\sqrt{X}}{N} = \frac{\sqrt{N\bar{x}}}{N}$$

$$\sigma_{\bar{x}} = \sqrt{\frac{\bar{x}}{N}}$$

Meaning the precision of a a series of measurements increases with $\sqrt{N}$

(need 4 measurements to double the precision)

Specific Uncertainty Rules

Combinations of independent measurements with unequal uncertainties

$$X = x_1 + x_2 + x_3 \ldots + x_N$$

The mean is

$$\langle x \rangle = \frac{\sum_{i=1}^{N} a_i x_i}{\sum_{i=1}^{N} a_i} \qquad \text{where} \quad a_i = \frac{1}{\sigma_{x_i}^2} \left(\sum_{i=1}^{N} \frac{1}{\sigma_{x_i}^2} \right)^{-1}$$

$$\langle x \rangle = \frac{\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

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

Meaning the individual measurements have weightings inversely proportional to their variance

Estimation

Imagine, we have a data set $\{x_1 + x_2 + x_3 \ldots + x_N\}$

Consider that we are trying to extract a variable a (further - true value) from a set of data. We call our estimator $\hat{a}$

- The concept of estimation in statistics is not synonymous with the common definition - it is not approximate procedure!
- Rather, it is a precise and accurate procedure and, though the result may be imprecise, the extent of the imprecision is quantified
- *An estimator is a procedure (when applied to a data sample) which gives a numerical value for a property of the parent population, or a property or parameter of the parent distribution function*
- An good estimator has the following properties:
- *Consistent, unbiased and efficient*

Estimation

Consider that we are trying to extract a variable a (further - true value) from a set of data. We call our estimator $\hat{a}$

Consistency

- $\hat{a}$ tends to the true value with infinite measurements

$$\lim_{N\to\infty} \hat{a} = a$$

Bias

- The **expectation value** of $\hat{a}$ is equal to **the true value**

$$\langle \hat{a} \rangle = a$$

Efficiency

- The variance of $\hat{a}$ is small

Estimation

An example from Barlow

Suppose we are trying to estimate the average height of students at a university. We take an unbiased sample from the total student population N. There are several estimators that one could conceive, of varying quality, such as:

1) Sum the sample heights and divide by N
2) Sum the first 10 heights and divide by 10 (ignoring the rest of the sample)
3) Sum the heights and divide by $N - 1$
4) Ignore the data entirely, and adopt the value 1.8 m
5) Multiply all the heights and take the N^{th} root
6) Adopt the most popular height (the mode)
7) Sum the tallest and shortest heights, and divide by 2
8) Sum the second, fourth, sixth… etc, heights and divide by $N/2$

Estimation

Consistency

$\hat{a}$ tends to the true value with infinite measurements $\lim_{N\to\infty} \hat{a} = a$

For case 1, $$\hat{\mu} = \frac{x_1 + x_2 + x_3 \ldots + x_N}{N} = \bar{x}$$

$$\bar{x} \to \mu \quad \text{when } N \to \infty$$

Clearly, This is also true for case 3

$$(N-1) \to N \quad \text{when } N \to \infty$$

and for case 8

Estimation

An example from Barlow

Suppose we are trying to estimate the average height of students at a university. We take an unbiased sample from the total student population N. There are several estimators that one could conceive, of varying quality, such as:

1) Sum the sample heights and divide by N

~~2) Sum the first 10 heights and divide by 10 (ignoring the rest of the sample)~~

3) Sum the heights and divide by $N - 1$

~~4) Ignore the data entirely, and adopt the value 1.8 m~~

~~5) Multiply all the heights and take the N^{th} root~~

~~6) Adopt the most popular height (the mode)~~

~~7) Sum the tallest and shortest heights, and divide by 2~~

8) Sum the second, fourth, sixth… etc, heights and divide by $N/2$

Excluded based on "Consistency"

Estimation

Bias

- The expectation value of $\hat{a}$ is equal to the true value $\langle \hat{a} \rangle = a$

For case 1,

$$\langle \hat{\mu} \rangle = \left\langle \frac{x_1 + x_2 + x_3 \ldots + x_N}{N} \right\rangle \qquad \langle \hat{\mu} \rangle = \frac{N\langle x \rangle}{N} = \mu$$

But for case 3 $$\langle \hat{\mu} \rangle = \frac{N}{N-1}\mu \neq \mu$$

Must be excluded as well

Estimation

An example from Barlow

Suppose we are trying to estimate the average height of students at a university. We take an unbiased sample from the total student population N. There are several estimators that one could conceive, of varying quality, such as:

1) Sum the sample heights and divide by N

~~2) Sum the first 10 heights and divide by 10 (ignoring the rest of the sample)~~

~~3) Sum the heights and divide by $N - 1$~~

Consistency

~~4) Ignore the data entirely, and adopt the value 1.8 m~~

Bias

~~5) Multiply all the heights and take the N^{th} root~~

~~6) Adopt the most popular height (the mode)~~

~~7) Sum the tallest and shortest heights, and divide by 2~~

8) Sum the second, fourth, sixth… etc, heights and divide by $N/2$

Estimation

An example from Barlow

Suppose we are trying to estimate the average height of students at a university. We take an unbiased sample from the total student population N. There are several estimators that one could conceive, of varying quality, such as:

1) Sum the sample heights and divide by N

~~2) Sum the first 10 heights and divide by 10 (ignoring the rest of the sample)~~

~~3) Sum the heights and divide by $N - 1$~~

~~4) Ignore the data entirely, and adopt the value 1.8 m~~

~~5) Multiply all the heights and take the N^{th} root~~

~~6) Adopt the most popular height (the mode)~~

~~7) Sum the tallest and shortest heights, and divide by 2~~

~~8) Sum the second, fourth, sixth… etc, heights and divide by $N/2$~~

Consistency

Bias

Efficiency- (The variance of is small)

Likelihood Function

When using estimators, one applies estimator $\hat{a}$ to a data sample $\{x_1 + x_2 + x_3 \ldots + x_N\}$ in order to estimate a

But in practice, to consider the properties of an estimator, the **reverse approach** is used. That is, **start with a known distribution** from which the x_i are drawn

$$P(x;a)$$

The probability of a particular data set is the product of the individual probabilities, and is called the *likelihood*

$$L(x_1 + x_2 + x_3 \ldots + x_N; a) = P(x_1; a)P(x_2; a)P(x_3; a)\cdots P(x_N; a)$$

$$= \prod P(x_i; a)$$

Maximum likelihood

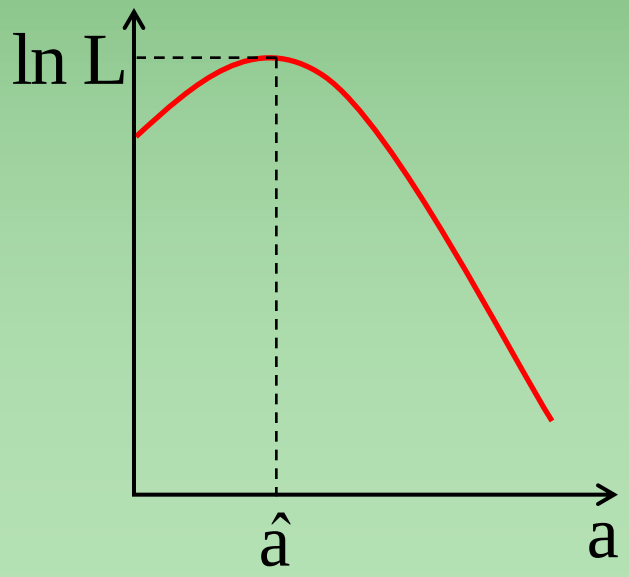
The **principle of maximum likelihood** is a method for the estimation of parameters. Consider the case where we wish to determine $\hat{a}$ for a set of data $\{x_1 + x_2 + x_3 \ldots + x_N\}$

The most probable value for $\hat{a}$ is where the likelihood is at a maximum

$$L(x_1 + x_2 + x_3 \ldots + x_N ; a)$$

Maximise the logarithmic likelihood function

$$\left.\frac{d \ln L}{d a}\right|_{a=\hat{a}} = 0$$

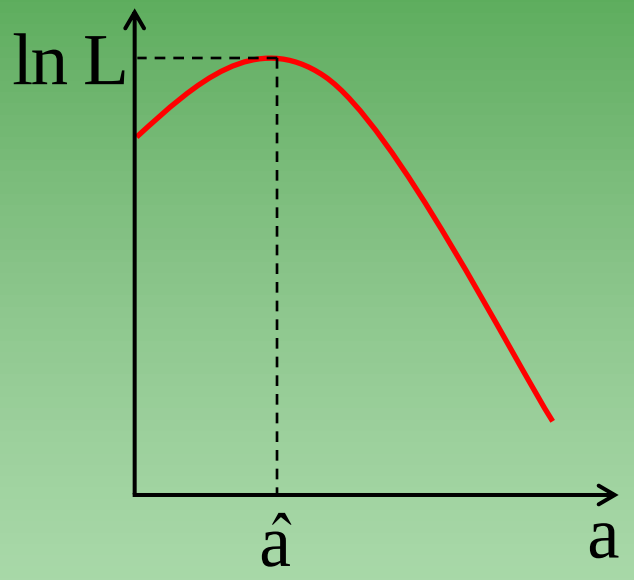


Minimum Variance Bound

It can be shown (see Barlow pg 74) that there is fundamental limit to the precision of any estimator, called the *minimum variance bound* (MVB), given by

$$V(\hat{a}) \geq \frac{1}{\left\langle (d \ln L / d a)^2 \right\rangle}$$

$$V(\hat{a}) \geq \frac{-1}{\left\langle (d^2 \ln L / d a^2) \right\rangle}$$



If an estimator $\hat{a}$ has a variance $V(\hat{a})$ equal to the MVB, then it is said to be efficient. Else, it's efficiency is given by

$$MVB / V(\hat{a})$$

Minimum Variance Bound for the Mean

For a Gaussian, the sample mean is an efficient estimate of μ

For any given x_i:

$$P(x_i;\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}$$

$$\ln L = -\sum_i \frac{(x_i-\mu)^2}{2\sigma^2} - N\ln\left(\sigma\sqrt{2\pi}\right)$$

$$\frac{d^2 \ln L}{d\mu^2} = -\frac{N}{\sigma^2}$$

MVB $$V(\hat{\mu}) \geq \frac{-1}{\left\langle (d^2 \ln L / d\mu^2) \right\rangle} = \frac{\sigma^2}{N}$$

Which is equal to the variance in $\hat{\mu}$ by the CLT

Minimum Variance Bound for the Mean

However, we could take the median as the estimator for μ. For a sample from a Gaussian parent distribution, as N tends to ∞, the variance tends to

$$\hat{\mathrm{V}} = \frac{\pi}{2\mathrm{N}}\sigma^2$$

Thus, the efficiency is

$$\frac{\mathrm{MVB}}{\hat{\mathrm{V}}} = \frac{\frac{\sigma^2}{\mathrm{N}}}{\frac{\pi}{2\mathrm{N}}\sigma^2} = \frac{1}{\pi/2} = \frac{2}{\pi} = 0.64$$

Estimating the variance

For the case where the true mean is known, μ

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

$$\left\langle \hat{V}(x) \right\rangle = \frac{N(x_i - \mu)^2}{N} = \left\langle (x-\mu)^2 \right\rangle = V(x)$$ (i.e. unbiased)

If we don't know the true mean, we have to substitute $\hat{\mu} = \bar{x}$ for μ

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

$$\left\langle \hat{V}(x) \right\rangle = \frac{N\left\langle x^2 - \bar{x}^2 \right\rangle}{N} = \left\langle x^2 \right\rangle - \left\langle \bar{x}^2 \right\rangle$$

Estimating the variance

$$\left\langle \hat{V}(x) \right\rangle = \frac{N\left\langle x^2 - \bar{x}^2 \right\rangle}{N} = \left\langle x^2 \right\rangle - \left\langle \bar{x}^2 \right\rangle$$

As the CLT states $\left\langle x \right\rangle = \left\langle \bar{x} \right\rangle$

$$\left\langle \hat{V}(x) \right\rangle = \left\langle x^2 \right\rangle - \left\langle x \right\rangle^2 - \left(\left\langle \bar{x}^2 \right\rangle - \left\langle \bar{x} \right\rangle^2 \right)$$

$$\left\langle \hat{V}(x) \right\rangle = \hat{V}(x) - \hat{V}(\bar{x})$$

The CLT also states $\hat{V}(\bar{x}) = \frac{\hat{V}(x)}{N}$

$$\left\langle \hat{V}(x) \right\rangle = \left(1 - \frac{1}{N}\right)\hat{V}(x) = \left(\frac{N-1}{N}\right)\hat{V}(x) \neq \hat{V}(\bar{x})$$ (i.e. biased)

Estimating the variance

$$\left\langle \hat{V}(x) \right\rangle = \left(1 - \frac{1}{N}\right)\hat{V}(x) = \left(\frac{N-1}{N}\right)\hat{V}(x) \neq \hat{V}(\bar{x})$$

Therefore, correcting by (Bessel's correction):

$$\left(\frac{N}{N-1}\right)$$

$$\left\langle \hat{V}(x) \right\rangle = s^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2$$

This expression is consistent and unbiased

The correlation coefficient

The correlation coefficient within a sample can be taken to be an estimate of the correlation of the parent distribution:

$$\hat{\rho} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}$$

When N is very large, it can be shown that the uncertainty is given by

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

An approximation which works to lower N is to calculate the variable z, which has an uncertainty s_z

$$z = \frac{1}{2}\ln\frac{1+\hat{\rho}}{1-\hat{\rho}} \qquad \sigma_z = \frac{1}{\sqrt{N-3}}$$

Maximum likelihood

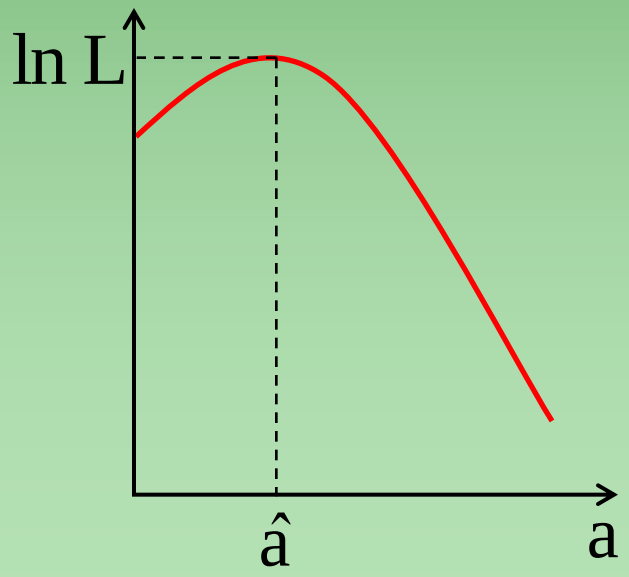
The principle of maximum likelihood is a method for the estimation of parameters. Consider the case where we wish to determine $\hat{a}$ for a set of data $\{x_1 + x_2 + x_3 \ldots + x_N\}$

The most probable value for $\hat{a}$ is where the likelihood is at a maximum

$$L(x_1 + x_2 + x_3 \ldots + x_N; a)$$

Maximise the logarithmic likelihood function

$$\left.\frac{d \ln L}{d a}\right|_{a=\hat{a}} = 0$$



Maximum likelihood

Gaussian weighted mean

Suppose we have a number of measurements of the same quantity x_i

$$P(x_i;\mu,\sigma_i)=\frac{1}{\sigma_i\sqrt{2\pi}}e^{-(x_i-\mu)^2/2\sigma_i^2}$$

$$\ln L=\sum_i -\ln\sigma_i\sqrt{2\pi}-\sum_i\frac{-(x_i-\mu)^2}{2\sigma_i^2}$$

$$\frac{d\ln L}{d\mu}=\sum_i\frac{(x_i-\hat{\mu})}{\sigma_i^2}=0 \quad \text{(for maximum likelihood)}$$

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

In agreement with our expression from the last lecture

$$\langle x\rangle=\frac{\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}$$

Maximum likelihood

Maximum likelihood estimators are usually *consistent*. However, they are not generally *unbiased*, even when one estimator is derived from another unbiased one.

Suppose we have an estimator $\hat{a}$ which is unbiased and symmetric about the true value $a_0 = 1.0$, with a width of $\sigma_a = 0.1$

Suppose we measure $\hat{a} = 1.1 a_0$. Then $\hat{a}^2 = 1.21 a_0^2$

Suppose we measure $\hat{a} = 0.9 a_0$. Then $\hat{a}^2 = 0.81 a_0^2$

That is, $\hat{a}$ is *symmetric*, an is an *unbiased* estimator

But, $\hat{a}^2$ is *asymmetric*, and is a *biased* estimator

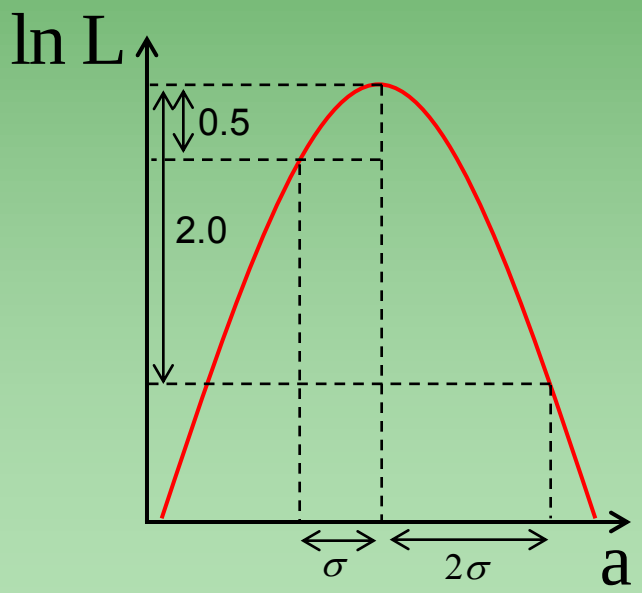
Maximum likelihood

As N tends becomes large, any consistent estimator becomes *unbiased*, and becomes *efficient*. It can also be shown that, due to the CLT, the probability distribution of $\hat{a}$ will be Gaussian, with a standard deviation equal to the standard deviation on the estimator $\hat{a}$

$$\Delta L = -0.5 \equiv \sigma_a$$

$$\Delta L = -2.0 \equiv 2\sigma_a$$

$$\Delta L = -4.5 \equiv 3\sigma_a$$



Sometimes $\ln L$ *and* dL/da *are solvable analytically. Other times, a numerical solution must be sought*

Maximum likelihood

Least Squares

Suppose we have a data sample consisting of (x,y) pairs, where the x_i are known precisely, and y_i are measured with resolution σ_i. It is understood that y is described by some function $f(x;a)$ and we wish to obtain a. The CLT states that the probability of a particular y_i at any x_i is

$$P(y_i;a) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-[y_i - f(x_i;a)]^2/2\sigma_i^2}$$

$$\ln L = -\frac{1}{2}\sum_i \left[\frac{y_i - f(x_i;a)}{\sigma_i}\right]^2 - \sum_i \ln \sigma_i\sqrt{2\pi}$$

To maximise L, one has to minimise

$$\sum_i \left[\frac{y_i - f(x_i;a)}{\sigma_i}\right]^2$$

Stratified Sampling

Suppose we wish to determine a property of some data which have two distinct sub-categories, such as the average height of students at your university (who of course can be male or female). The average height of male and female students is different. Variations in this ratio in your sample will add to the fluctuations in the determined average height.

If you sample randomly, where $f_1 + f_2 = 1$

$$V(x) = \int \left[f_1 P_1(x) + f_2 P_2(x) \right] (x - \mu)^2 \, dx$$

$$V(x) = f_1 V_1(x) + f_2 V_2(x) + f_1 f_2 (\mu_1 - \mu_2)^2$$

The last term describes the contribution to the variance due to fluctuations in the ratio of types 1 and 2, and drops out if the sampling ratio is held constant (stratified sampling)

Stratified Sampling

What is the best ratio to adopt when using stratified sampling?

Suppose we take m_1 measurements of type 1 and m_2 of type 2

$$\hat{\mu} = f_1\hat{\mu}_1 + f_2\hat{\mu}_2 \qquad V = \frac{f_1^2 V_1}{m_1} + \frac{f_2^2 V_2}{m_2}$$

$$\frac{m_1}{m_2} = \frac{f_1\sqrt{V_1}}{f_2\sqrt{V_2}} = \frac{f_1\sigma_1}{f_2\sigma_2}$$

For the case where both σ are the same, the optimum ratio of the sample is the ratio of the parent population. Otherwise, the ratios should be weighted by the σ of each distribution

Summary

- Estimation
  - Consistency
  - Bias
  - Efficiency
  - Minimum Variance Bound
  - Estimating Variance
  - Maximum Likelihood
  - Stratified sampling

UNIVERSITY OF THE WEST of SCOTLAND

Introductory Data Analysis

Lecture 5

χ^2 and Least-squares fitting

SUPA Graduate School 2009
Prof. A. Andreyev

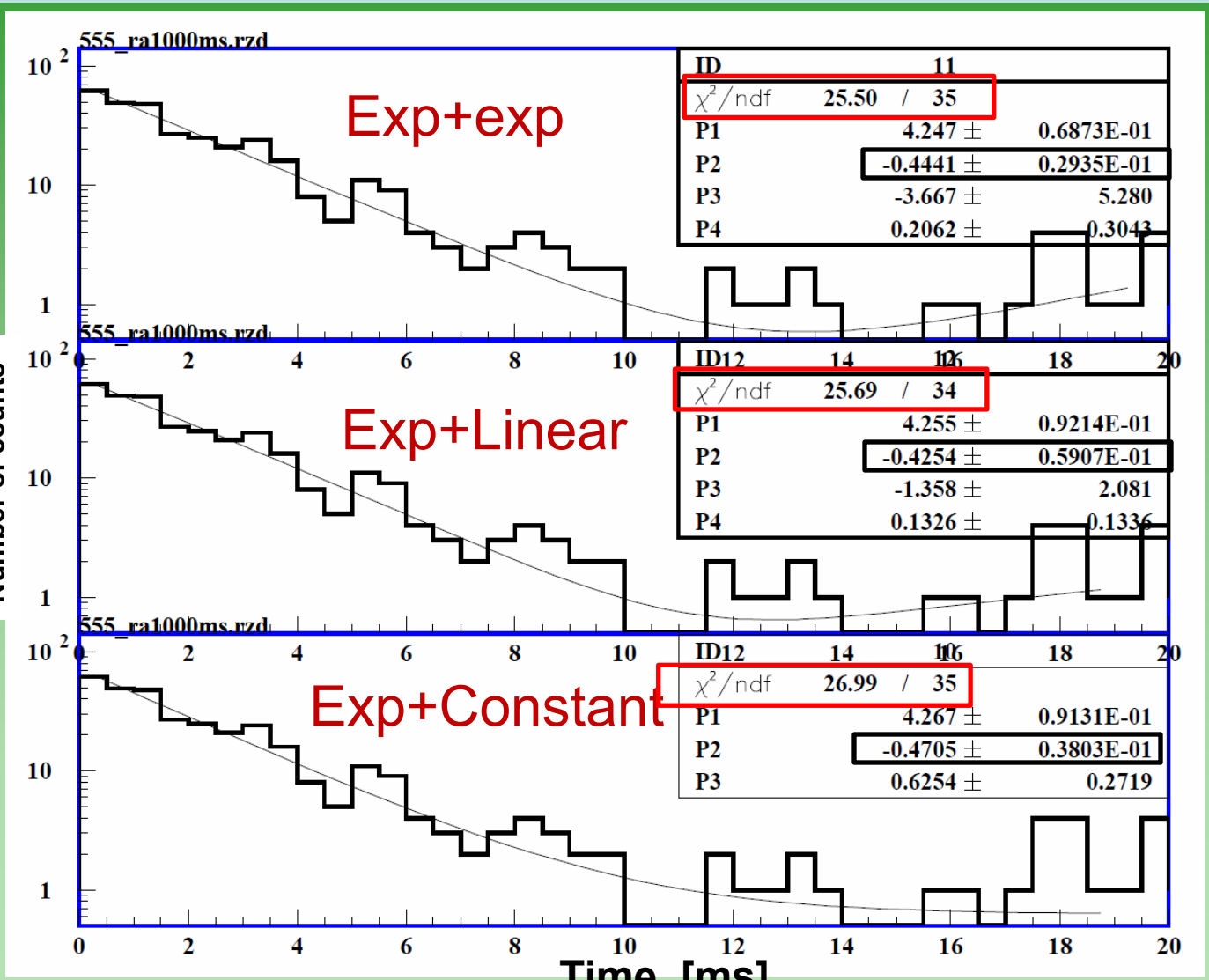
Nuclear Physics Research Group
School of Engineering and Science

Lecture 5 outline

- An introduction (a simple example)
- χ^2 and reduced χ^2
- Fitting with a linear function
- More complex examples

A reminder from Lecture 2

Example: **Half-life determination of a nucleus by fitting its decay curve (the same data!) with 3 variants of functions**



$T_{1/2}=\ln 2/P2$

$T_{1/2}=1.56$ ms
smallest fit uncertainty

$T_{1/2}=1.63$ ms

$T_{1/2}=1.47$ ms

Least Squares Fitting

- **The least squares method** is a mechanism for extracting parameters from an experimental data set

- In its simplest form, one has two variables, x and y, of which **x are known** and **y are measured with some precision σ**. The *x* and *y* are related by an 'expected' function

$$y = f(x;a)$$

where a is an unknown parameter (to be determined)

- In **the Least-Squares Procedure** one wants to establish both the function $f(x,a)$ and the parameter *a* in such a way, that the 'expected' values of $y_{expected}=f(x,a)$ are close to the experimentally measured values y_i

Least Squares Fitting

- The data are a set of **precise values of x** $\{x_1 + x_2 + x_3 \ldots + x_N\}$

for which there are a corresponding set of measurements of $y \quad \{y_1 + y_2 + y_3 \ldots + y_N\}$ with precision σ_i

- The least squares method requires the calculation and minimization of a quantity χ^2, which represents the total deviation of data y_i from **the expected values** given by $y = f(x;a)$

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i;a)}{\sigma_i} \right]^2 \qquad \chi^2 = \sum_{i=1}^{N} \left[\frac{y_i^{\text{measured}} - y_i^{\text{ideal/expected}}}{\text{estimated uncertainty}} \right]^2$$

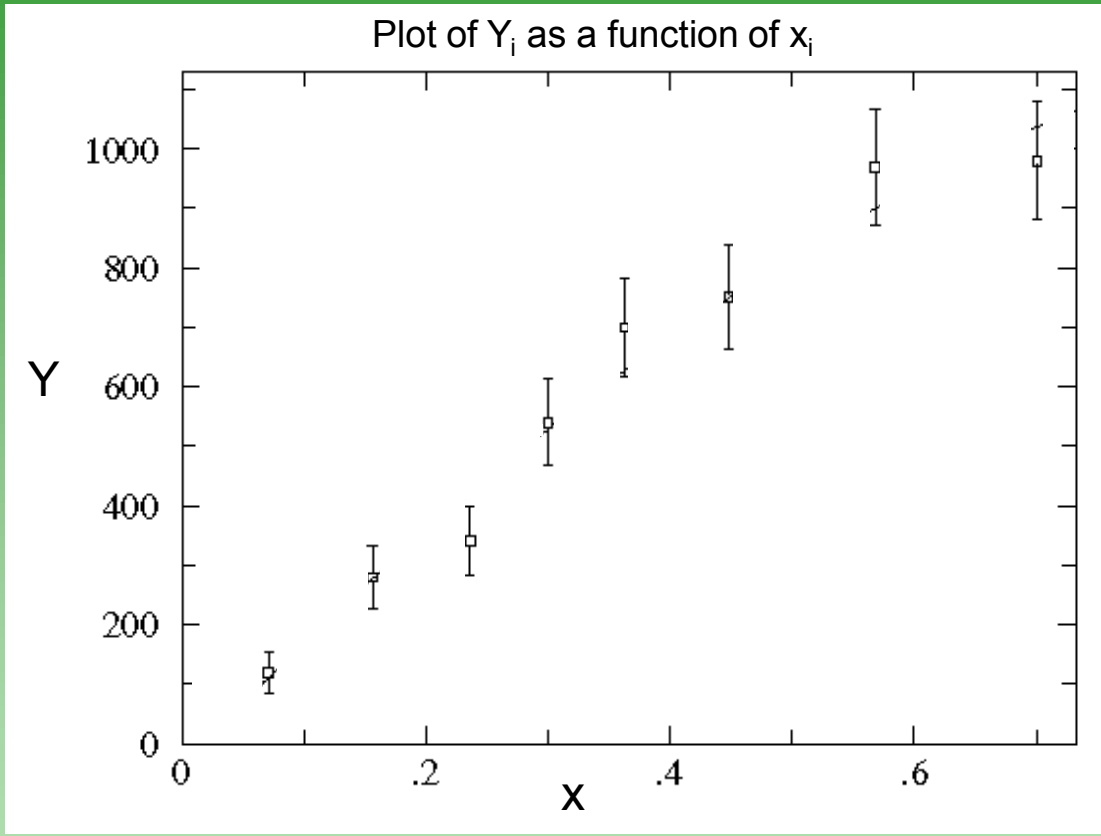
If all the data have the same σ_i then this simplifies to:

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^{N} \left[y_i - f(x_i;a) \right]^2$$

A simple Example

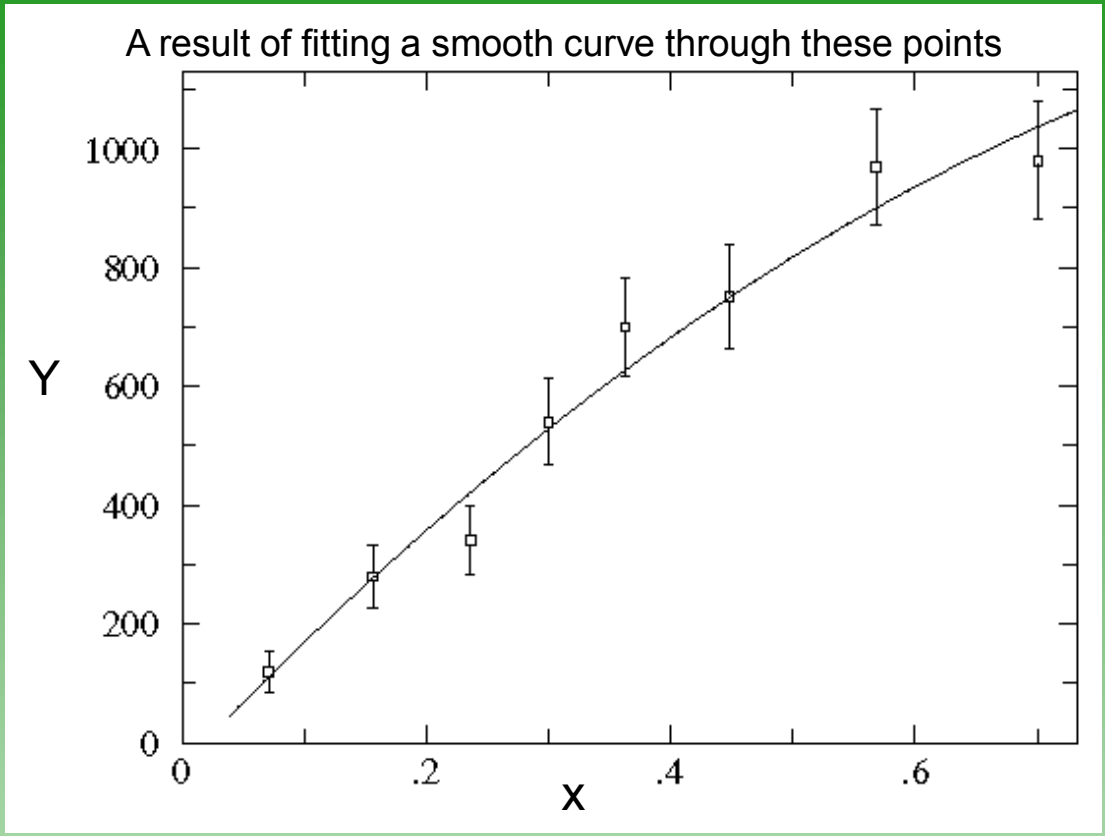
Imagine, we have a set of some data, measured **at precise values of x_i**, resulting in measured values y_i with a standard deviation of σ_i

Value (x_i)	*N*	*$N^{1/2}$*	*Measured y_i*
0.071	12	3.5	120 ± 35
0.156	28	5.3	280 ± 53
0.236	34	5.8	340 ± 58
0.300	54	7.3	540 ± 73
0.363	70	8.4	700 ± 84
0.448	75	8.7	750 ± 87
0.568	97	9.8	970 ± 98
0.701	98	9.9	980 ± 99



- **A question: what function describes the measured dependence?**
- **Linear? Exponential? Parabolic?....**
- **Sometimes we know the 'expected' dependence, sometimes – not and want to deduce it**

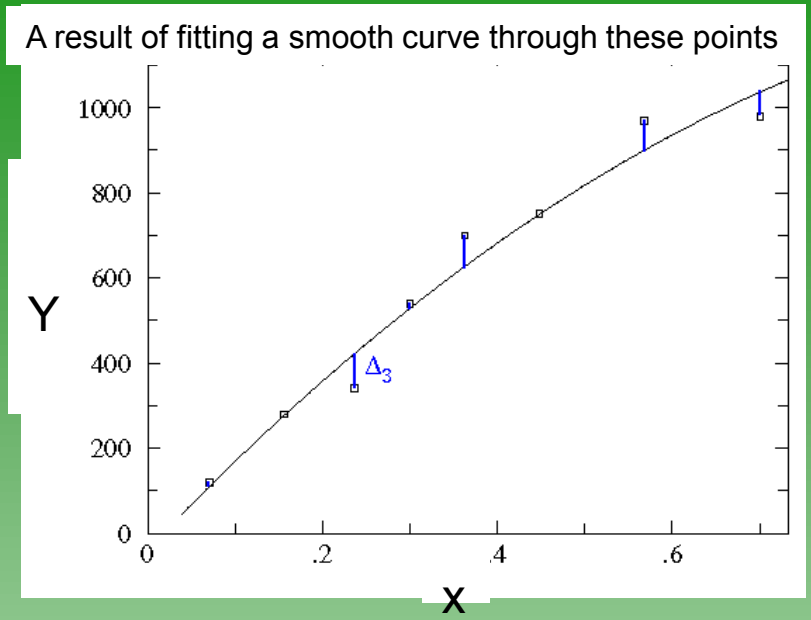
Let's try a 'smooth' curve



- The "smooth curve" is a curve that **mostly passes through the data themselves (and the error bars σ_i)**

- **Note, that if error bars are large, then we can draw many such curves, and will need to find a mechanism to tell which of the curves is the best**

“Good” and “Bad” points



- If the error bars σ_i **are large**, the **deviation from the curve** (denoted by Δ_i, shown in blue) **can also be large.**

- **What counts is the relative size of the deviation Δ_i and the error bar σ_i**

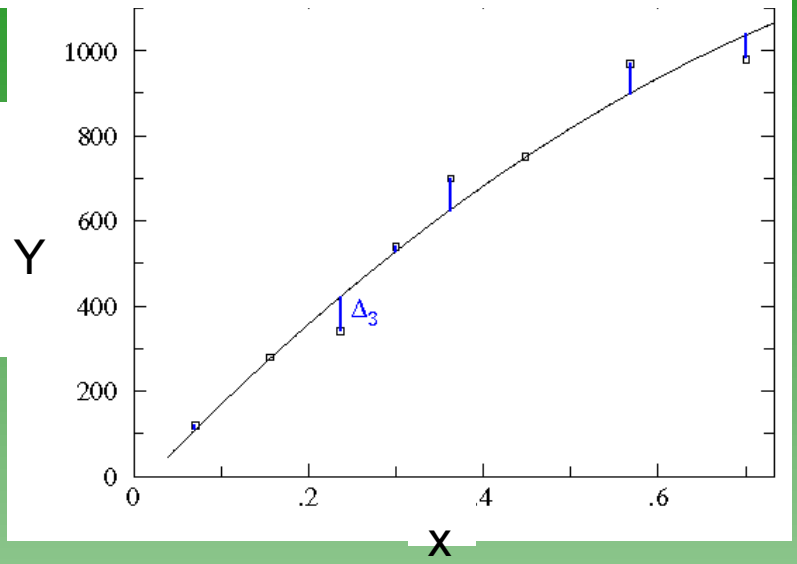
- "Good" points have a small (less than 1) ratio of deviation Δ_i **and the error** σ_i

- "Bad" points have a ratio of deviation to error **larger than one**, and hence the curve fails to go through the error bar (as in the third data point)

- On average a good fit will have as many unusually large deviations as unusually small deviations, that is, on average the ratio of deviation to error will be about 1.
- Of course, in a perfect fit the curve will go right through every data point: zero deviation Δ_i (seldom happens)

χ^2 and degree of freedom

A result of fitting a smooth curve through these points



$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i; a)}{\sigma_i} \right]^2 \qquad \Delta_i$$

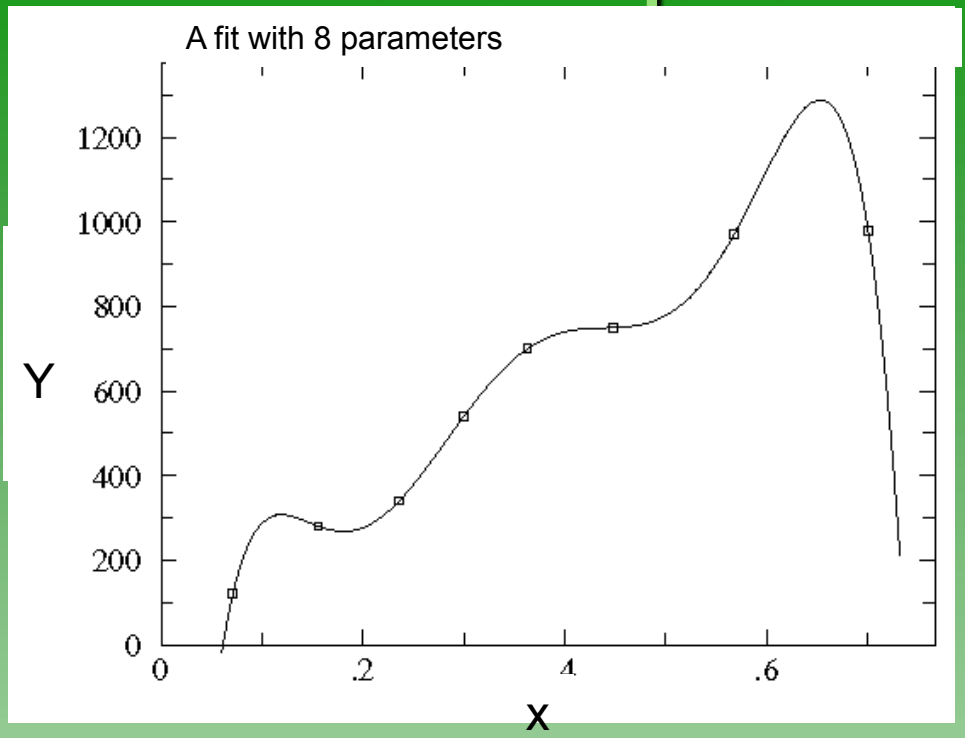
Thus, χ^2 is defined as the sum of the square of each data point's ratio of deviation to error

$$\chi^2 = \left(\frac{\Delta_1}{\sigma_1}\right)^2 + \left(\frac{\Delta_2}{\sigma_2}\right)^2 + \left(\frac{\Delta_3}{\sigma_3}\right)^2 + \cdots + \left(\frac{\Delta_N}{\sigma_N}\right)^2$$

- **On average** we expect each term in the sum to be about 1 so the total χ^2 should be **about equal the number of data points N_i (in our case: 8)**

- Note, that by selecting a fitting-curve with as many adjustable parameters as data points, one can usually force the curve to exactly hit every data point: **a perfect fit but probably of no significance**

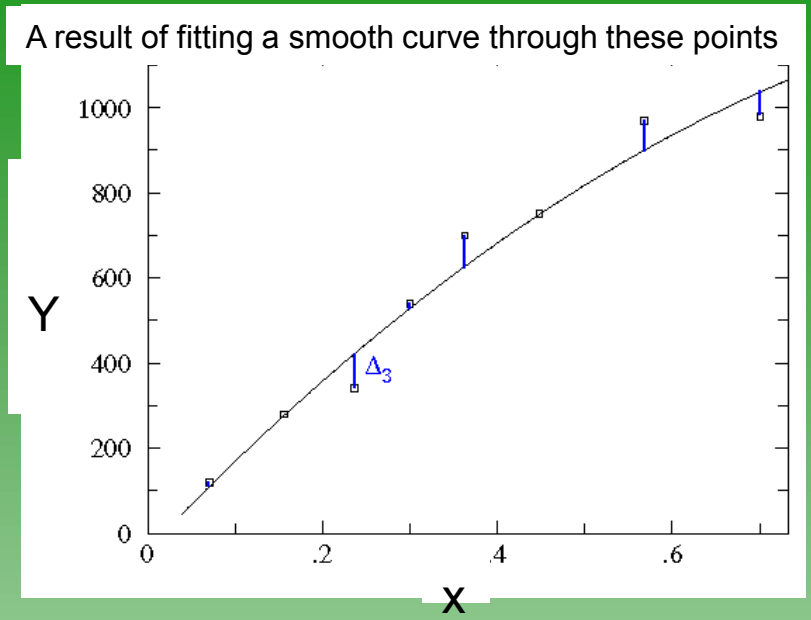
An example of wrong fitting



$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i; a)}{\sigma_i} \right]^2$$

- Here our **eight data points** are exactly hit by a polynomial with **eight adjustable parameters**
- Thus we can make χ^2 =0 by selecting a curve that twists and turns to hit every point
- But, most probably, no one would think that the actual relationship is this bizarre **and it is useless to fit a curve exactly through inexact data**!

χ^2 and degree of freedom



$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i; a)}{\sigma_i} \right]^2$$

(Δ_i = $y_i - f(x_i; a)$)

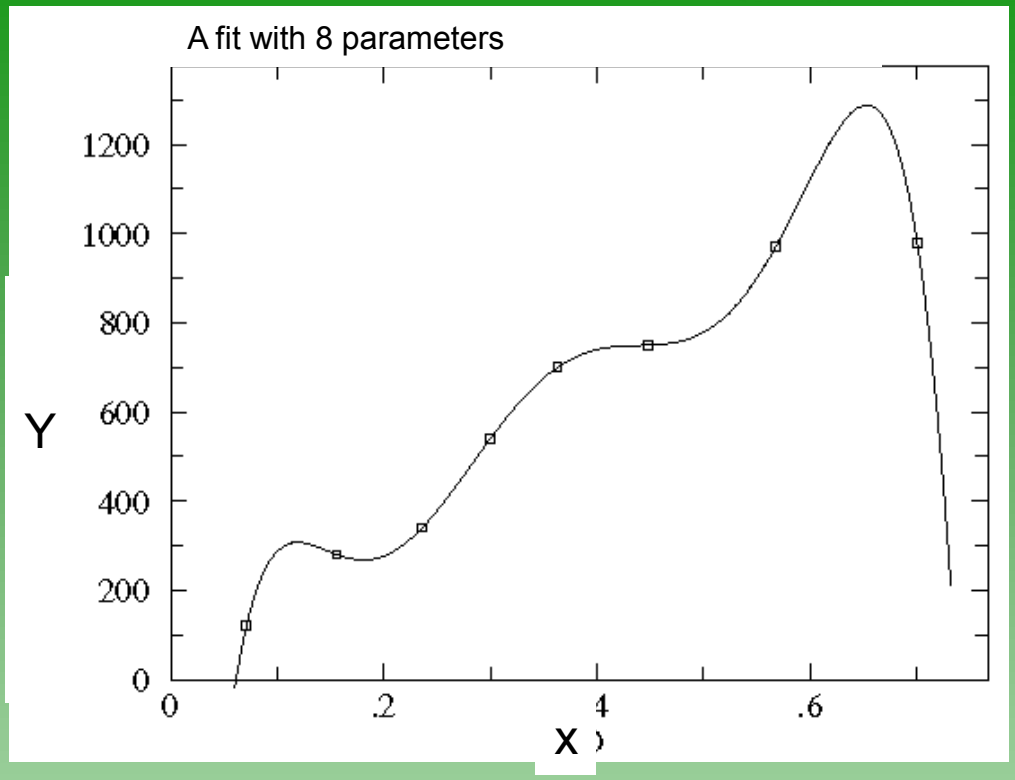
Thus, χ^2 is defined as the sum of the square of each data point's ratio of deviation to error

$$\chi^2 = \left(\frac{\Delta_1}{\sigma_1}\right)^2 + \left(\frac{\Delta_2}{\sigma_2}\right)^2 + \left(\frac{\Delta_3}{\sigma_3}\right)^2 + \cdots + \left(\frac{\Delta_N}{\sigma_N}\right)^2$$

- The number of "effective" data points, i.e., those that could not be automatically hit by the curve, **is called the number of "degrees of freedom“**

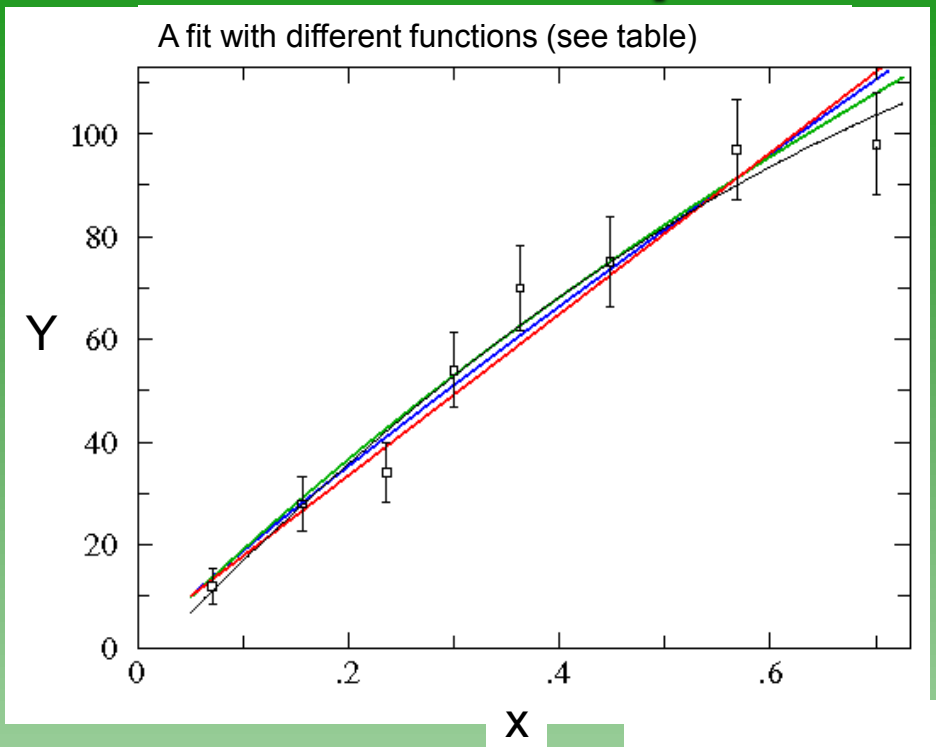
- Degrees of freedom (d.f.)= number of data points - number of adjustable parameters

Reduced χ^2



- We want a curve with as few twists and turns as possible that comes near (or better yet: inside) each error bar.
- We focus then on the χ^2 **per degree of freedom**: *reduced chi-square*= χ^2 /(d.f.)
- This number should be expect to be ~1. (If it is less than one, we have an unexpectedly good fit; If it is much greater than one, the curve is missing too many data points to be believed)

Examples of Good Fits

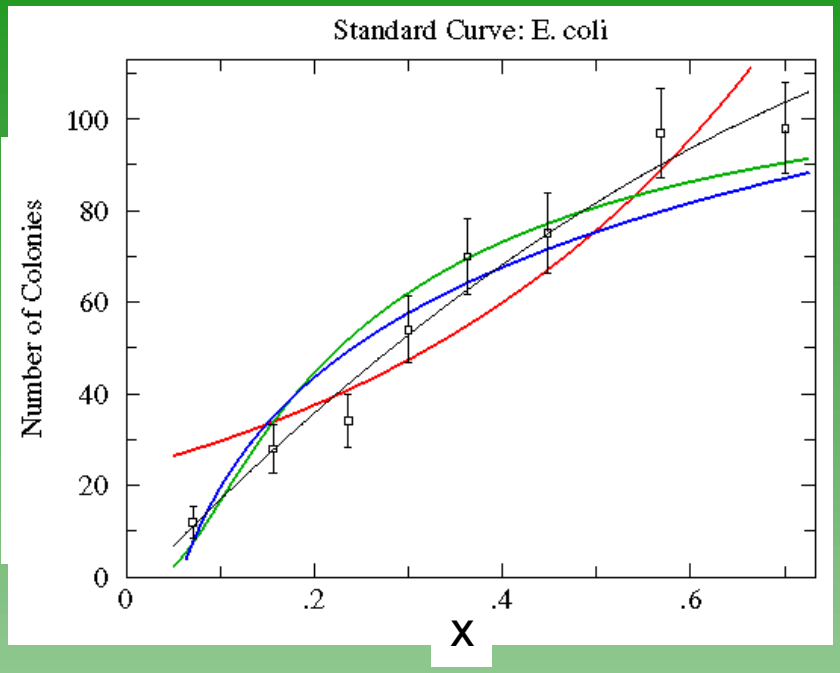


These functions have essentially the **same reduced χ^2** values over the region covered by the data; any of them would make a fine choice.

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i; a)}{\sigma_i} \right]^2$$

Curve	Color	Reduced χ^2	Parameter Values *a*	*b*	*c*
Linear y=a+bx	red	0.93	2.11	157	
Power $y=a\,x^b$	blue	0.85	153	0.911	
Inverse X & Y 1/y=a+b/x	green	0.77	0.00207	0.00504	
Quadratic $y=a+bx+cx^2$	black	0.73	-3.7	215	-88

Examples of Bad Fits



Here are some other functions **which do not make good fits** (reduced χ^2~2.5-5; green=Arrhenius, blue=Natural Log, red=Exponential)

Least Squares Fitting

- **The least squares method** is a mechanism for extracting parameters from an experimental data set

- In its simplest form, one has two variables, x and y, of which **x are known** and **y are measured with some precision σ**. The *x* and *y* are related by an 'expected' function

$$y = f(x;a)$$

where a is an unknown parameter (to be determined)

- In **the Least-Squares Procedure** one wants to establish both the function $f(x,a)$ and the parameter *a* in such a way, that the 'expected' values of $y_{expected}=f(x,a)$ are close to the experimentally measured values y_i

Least Squares Fitting

- The data are a set of **precise values of x** $\{x_1 + x_2 + x_3 \ldots + x_N\}$

for which there are a corresponding set of measurements of $y \quad \{y_1 + y_2 + y_3 \ldots + y_N\}$ with precision σ_i

- The least squares method requires the calculation and minimization of a quantity χ^2, which represents the total deviation of data y_i from **the expected values** given by $y = f(x;a)$

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i;a)}{\sigma_i} \right]^2 \qquad \chi^2 = \sum_{i=1}^{N} \left[\frac{y_i^{\text{measured}} - y_i^{\text{ideal/expected}}}{\text{estimated uncertainty}} \right]^2$$

If all the data have the same σ_i then this simplifies to:

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^{N} \left[y_i - f(x_i;a) \right]^2$$

Minimization of χ^2

By choosing a value of a which **minimises χ^2**, the best fit of the function to the data is obtained, thus implying the best estimator of a

If the derivatives of f with respect to a are know, then one just has to find the solution for which

$$\frac{d\chi^2}{da} = 0 \qquad \chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{N}\left[y_i - f(x_i;a)\right]^2$$

Which is $$\frac{d\chi^2}{da} = -\sum_i \frac{1}{\sigma^2}\frac{df(x_i;a)}{da}\left[y_i - f(x_i;a)\right] = 0$$

A simple case – direct proportionality

Let us consider a case for which we expect the data to be described by a **linear dependence, with no offset**. Therefore,

$$f(x;a) = ax$$

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - ax_i)^2}{\sigma_i^2}$$

Differentiating this gives:

$$\frac{d\chi^2}{da} = \sum_i -2x_i \frac{y_i - ax_i}{\sigma_i^2}$$

And for uniform σ:

$$\frac{d\chi^2}{da} = -\frac{2}{\sigma^2} \sum_i \left(x_i y_i - a x_i^2 \right)$$

A simple case – direct proportionality

For the least-squares estimate of the slope, $\hat{a}$

$$\sum_i \left(x_i y_i - \hat{a} x_i^2 \right) = 0$$

$$\sum_i x_i y_i = \hat{a} \sum_i x_i^2$$

If we divide both sides by N, we have averages in place of the summations

$$\overline{xy} = \hat{a}\, \bar{x}^2$$

$$\hat{a} = \frac{\overline{xy}}{\bar{x}^2}$$

A simple case – direct proportionality

$$\hat{a} = \frac{\overline{xy}}{\overline{x}^2}$$

What is the uncertainty on this?

We can write it in the following form $\hat{a} = \sum \frac{x_i}{N\overline{x}^2} y_i$

We know from our previous error analysis that:

$$V(\hat{a}) = \left(\frac{df}{dx}\right)^2 V(x) + \left(\frac{df}{dy}\right)^2 V(y)$$

$$V(\hat{a}) = \sum \left(\frac{x_i}{N\overline{x^2}}\right)^2 \sigma^2$$

$$V(\hat{a}) = \frac{\sigma^2}{N\overline{x^2}}$$

A common case – a general linear fit

Let us now consider a case for which we expect the data to be described by a **linear fit, with a possible offset**. The uncertainties on each point are equal (for simplicity only). Therefore,

$$f(x;m,c) = mx + c$$

$$\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y_i - mx_i - c\right)^2$$

The term to be minimised is thus $\sum_{i=1}^{N}\left(y_i - mx_i - c\right)^2$

Differentiating this **with respect to c** and substituting estimators

$$\sum_i -2\left(y_i - \hat{m}x_i - \hat{c}\right) = 0$$

A common case – a general linear fit

$$\sum_i -2\left(y_i - \hat{m}x_i - \hat{c}\right) = 0$$

Dividing this by N gives

$$\bar{y} - \hat{m}\bar{x} - \hat{c} = 0$$

Now, differentiating the sum $\sum_{i=1}^{N}\left(y_i - mx_i - c\right)^2$

with respect to m gives

$$\sum_i -2x_i\left(y_i - \hat{m}x_i - \hat{c}\right) = 0$$

$$-2\sum_i \left(x_i y_i - \hat{m}x_i^{\,2} - \hat{c}x_i\right) = 0$$

Again, dividing this by N gives

$$\overline{xy} - \hat{m}\overline{x^2} - \hat{c}\bar{x} = 0$$

A common case – a general linear fit

$$\bar{y} - \hat{m}\bar{x} - \hat{c} = 0$$

$$\overline{xy} - \hat{m}\overline{x^2} - \hat{c}\bar{x} = 0$$

Combining these expressions to eliminate $\hat{c} = \bar{y} - \hat{m}\bar{x}$ gives the result for the slope

$$\overline{xy} - \hat{m}\overline{x^2} - (\bar{y} - m\bar{x})\bar{x} = 0$$

$$\overline{xy} - \hat{m}\overline{x^2} + \hat{m}\bar{x}^2 - \bar{x}\bar{y} = 0$$

$$\overline{xy} - \hat{m}\left(\overline{x^2} - \bar{x}^2\right) - \bar{x}\bar{y} = 0$$

$$-\hat{m}\left(\overline{x^2} - \bar{x}^2\right) = \overline{xy} - \bar{x}\bar{y}$$

$$\hat{m} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$\hat{m} = \frac{\mathrm{cov}(x, y)}{V(x)}$$

Thus, we can now solve for $\hat{c}$

$$\hat{c} = \bar{y} - \hat{m}\bar{x}$$

A common case – a general linear fit

If we return to our previous expression **for the slope**

$$\hat{m} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

we can consider the uncertainty on this parameter by writing it in the form

$$\hat{m} = \sum_i \frac{x_i - \bar{x}}{N(\overline{x^2} - \bar{x}^2)} y_i$$

Once again, our variance is

$$V(\hat{a}) = \left(\frac{df}{dx}\right)^2 V(x) + \left(\frac{df}{dy}\right)^2 V(y)$$

$$V(\hat{m}) = \sum_i \left(\frac{x_i - \bar{x}}{N(\overline{x^2} - \bar{x}^2)}\right)^2 \sigma^2$$

$$V(\hat{m}) = \frac{\sigma^2}{N(\overline{x^2} - \bar{x}^2)}$$

A common case – a general linear fit

The expression for the intercept $\hat{c} = \bar{y} - \hat{m}\bar{x}$

can also be expressed as $\hat{c} = \dfrac{\overline{x^2}\,\bar{y} - \bar{x}\,\overline{xy}}{\overline{x^2} - \bar{x}^2}$

Again, our variance is $V(\hat{a}) = \left(\dfrac{df}{dx}\right)^2 V(x) + \left(\dfrac{df}{dy}\right)^2 V(y)$

Which can be shown to be

$$V(\hat{c}) = \sum_i \left(\frac{\overline{x^2} - \bar{x}x_i}{N(\overline{x^2} - \bar{x}^2)} \right)^2 \sigma^2$$

$$V(\hat{c}) = \frac{\sigma^2 \overline{x^2}}{N(\overline{x^2} - \bar{x}^2)}$$

A common case – a general linear fit

We can now consider a case for which we expect the data to be described by a linear fit, and where the **uncertainties on each point vary**. Therefore,

$$f(x;m,c) = mx + c$$

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - mx_i - c)^2}{\sigma_i^2}$$

This will result in the same expressions

$$\hat{m} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \qquad \hat{c} = \frac{\overline{x^2}\,\bar{y} - \bar{x}\,\overline{xy}}{\overline{x^2} - \bar{x}^2}$$

But this time the values have been determined **by weighting them by** $1/\sigma_i^2$

And the normalisation **is not N** but $\sum \frac{1}{\sigma_i^2}$

A common case – a general linear fit

Furthermore, for the cases before where we had a generic σ

This has to be replaced with the mean standard deviation

$$\overline{\sigma^2} = \frac{\sum \frac{\sigma_i^2}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}} = \frac{N}{\sum \frac{1}{\sigma_i^2}}$$

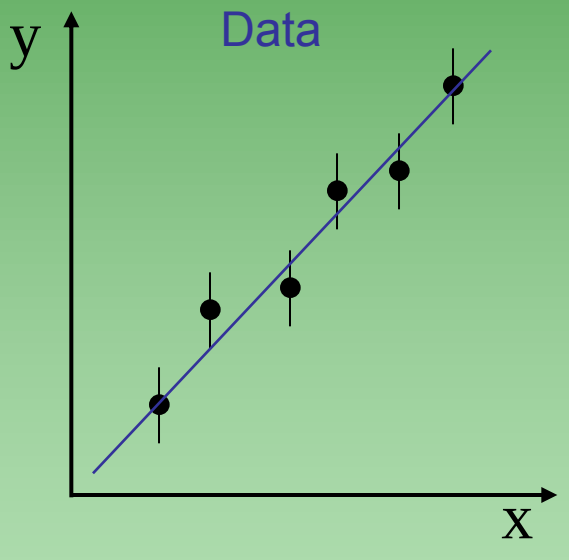
This will result in the expressions

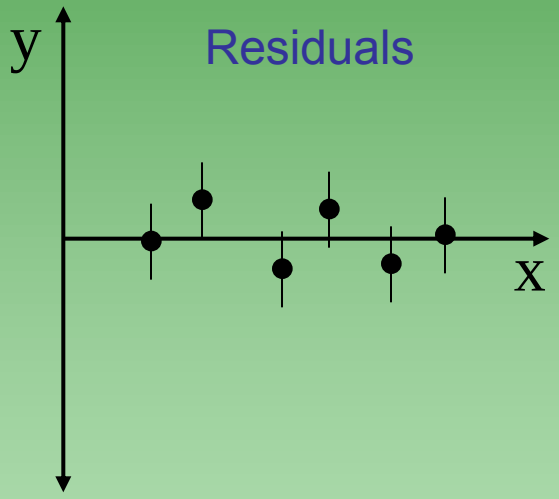
$$V(\hat{m}) = \frac{\overline{\sigma^2}}{N(\overline{x^2} - \overline{x}^2)}$$

$$V(\hat{c}) = \frac{\overline{\sigma^2}\,\overline{x^2}}{N(\overline{x^2} - \overline{x}^2)}$$

Residuals

Having fitted your data, it is possible **to calculate the residuals**, which are the differences between each of your data points and the fitted function

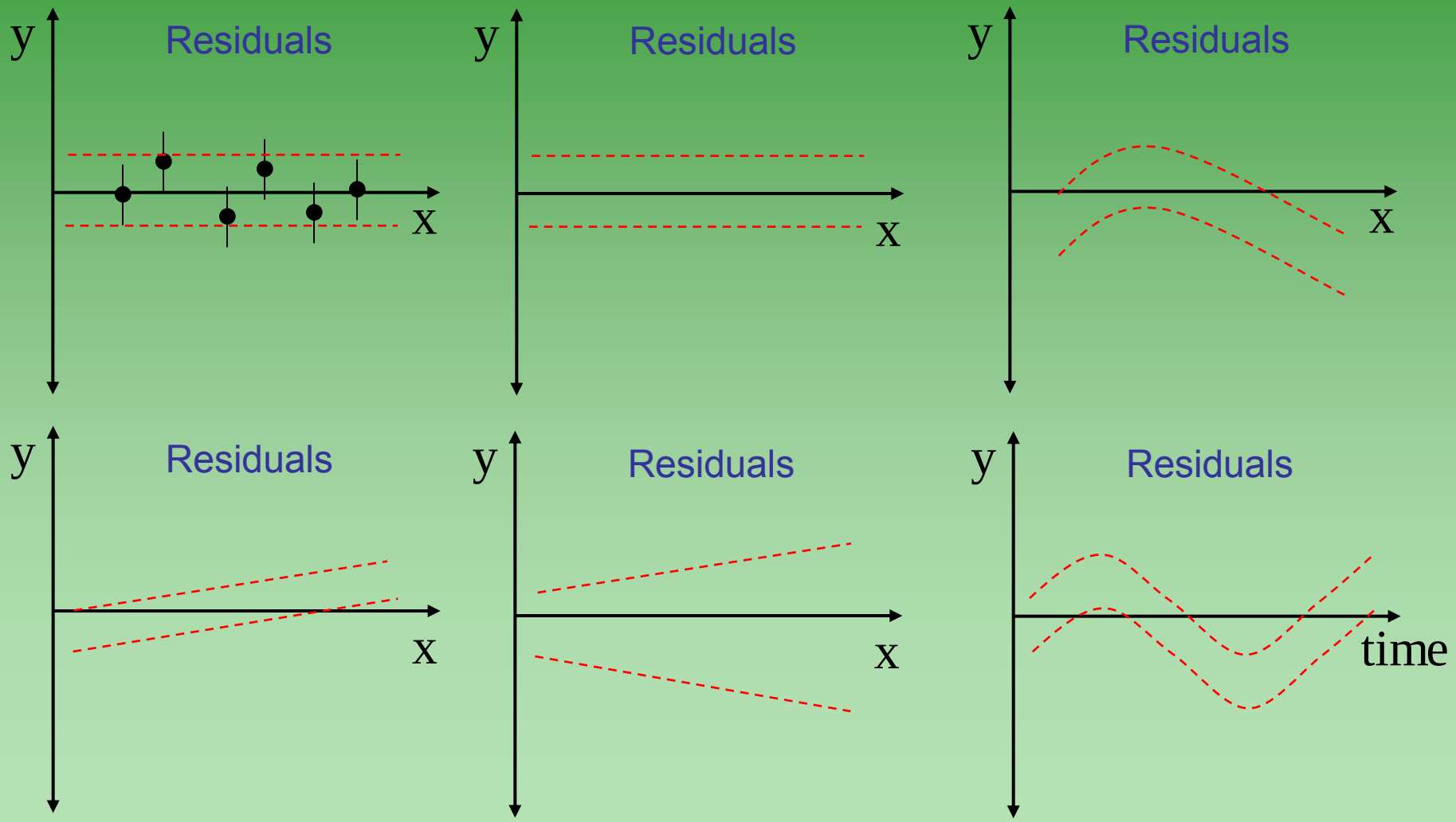




The distribution of your residuals gives you some idea of the appropriateness of your fit

Residuals

The distribution of your residuals gives you some idea of the appropriateness of your fit



What should χ^2 be?

Having fitted your data, and minimised χ^2, how do you know if the fit is good?

The χ^2 is effectively $\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i^{\,actual} - y_i^{\,ideal}}{estimated uncertainty} \right]^2$

Thus, the **χ^2 should be small**, but if it is very small, then it suggests that you've over-estimated your uncertainties.

For a quantitative analysis, we can turn to the χ^2 distribution

$$P(\chi^2; n) = \frac{2^{-n/2}}{\Gamma(n/2)} \chi^{n-2} e^{-\chi^2/2}$$

Here, $\Gamma(x)$ is the standard gamma function, and **n is the number of degrees of freedom**, that is N minus the number of free parameters.

This distribution has a **mean n and a variance $2n$**

What should χ^2 be?

The χ^2 distribution has a mean n and a variance $2n$

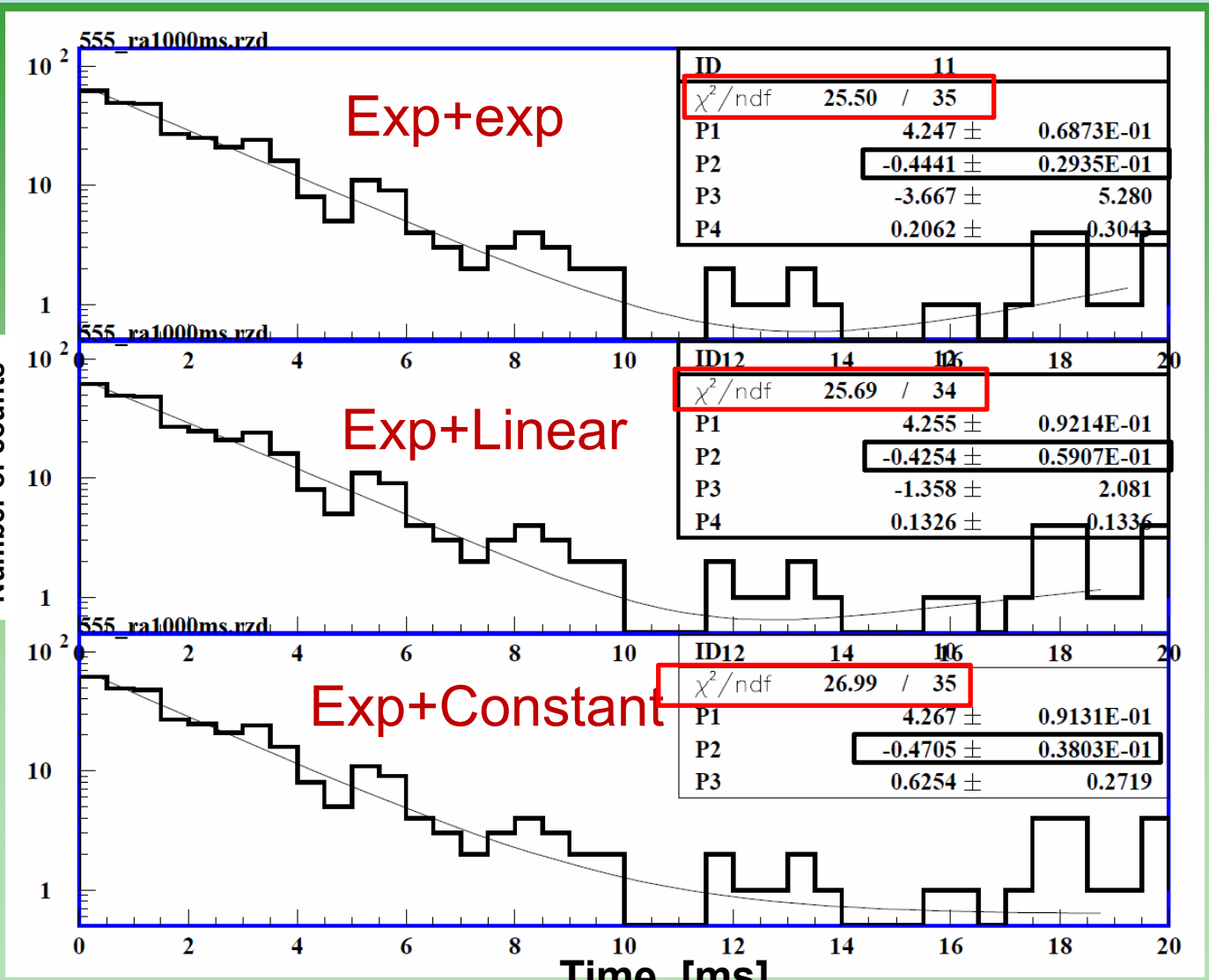
This means that the χ^2 per degree of freedom should be approximately equal to 1

$$\frac{\chi^2}{n} \approx 1$$

This will be true for a good fit. A χ^2 per degree of freedom that is much greater than 1 suggests that there is something wrong with the fit (or the uncertainties are not correct)

A reminder from Lecture 2

Example: **Half-life determination of a nucleus by fitting its decay curve (the same data!) with 3 variants of functions**



$T_{1/2}=\ln 2/P2$

$T_{1/2}=1.56$ ms
smallest fit uncertainty

$T_{1/2}=1.63$ ms

$T_{1/2}=1.47$ ms

More complex functions

It may be that we expect the data to be described by more complex function, perhaps a quadratic.

$$f(x; m, c) = ax^2 + bx + c$$

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(y_i - ax_i^2 - bx_i - c\right)^2}{\sigma_i^2}$$

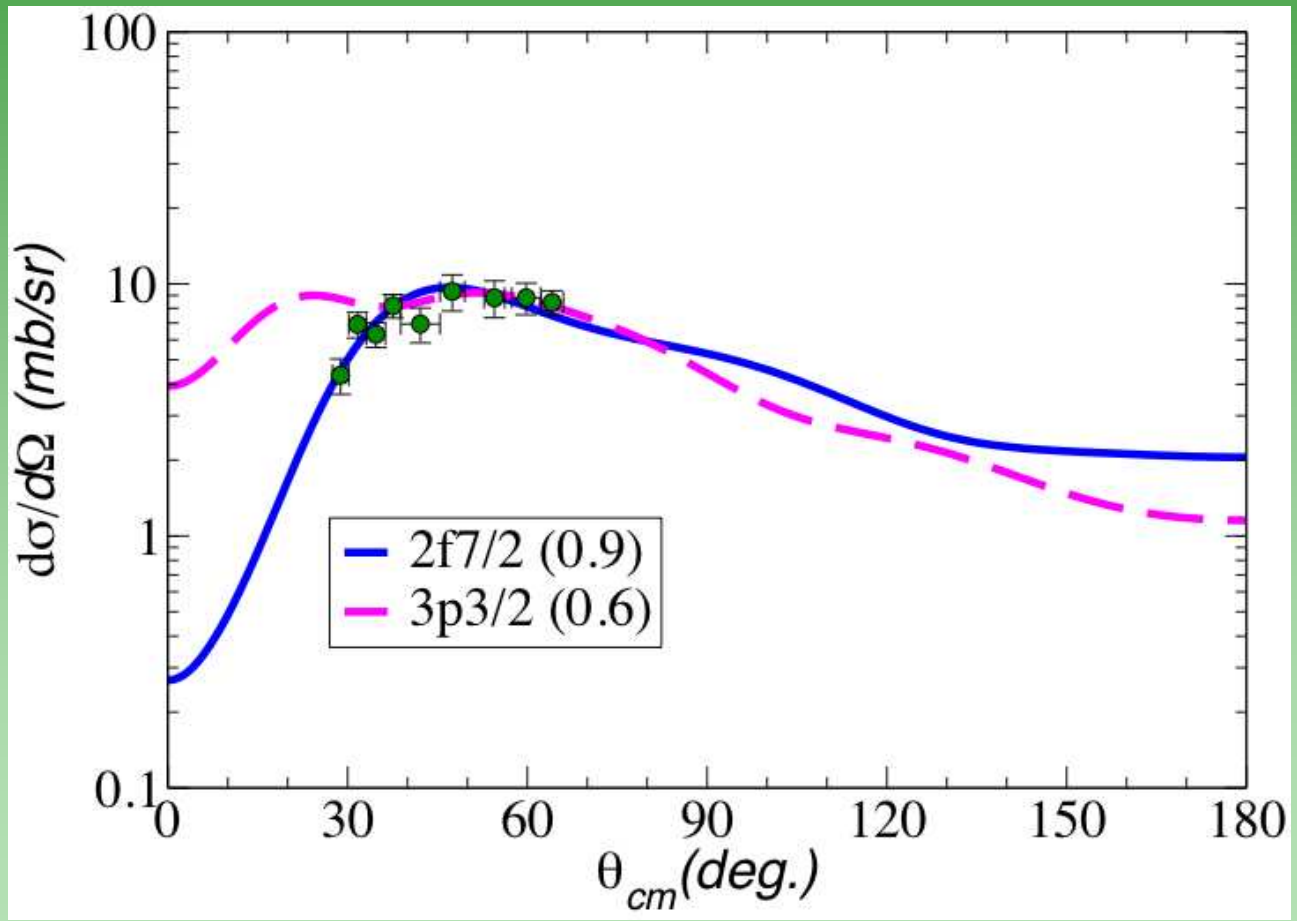
Or it may be that are trying to fit a Gaussian curve to a peak in a spectrum.

$$f(x; A, \mu, \sigma) = Ae^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

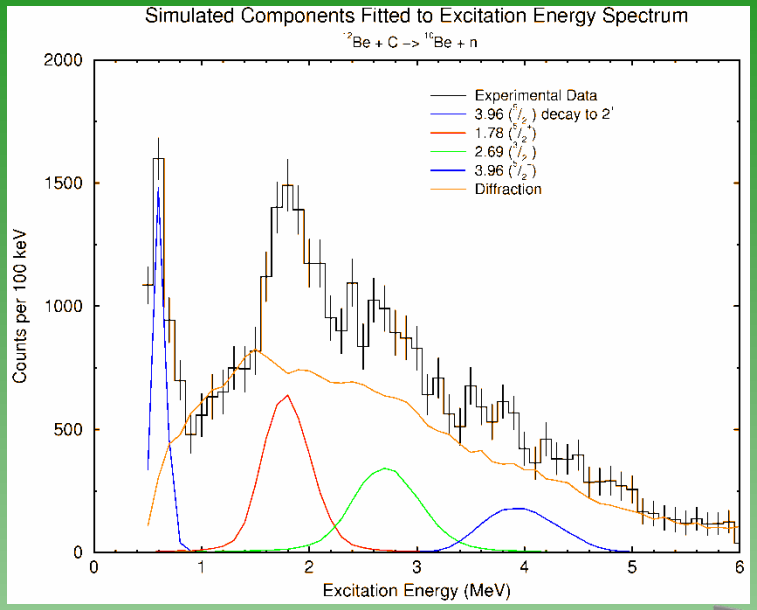
$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - Ae^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma_i} \right]^2$$

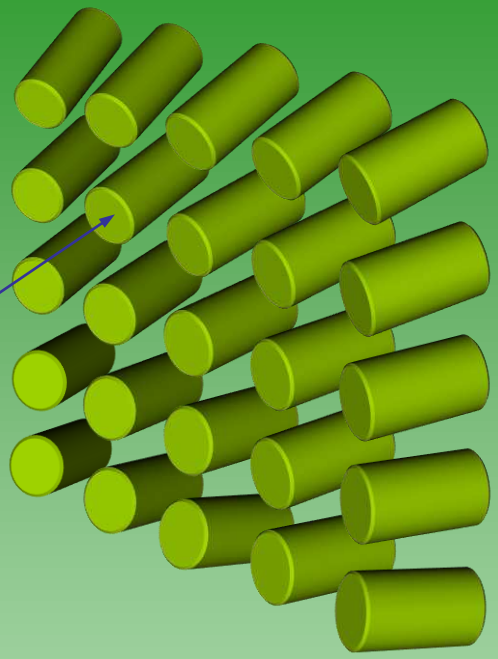
More complex functions

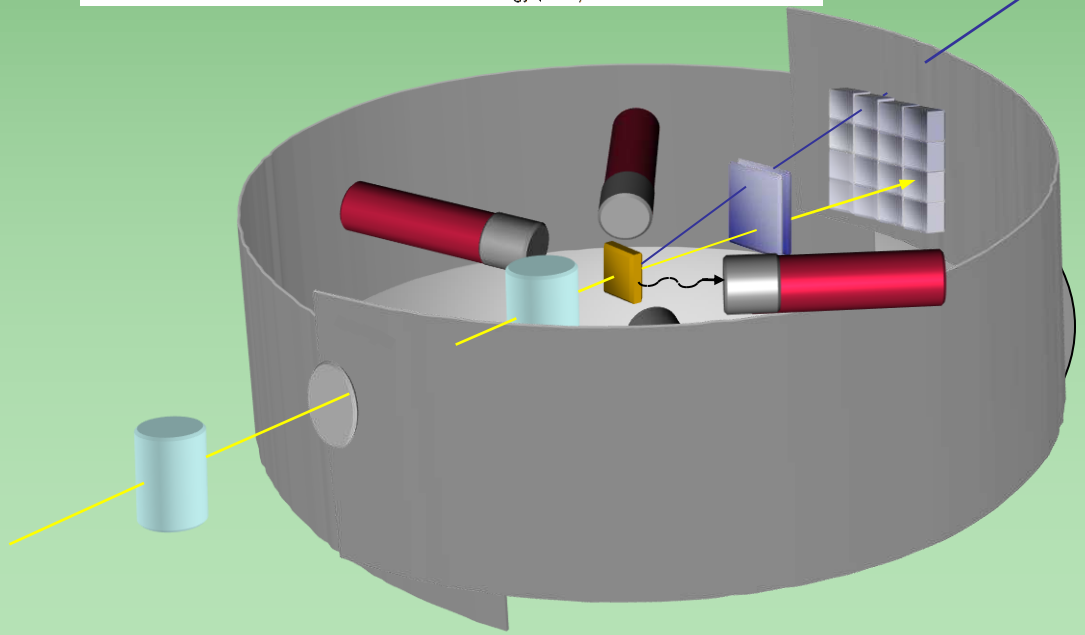
We might run a code to calculate something like a differential cross section



More complex functions

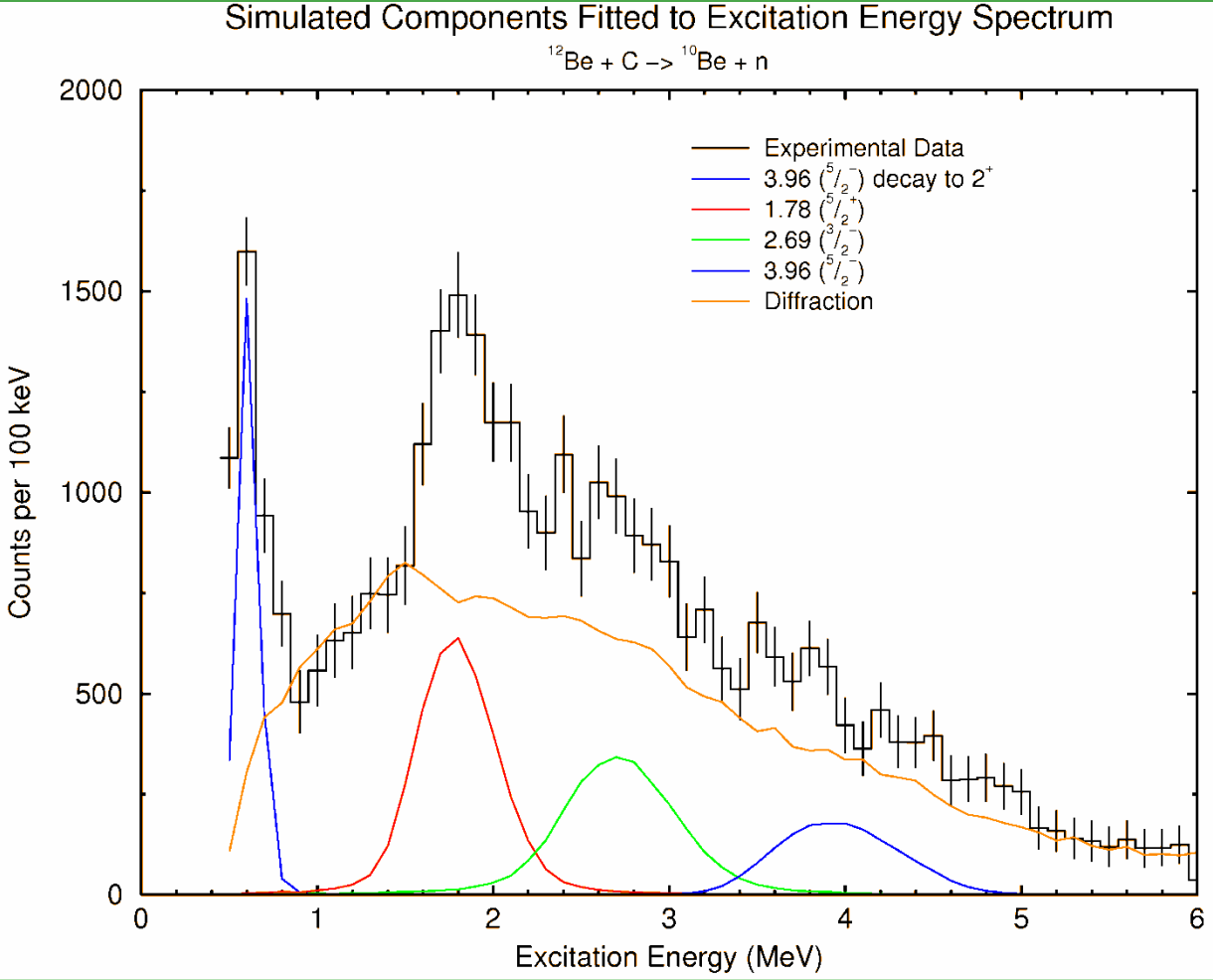






More complex functions

We might have a number of line-shapes we wish to fit to a spectrum



More complex functions

What if you know the function for the fit, but it is not straightforward to deal with it analytically?

$$y = f(x; a, b)$$

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(y_i - f(x; a, b)\right)^2}{\sigma_i^2}$$

We can then write a code to vary the parameters a and b, until we find a minimum for χ^2

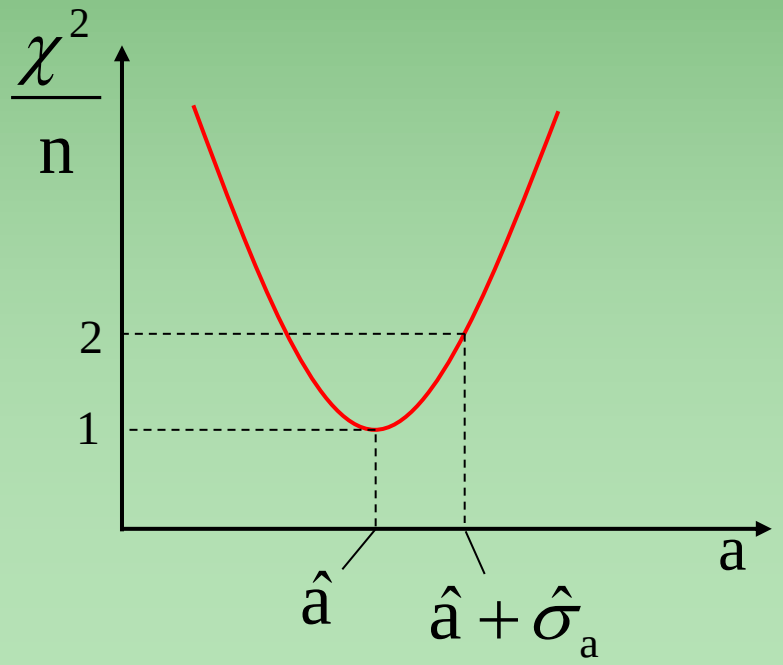
There are routines which will do this efficiently (though there is often associate risk of finding local minima). In the most general case, you can simply do a grid-search over your parameters

What about our uncertainties in our parameters from our χ^2 fit?

Let's return to our case where we have a function of just one parameter

$$f(x;a) = ax$$

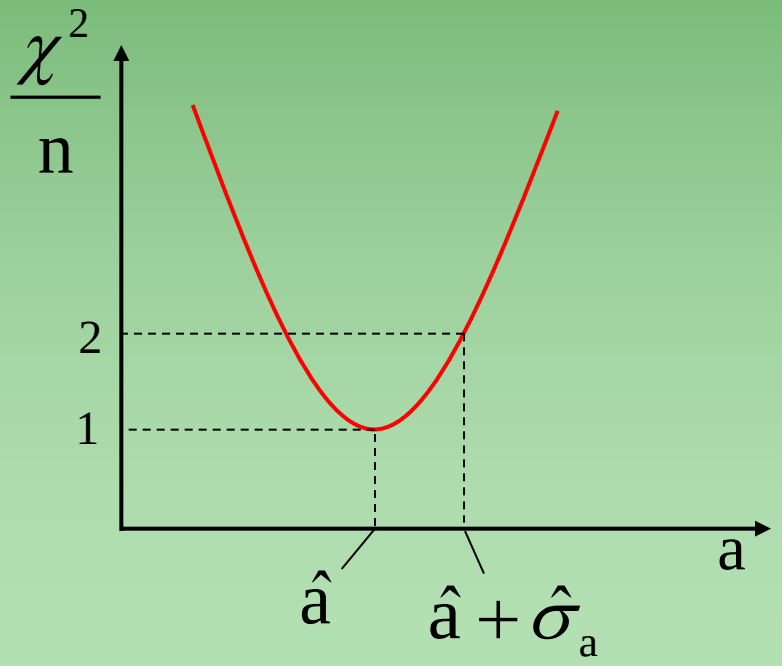
For a good fit, our χ^2 per degree of freedom should be ≈1. If we plot this, we should get



What about our uncertainties in our parameters from our χ^2 fit?

What if we have a function of 2 or more parameters?

$$f(x; a, b)$$



We now have a two-dimensional problem. One solution is the following. Find the best-fit parameters a and b, by letting them both vary independently. Then, to find the uncertainty in parameter a, vary it 'manually' whilst letting b minimise, until you find the minimum χ^2+1 value. This is your σ_a. Repeat the process, but varying b manually, and leaving a free, to get σ_b.

Summary

- Least Squares fitting
  - Specific example of maximum likelihood
  - Describe our data with a function
  - Define χ^2 to represent the difference between the data and the function
  - Extrapolation
  - Residuals
  - χ^2 per degree of freedom
  - Estimating uncertainties in parameters

Extrapolation/Interpolation

Suppose we have determined the slope and intercept of the linear fit, and we now wish to determine some value of y, Y at a given x, X. Obviously, the value is easily calculated

$$Y = mX + c$$

The uncertainty is given by

$$V(Y) = V(\hat{c}) + X^2 V(\hat{m}) + 2X\,\mathrm{cov}(\hat{m}, \hat{c})$$

The covariance term in this expression is important, and tends to reduce the uncertainty. If $\bar{x}$ is zero, then the covariance term is also zero. Alternatively, one can calculate parameters that force this to be the case

$$y = \hat{m}(x - \bar{x}) + \hat{c}'$$

Extrapolation/Interpolation

In this case, the variance on $\hat{m}$ is as before

$$V(\hat{m}) = \frac{\overline{\sigma^2}}{N(\overline{x^2} - \bar{x}^2)}$$

But the variance on $\hat{c}$ is now

$$V(\hat{c}) = \frac{\overline{\sigma^2}}{\sqrt{N}}$$

The uncertainty on an extrapolation is then

$$V(Y) = \frac{\sigma^2 (X - \bar{x})^2}{N(\overline{x^2} - \bar{x}^2)} + \frac{\sigma^2}{N}$$

UWS UNIVERSITY OF THE WEST of SCOTLAND

Introductory Data Analysis

Lecture 6

Maximum Likelihood and Log Likelihood methods

SUPA Graduate School 2010
Prof. A. Andreyev
Andrei.Andreyev@uws.ac.uk

Nuclear Physics Research Group
School of Engineering and Science

Lecture 6 outline

- Reminder: Least squares method (LSM)
- Fitting low-statistics data with LSM: problems
- Maximum likelihood method
- Log-likelihood method

Very useful:
http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_likelihood.htm
http://www-structmed.cimr.cam.ac.uk/Course/Likelihood/likelihood.html

Reminder: Least Squares Fitting

- The data are a set of **precise values of x** $\{x_1 + x_2 + x_3 \ldots + x_N\}$

for which there are a corresponding set of measurements of $y \quad \{y_1 + y_2 + y_3 \ldots + y_N\}$ with precision σ_i

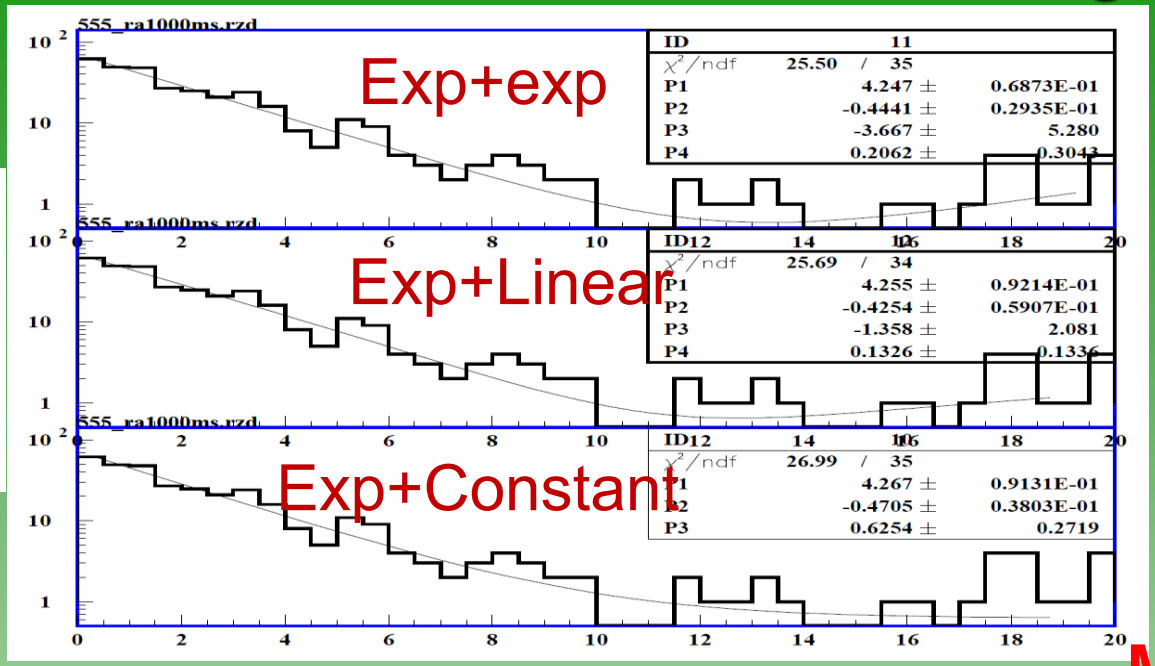
- The least squares method requires the calculation and minimization of a quantity χ^2, which represents **the total deviation of data** y_i from **the expected values** given by $f(x;a)$

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i;a)}{\sigma_i} \right]^2 \qquad \chi^2 = \sum_{i=1}^{N} \left[\frac{y_i^{\text{measured}} - y_i^{\text{ideal/expected}}}{\text{estimated uncertainty}} \right]^2$$

If all the data have the same σ_i then this simplifies to:

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^{N} \left[y_i - f(x_i;a) \right]^2$$

From Lec. 3: Fit with Least-squares method and check χ^2 value for different fitting functions



$T_{1/2}$=1.56 ms

$T_{1/2}$=1.63 ms

$T_{1/2}$=1.47 ms

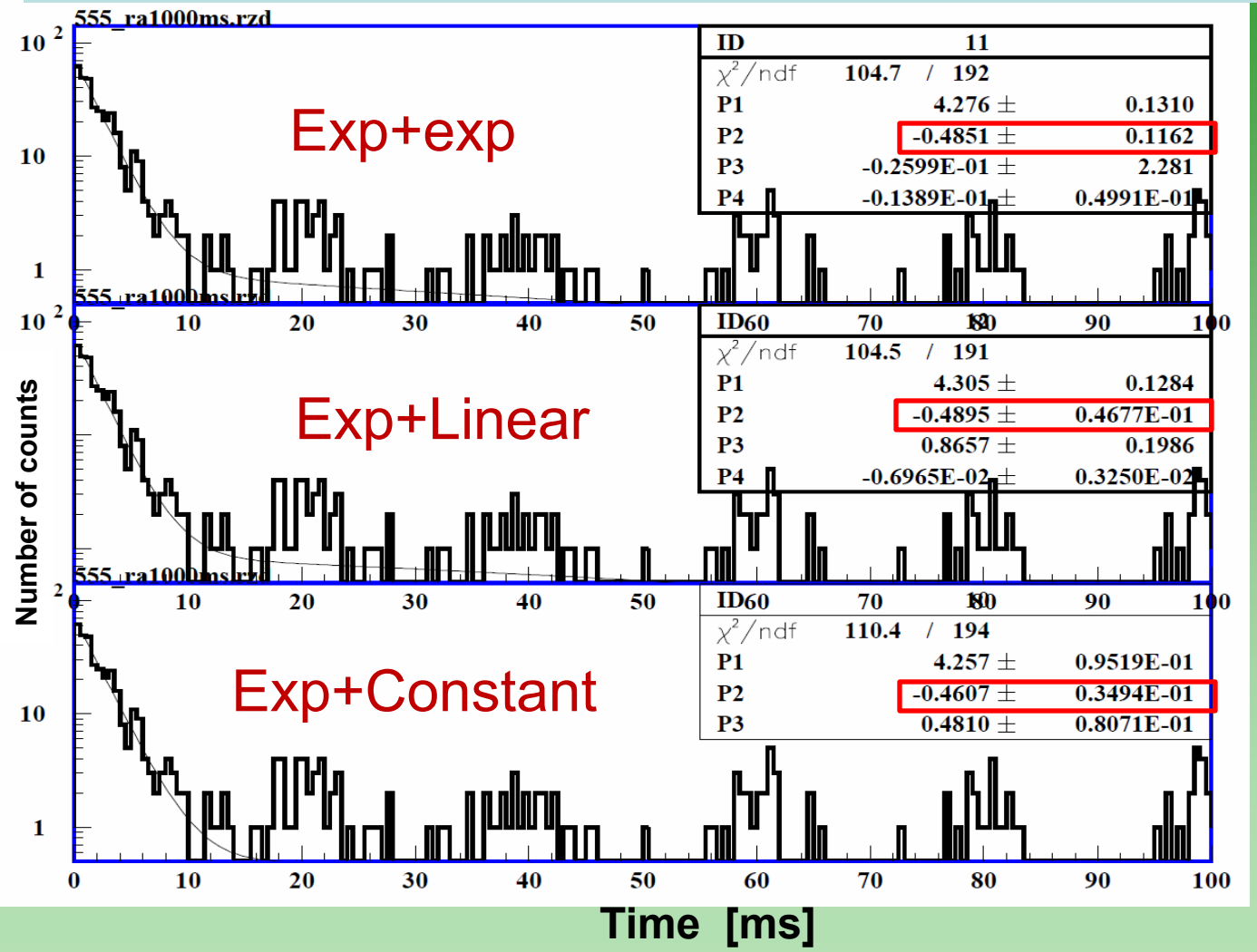
Max Difference 0.16 ms

Why different $T_{1/2}$ values?

- One of the reasons is that, in this case, the time **interval of 0-20 ms is too short to determine the background's behaviour with the necessary precision**
- **Need a longer interval for fitting!?**

Try a longer time interval (0-100 ms)

Now, try the “same” data, but **in the larger range** (up to 100 ms) – this should make the ‘background’ determination more ‘stable’



$T_{1/2}=\ln 2/P2$

$T_{1/2}=1.43$ ms

$T_{1/2}=1.42$ ms

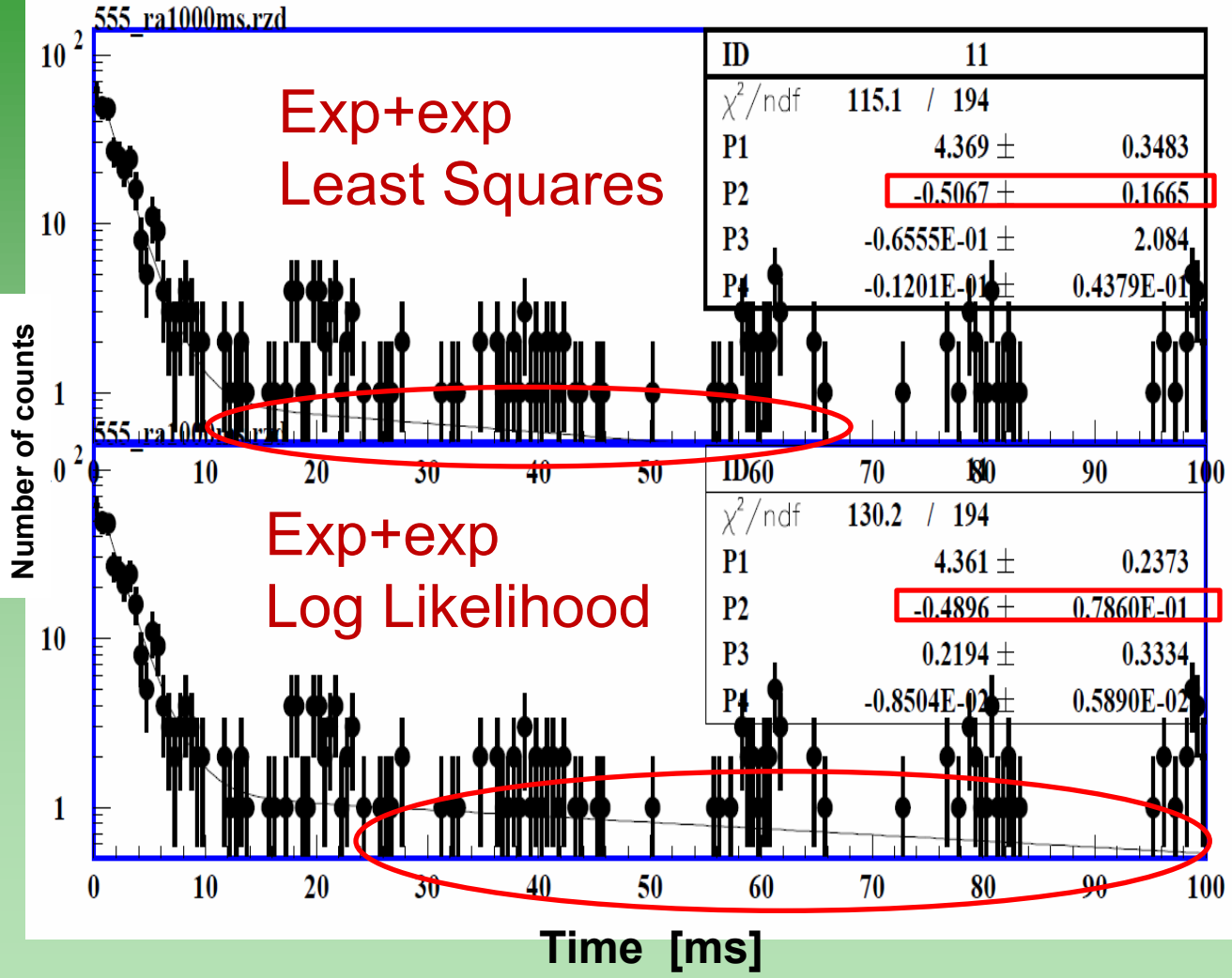
$T_{1/2}=1.50$ ms

Max Difference now is 0.08 ms!

It seems, the results are now much closer to each other!

Least Squares vs Log. Likelihood

Now, the “same” data (range 0-100 ms), fit with 2 exponents, but **using either Least Squares or Log Likelihood method**



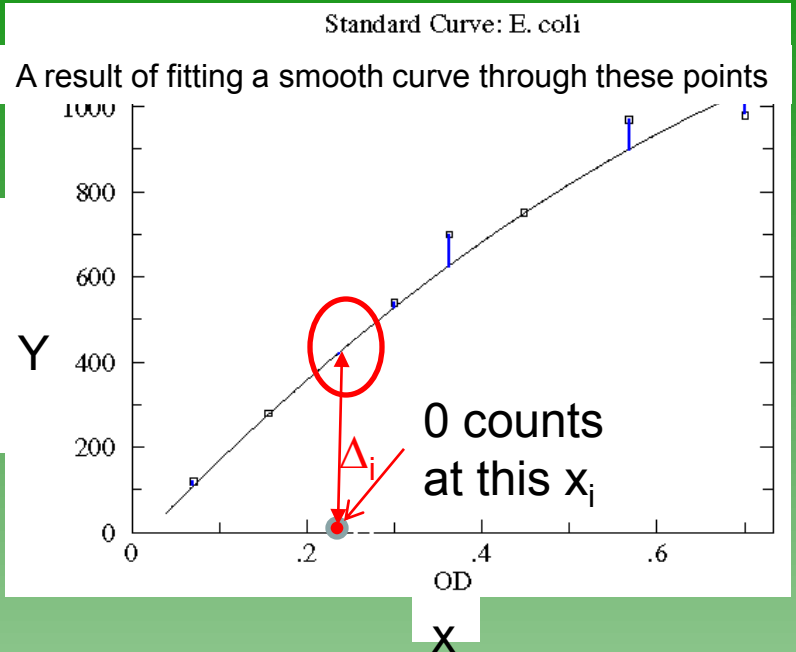
$T_{1/2}=\ln 2/P2$

$T_{1/2}=1.37(35)$ ms

$T_{1/2}=1.42(25)$ ms

- **Difference now is 0.05 ms**
- **Uncertainty is smaller in case of log-likelihood**
- **Higher level for exp2!**

Problems for LS fit for low statistics data



$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i; a)}{\sigma_i} \right]^2$$

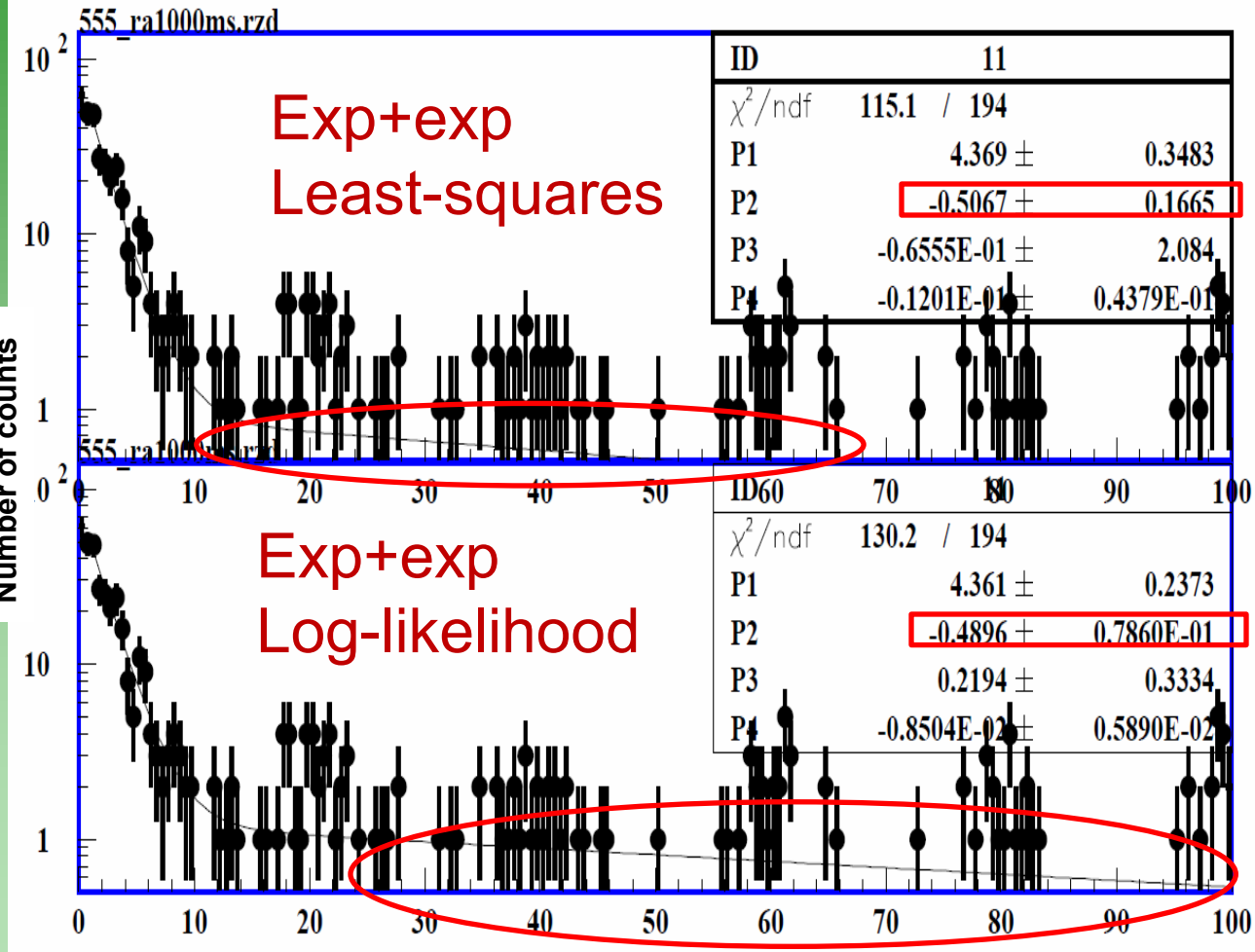
Thus, χ^2 is defined as the sum of the square of each data point's ratio of deviation to uncertainty

$$\chi^2 = \left(\frac{\Delta_1}{\sigma_1}\right)^2 + \left(\frac{\Delta_2}{\sigma_2}\right)^2 + \left(\frac{\Delta_3}{\sigma_3}\right)^2 + \cdots + \left(\frac{\Delta_N}{\sigma_N}\right)^2$$

- Now, imagine that at some x_i, no counts were observed
- First problem: what do we do about σ_i at this x_i value?
- Usual solution, - assume, **an upper limit of number of counts =1**, thus $\sigma_i=1$, thus 1 ± 1 counts at this x_i
- Second problem: at this x_i value, one gets **Δ_i- large!**
- All this leads to un artificial increase of χ^2 value!
- As a result, the fitting procedure will try to 'lower' the fitting curve to compensate for this effect (to make Δ_i lower at this point)

Least Squares vs Log. Likelihood

Now, the “same” data (range 0-100 ms), fit with 2 exponents, but **using either Least Squares or Log Likelihood method**



$T_{1/2}$=ln2/P2

$T_{1/2}$=1.37**(35)** ms

$T_{1/2}$=1.42**(25)** ms

- **LSM does NOT ‘like’ zero data**
- **LSM is not good for low statistics’s data**

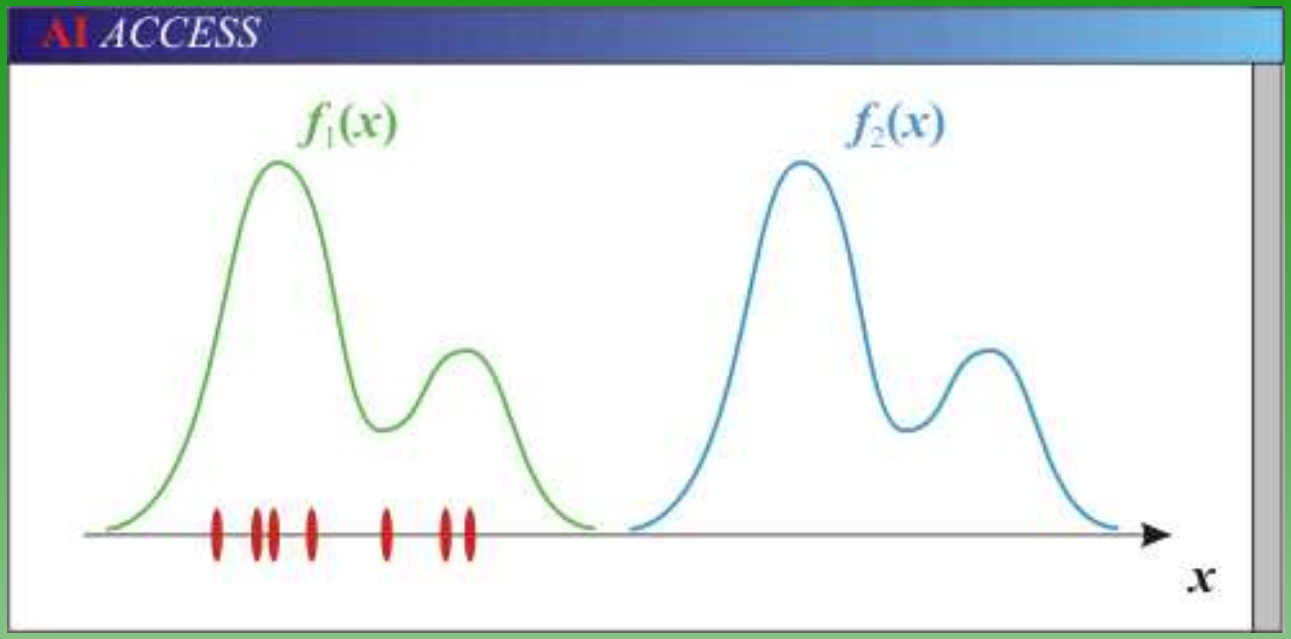
Maximum Likelihood Method

- The data are a set of **precise values of x** $\{x_1 + x_2 + x_3 \ldots + x_N\}$

for which there are a corresponding set of measurements of $y \quad \{y_1 + y_2 + y_3 \ldots + y_N\}$ with precision σ_i

Maximum likelihood method, is a procedure of finding the value of one or more parameters for a given statistic which makes the ***known*** **likelihood distribution a maximum**.

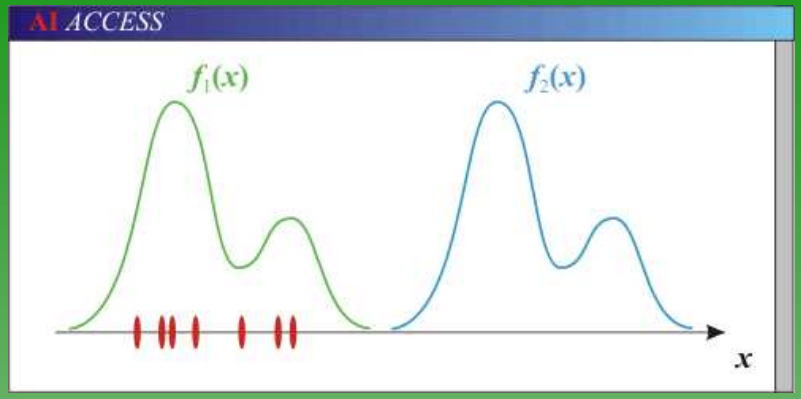
Maximum Likelihood Method



The illustration above shows a sample of n independent observations, and **two continuous distributions** (e.g., density distribution) $f_1(x)$ and $f_2(x)$, with $f_2(x)$ being just $f_1(x)$ translated by a certain amount.

- Of these two distributions, which **one is the most likely to have generated the sample** ?
- Clearly, the answer is $f_1(x)$, (but we would like to formalize this intuition.

Likelihood of the distribution



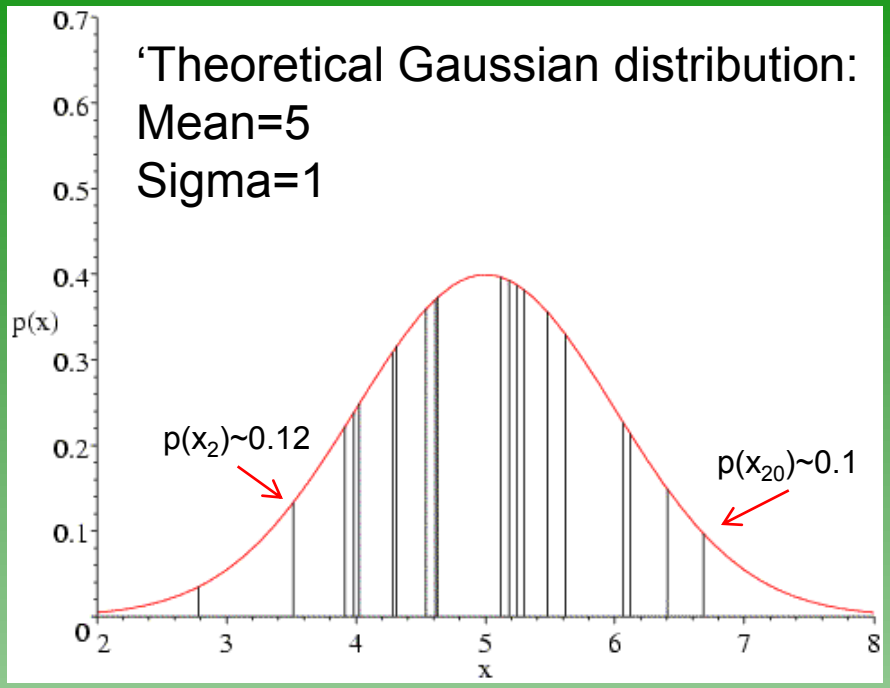
To quantify this intuitive judgement:
for any probability distribution $f(x)$, just **multiply the values of $f(x)$ for each of the observations** of the sample, denote the result L, and call it the **likelihood** of the distribution $f(x)$ for this particular sample :

$$\textbf{Likelihood} = L = \prod_i f(x_i) \qquad i = 1, 2, \ldots, n$$

- **Clearly, the likelihood L can have a large value only if all the observations are in regions where $f(x)$ is not very small.**

- **http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_likelihood.htm**

Example: Maximum Likelihood Method



•In this figure, the Gaussian probability distribution is plotted for a mean of 5 and standard deviation of 1.

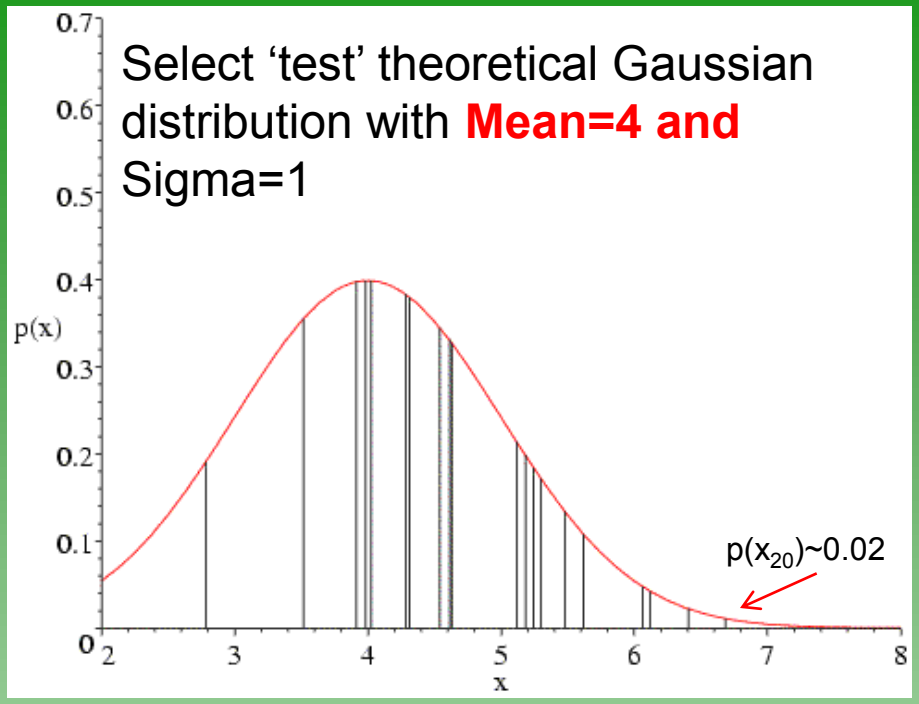
•The twenty vertical bars correspond to the twenty 'measured' data points (here derived from the above-mentioned theoretical distribution!)

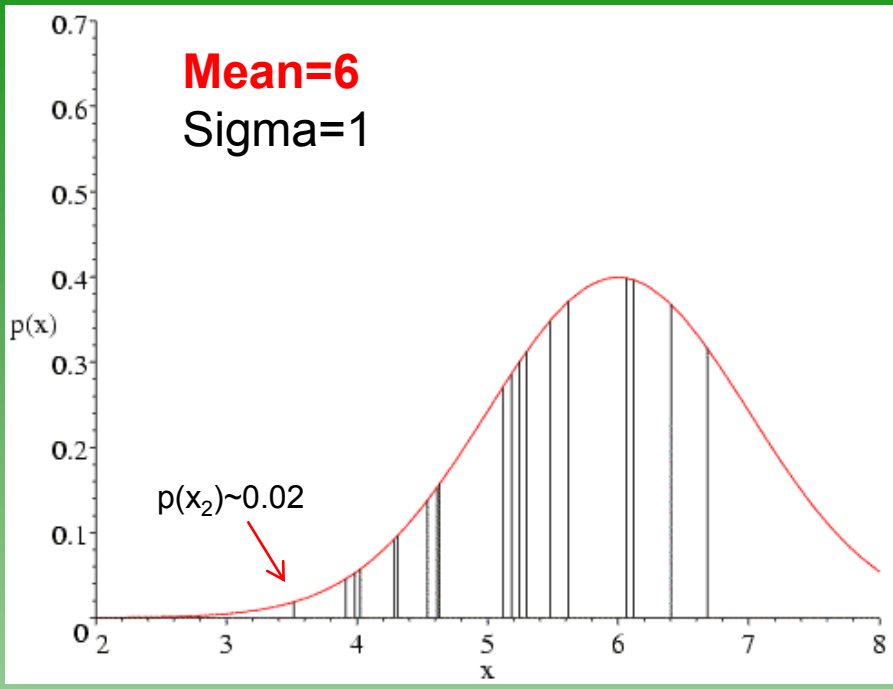
•The height of each bar represents the probability **$p(x_i)$** of that measurement, given the assumed mean and standard deviation. (e.g. $p(x_2)$~0.12, $p(x_{20})$~0.1)

•The likelihood function is the product of all those probabilities (heights) $L=\prod p(x_i)$

•None of the probabilities **$p(x_i)$** is particularly low, and they are highest in the centre of the distribution, which is most heavily populated by the data.

Example: Maximum Likelihood Method

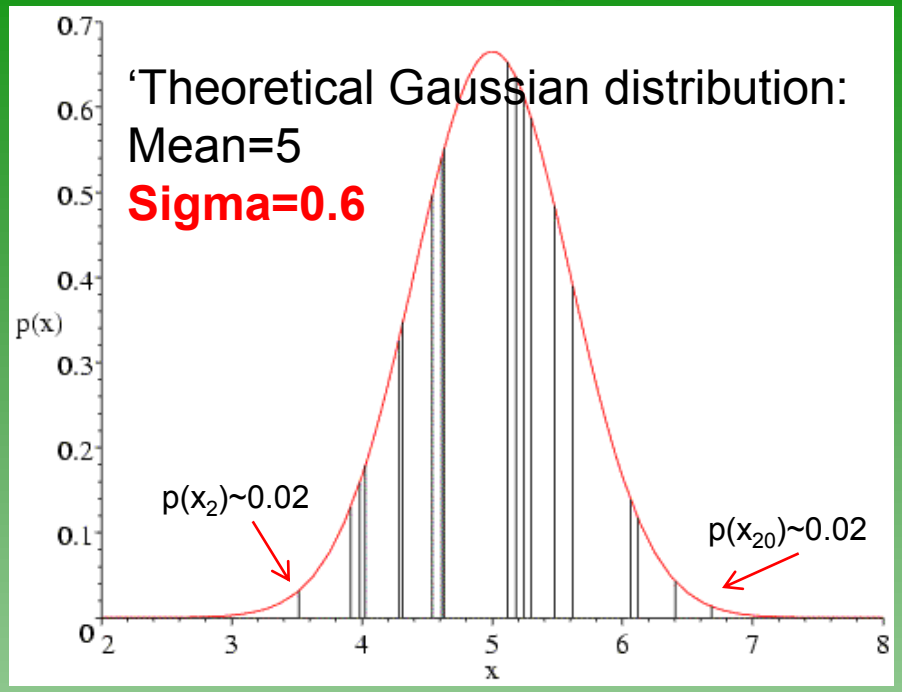


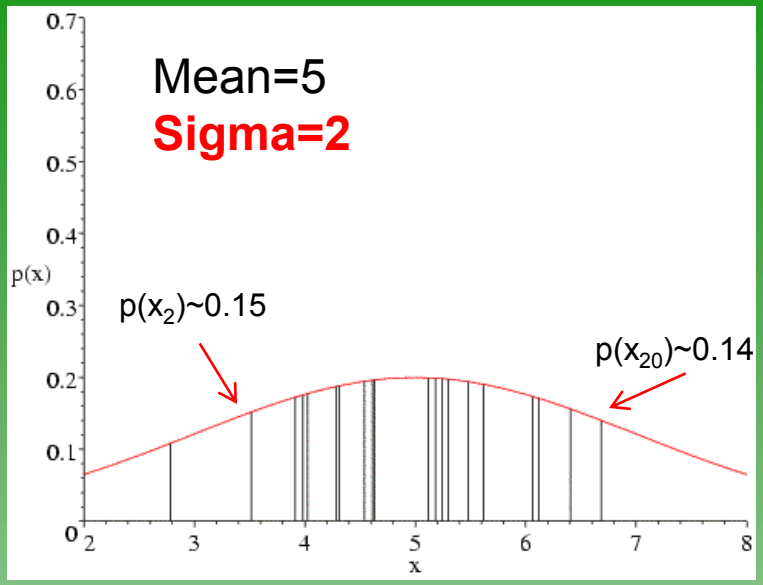


- if we **change the mean to 4**, **the data points on the high end of the distribution become very improbable (e.g.** $p(x_{20})\sim0.02$), which will **reduce the likelihood function significantly**.
- Also, **fewer of the data points are now in the peak region.**

The same sort of thing happens if **we change the mean to 6** (here $p(x_2)\sim0.02$ instead of $p(x_2)\sim0.12$ earlier)

Example: Maximum Likelihood Method

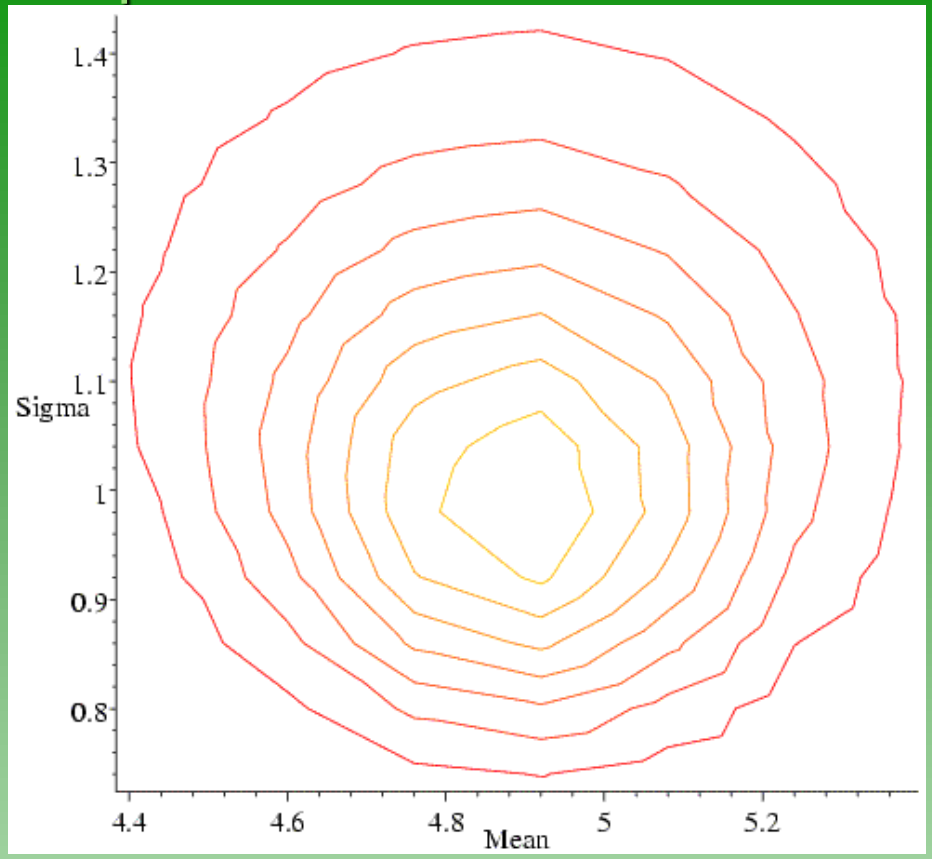




If we have **the correct mean but the wrong standard deviation?**
In this figure, σ=0.6.
In the heavily populated centre, the probability values go up, **but the values in the two tails go down even more**, so that the **overall value of the likelihood is reduced.**

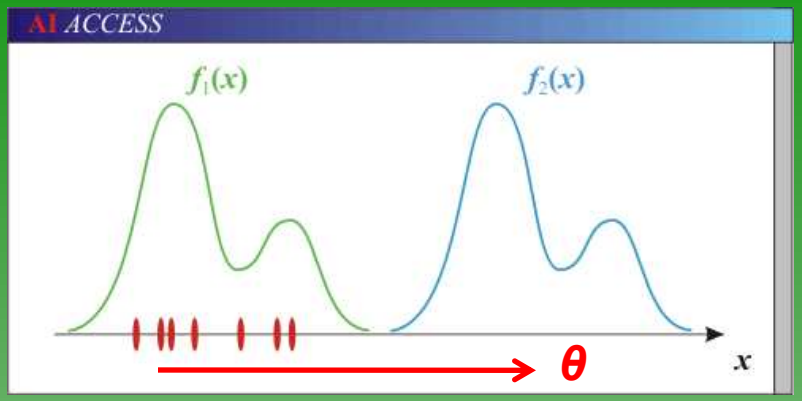
Similarly, if we choose too high a value for the standard deviation (two in this figure), **the probabilities in the tails go up**, but the **decrease in the heavily-populated centre means that the overall likelihood goes down.**

Example: Maximum Likelihood Method



- We can carry out a **likelihood calculation for all possible pairs** of mean and standard deviation, shown in the following contour plot.
- The peak in this distribution is close to the mean and standard deviation that were used in generating the data from a Gaussian distribution

Maximum Likelihood Method



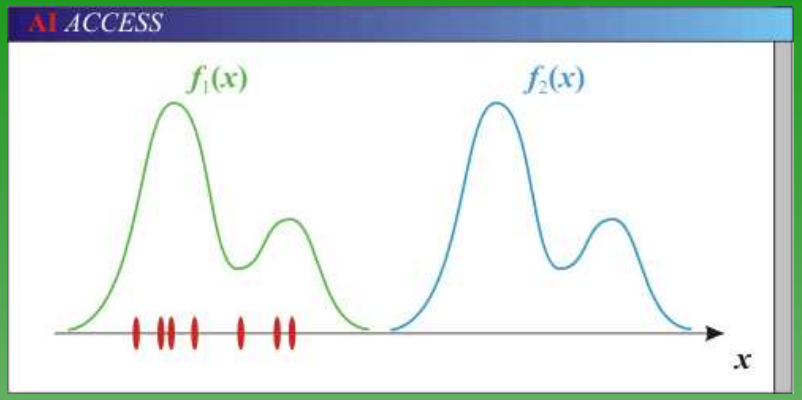
Likelihood = $L =: \Pi_i f(x_i)$
$i = 1, 2, ..., n$

- *Recall, that* $f_1(x)$ and $f_2(x)$ are assumed to belong to a family of distributions, all identical in shape and **differing only by their position horizontal position θ along the x axis**

- General question: which position θ of the generic distribution $f(x)$ gives **the largest likelihood**?

- If such θ found, then one may consider this value of θ as being **probably fairly close to the true (and unknown) value θ_0** of the parameter of the distribution that actually generated the sample (thus, we found an **estimator**)

Maximum Likelihood Method



Likelihood = $L =: \Pi_i f(x_i)$
$i = 1, 2, ..., n$

- Therefore, **the concept of likelihood leads to a method of parameter estimation**. The method consists in retaining as an estimate of θ_0 the value of θ resulting in the largest possible value of the sample likelihood.
- This method is thus called **Maximum Likelihood estimation**, which is, in fact, **the most powerful and widely used method of parameter estimation these days.**
- **An estimator θ^*** obtained by maximizing the likelihood of a probability distribution is called a **Maximum Likelihood estimator** and is usually denoted "MLE".
- The likelihood depends on both the sample $\mathbf{x} = \{x_i\}$ and the parameter θ, thus usually denoted as $L(\mathbf{x}, \theta)$.

Log- Likelihood Method

- The likelihood is defined as a product, **and maximizing a product is usually more difficult than maximizing a sum** (also the product of a lot of small numbers may be too small to represent on a computer)
- But if a function $L(\theta)$ is changed into a new function $L'(\theta)$ by a **monotonously increasing transformation**, then $L(\theta)$ and $L'(\theta)$ will clearly reach their maximum values for the **same** value of θ.
- In particular, if the monotonous transformation is **logarithmic**, maximization of a product is then turned **into an easier maximization of a sum (as** $\log_b(x \cdot y) = \log_b(x) + \log_b(y)$)

- The logarithm of the likelihood is called the **log-likelihood**, and will be denoted log-L. So, by definition :

Log-likelihood = **log-*L*** = : $\Sigma_i \log(f(x_i))$ $\qquad i = 1, 2, ..., n$

Note: the likelihood and log-likelihood reach their extrema for the same values of θ.

Maximasing Log- Likelihood

From high school:
to find a maximum of likelihood (log likelihood), we need to do:

$$\frac{d}{d\theta} L(\theta) = 0$$

•Even though most classical likelihoods are differentiable, there is no reason why the solutions of this equation should have simple analytical forms.

•As a matter of fact, more often than not, they don't, and it may then be necessary **to resort to computer numerical techniques to identify the extrema of the likelihood function**

Maximazing Log- Likelihood

$$\frac{d}{d\theta} L(\theta) = 0$$

- The above equation identifies extrema of $L(\theta)$, but says nothing about which of these extrema **are maxima (that we are interested in)** and which are minima (that we are not interested in).

- So, after the solutions of the equation have been identified, one must go through these solutions **to retain only those corresponding to maxima**

- There is no reason why the likelihood **would have a single maximum**. So once the maxima have been found, only the largest among them is retained

- Computer optimization techniques can be adjusted so as to identify only maxima (using the second derivative).

Summary: Maximazing Log- Likelihood

Imagine, we have a continuous random variable *x* which gives:

$$f(x;\theta_1,\theta_2,...,\theta_k)$$

where $\theta_1, \theta_2, ..., \theta_k$ are *k* unknown constant parameters which need to be estimated

Then the **likelihood function is given by the product**:

$$L(x_1,x_2,...,x_N|\theta_1,\theta_2,...,\theta_k)=L=\prod_{i=1}^{N} f(x_i;\theta_1,\theta_2,...,\theta_k)$$

$$i=1,2,...,N$$

The logarithmic likelihood function is given by:

$$\Lambda=\ln L=\sum_{i=1}^{N}\ln f(x_i;\theta_1,\theta_2,...,\theta_k)$$

The maximum likelihood estimators (MLE) of $\theta_1, \theta_2, ..., \theta_k$ are obtained by maximizing L or Λ

$$\frac{\partial(\Lambda)}{\partial\theta_j}=0,\quad j=1,2,...,k$$

Illustrating the MLE Method Using the Normal Distribution

To obtain the MLE estimates for the mean $\bar{T}$, and standard deviation, σ_T for the normal distribution**, start with the normal distribution which is given by**:

$$f(T) = \frac{1}{\sigma_T \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{T-\bar{T}}{\sigma_T}\right)^2}$$

If T_1, T_2, ... T_N are known times

$$L(T_1, T_2, ..., T_N | \bar{T}, \sigma_T) = L = \prod_{i=1}^{N} \left[\frac{1}{\sigma_T \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{T_i-\bar{T}}{\sigma_T}\right)^2} \right]$$

$$L = \frac{1}{(\sigma_T \sqrt{2\pi})^N} e^{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{T_i-\bar{T}}{\sigma_T}\right)^2}$$

$$\Lambda = \ln L = -\frac{N}{2}\ln(2\pi) - N \ln \sigma_T - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{T_i - \bar{T}}{\sigma_T}\right)^2$$

Illustrating the MLE Method Using the Normal Distribution

Then taking the partial derivatives of Λ with respect to each one of the parameters and setting it equal to zero yields:

$$\frac{\partial(\Lambda)}{\partial \bar{T}} = \frac{1}{\sigma_T^2}\sum_{i=1}^{N}(T_i - \bar{T}) = 0 \qquad \frac{\partial(\Lambda)}{\partial \sigma_T} = -\frac{N}{\sigma_T} + \frac{1}{\sigma_T^3}\sum_{i=1}^{N}(T_i - \bar{T})^2 = 0$$

Solving these equations simultaneously yields:

$$\bar{T} = \frac{1}{N}\sum_{i=1}^{N} T_i$$

$$\hat{\sigma}_T^2 = \frac{1}{N}\sum_{i=1}^{N}(T_i - \bar{T})^2$$

$$\hat{\sigma}_T = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(T_i - \bar{T})^2}$$

http://www.weibull.com/AccelTestWeb/mle_maximum_likelihood_parameter_estimation.htm