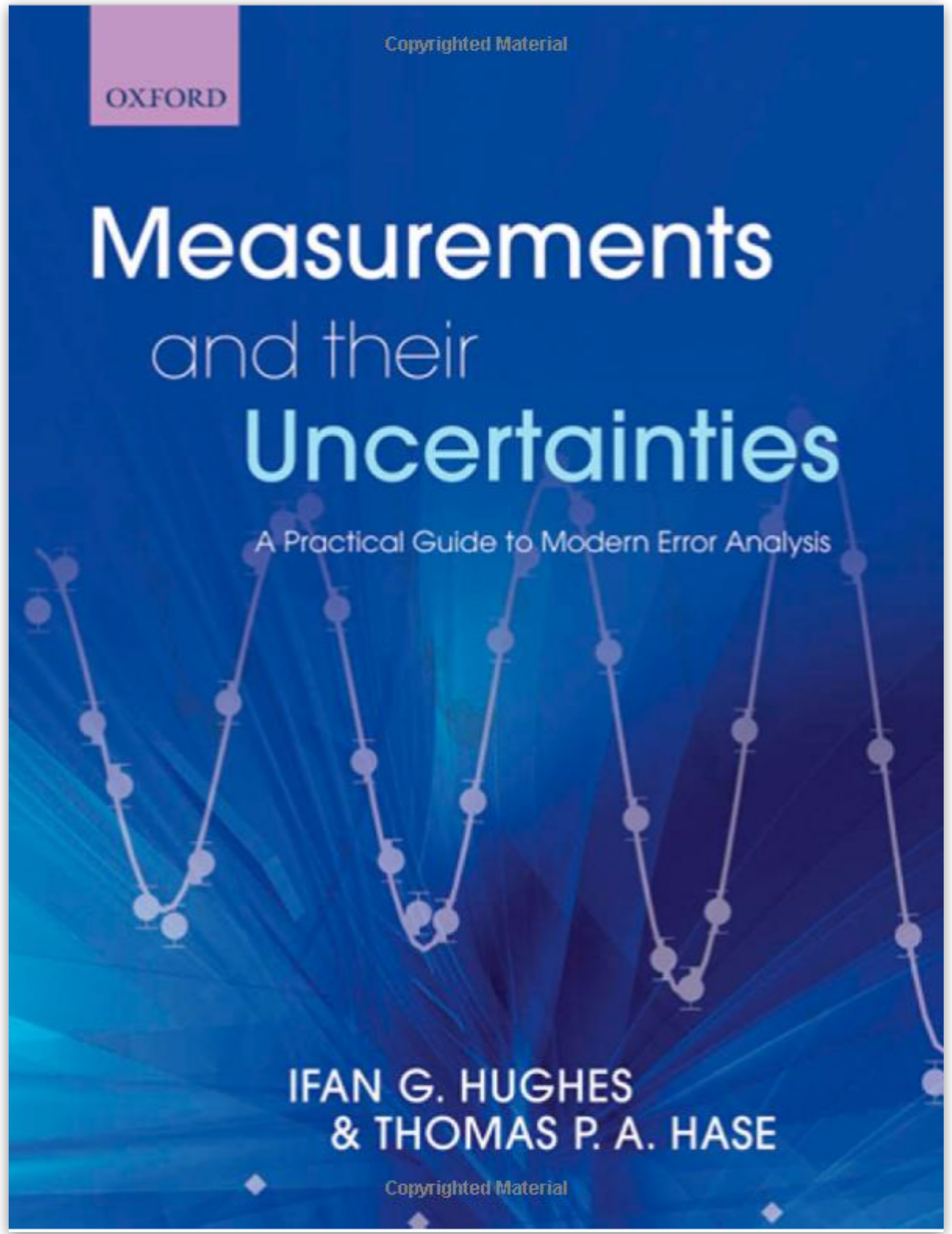
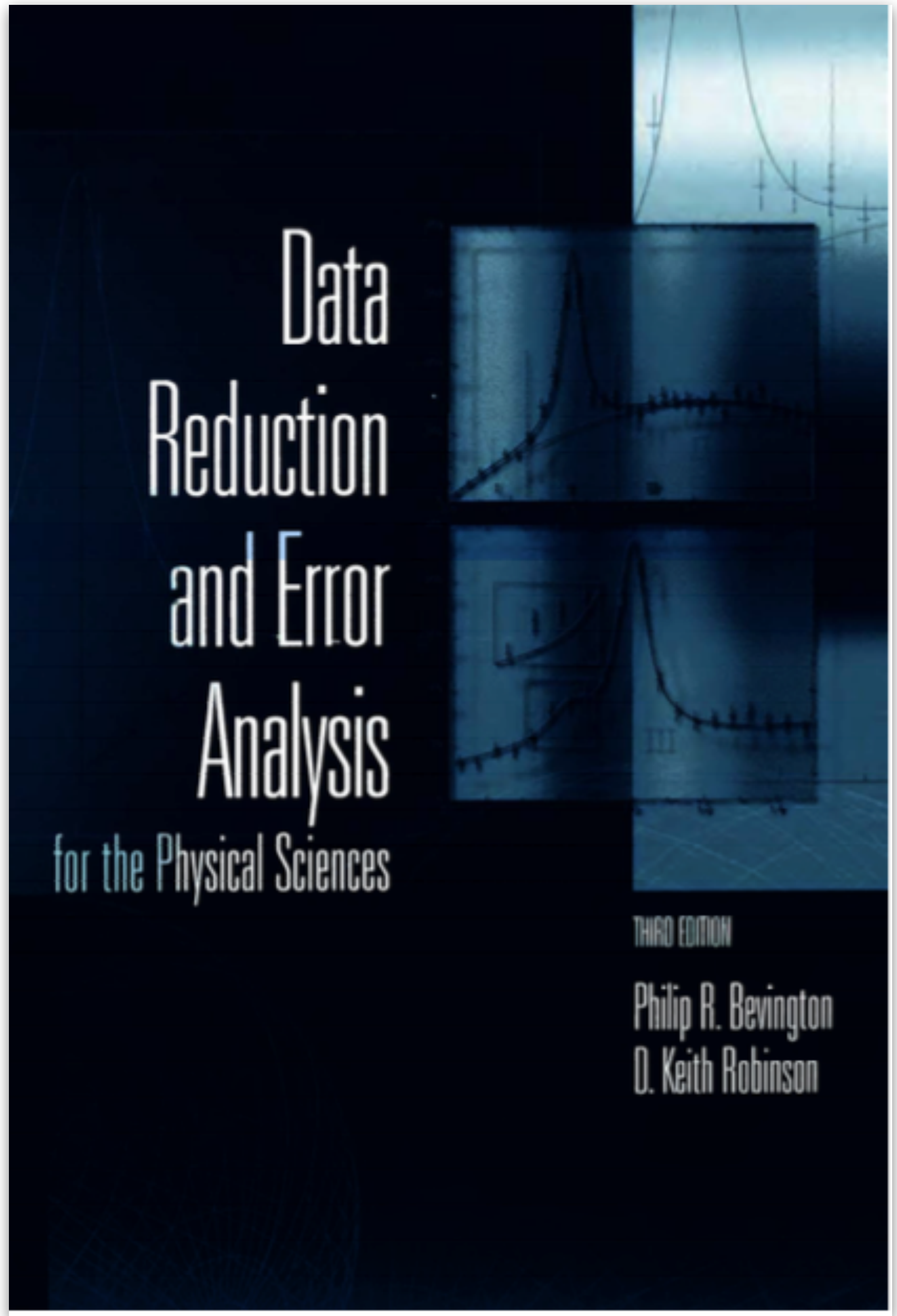
Data Analysis Quickstart

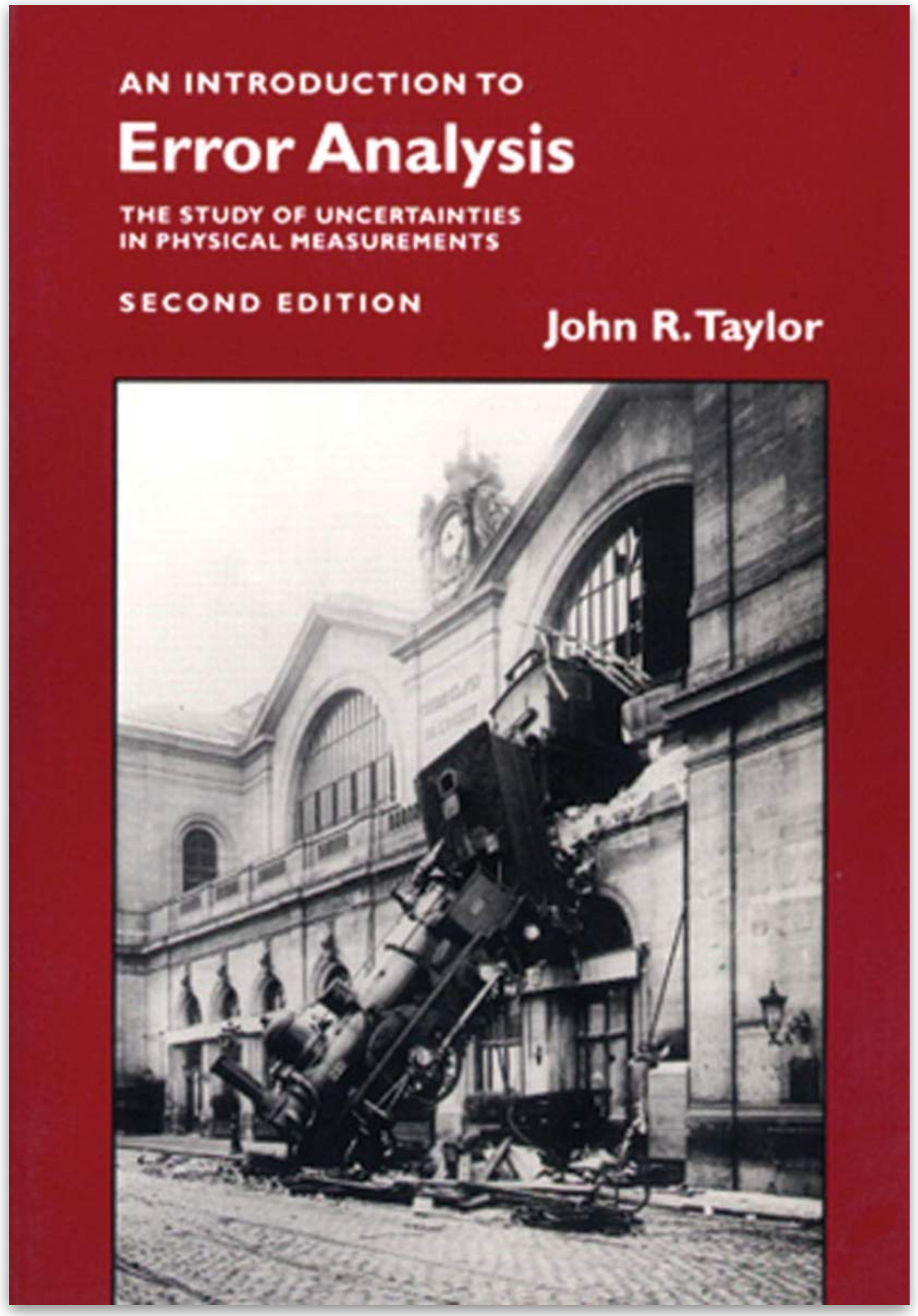
Stage 3 Advanced Labs

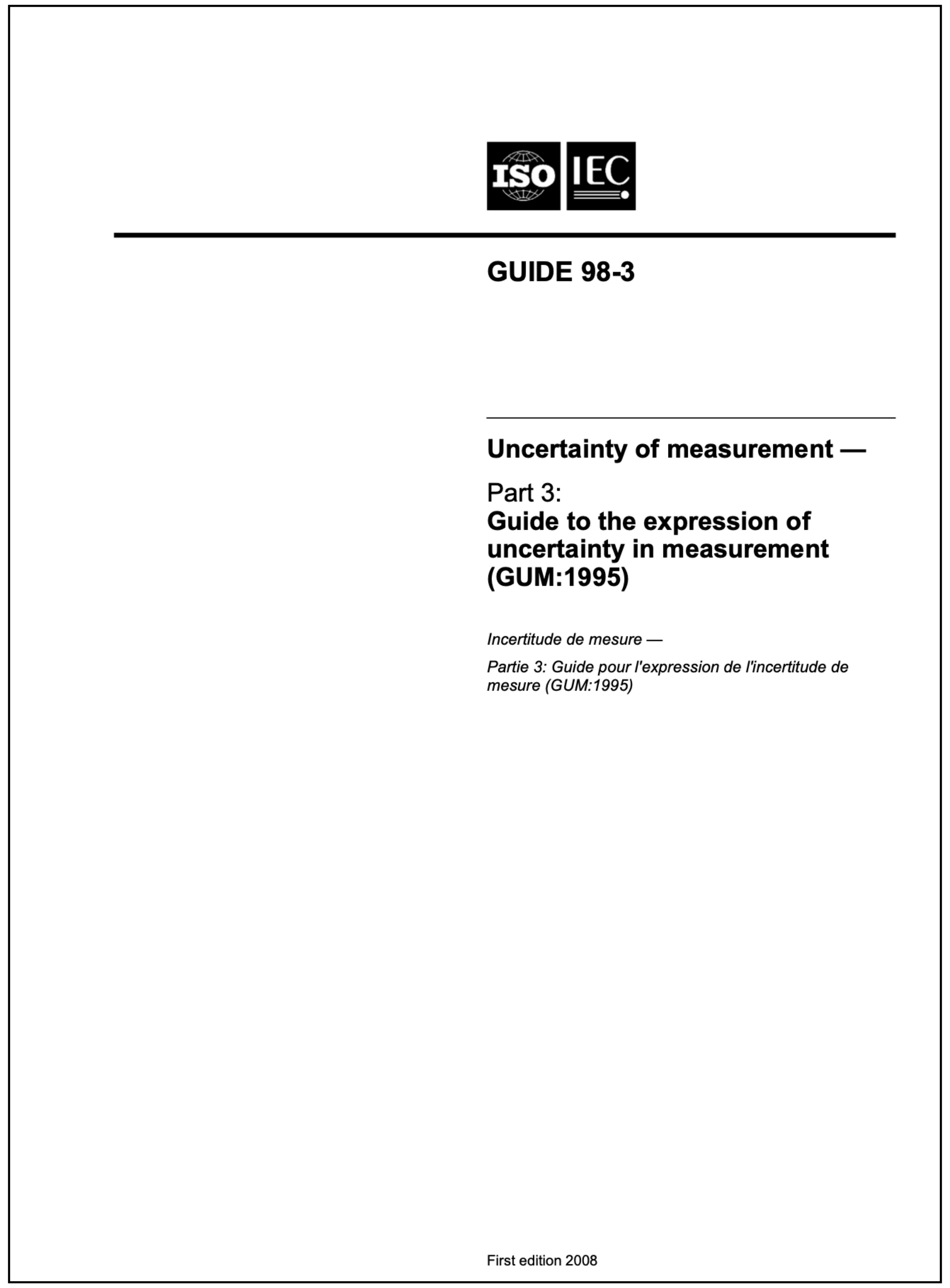
Assoc. Prof. John Quinn

Some Recommended Books







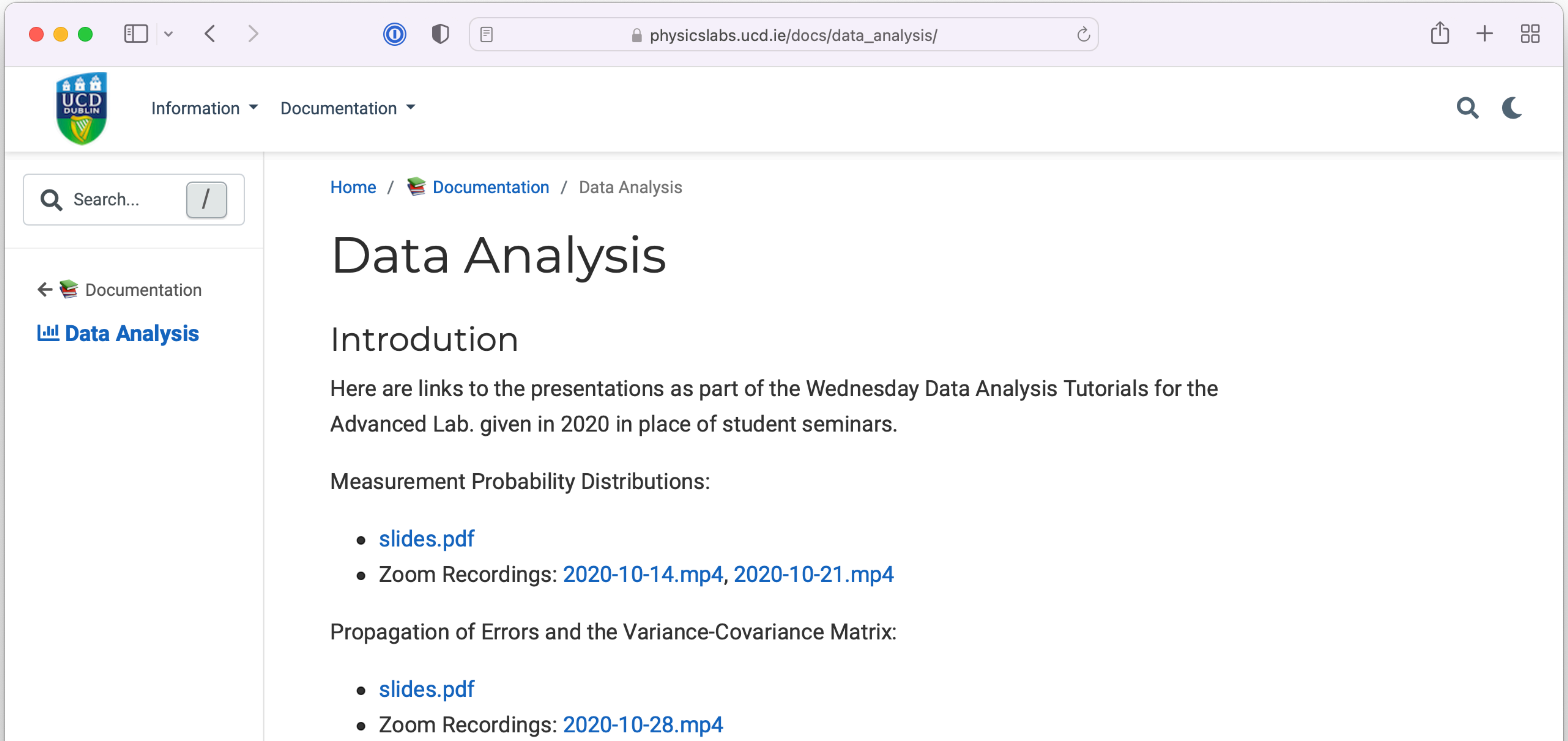


UCD Online:
http://lib.myilibrary.com/Open.aspx?id=273234#

https://www.iso.org/standard/50461.html

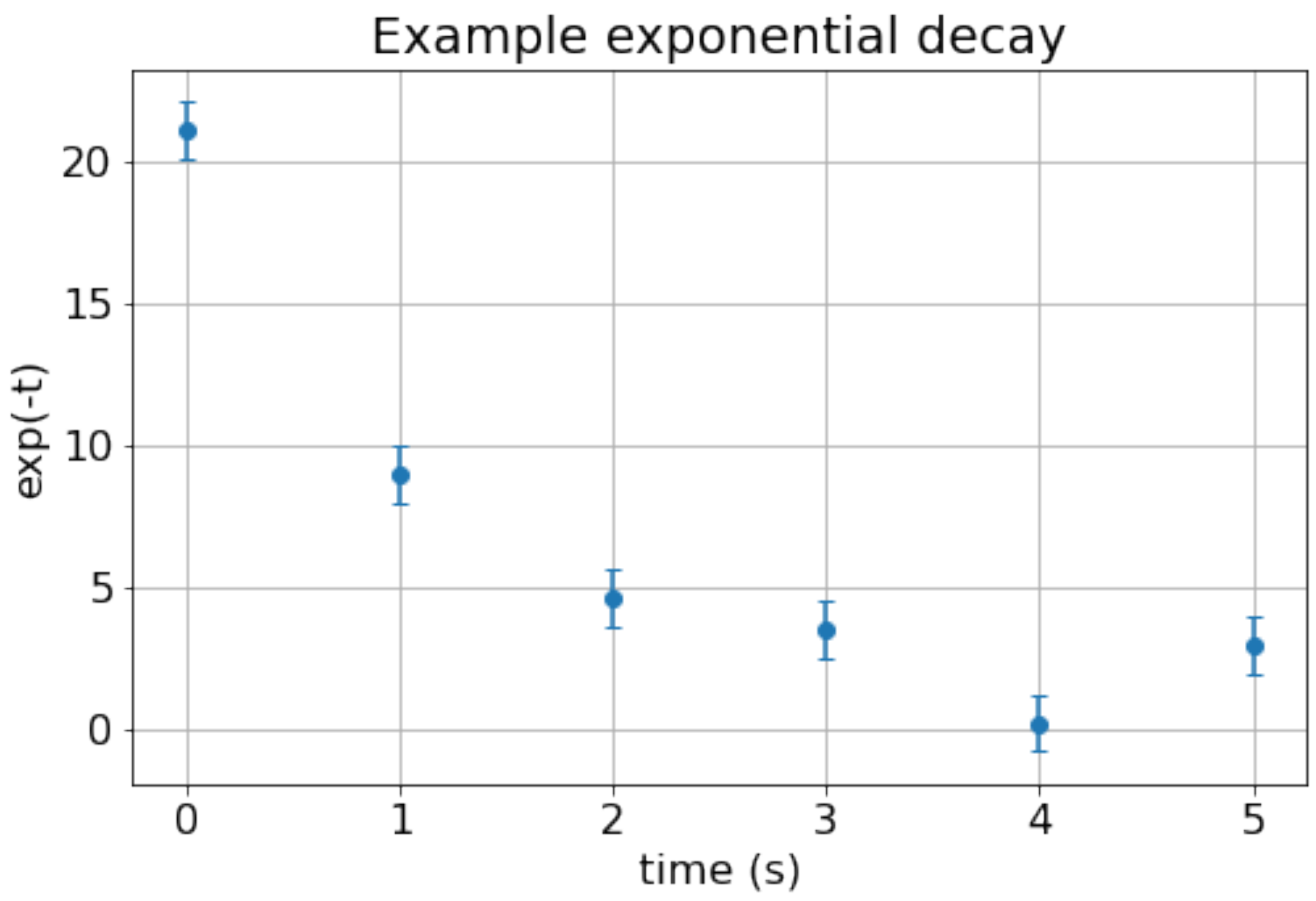
Lockdown Tutorials

- Slides and recordings at: https://physicslabs.ucd.ie/docs/data_analysis/



Graphs and Presentation of Data

- Please make graphs a reasonable size in your report!
- Make sure to:
  - include a title
  - label the axes and include units if appropriate
  - use an appropriate font size so that all text is clearly legible
- Experimental data points should be presented as points, preferably with error bars, and not connected by lines.
- Grid lines may be used.
- LaTeX can be used in titles and labels.
- Save as PDF for inclusion in reports.
- See HowTos.

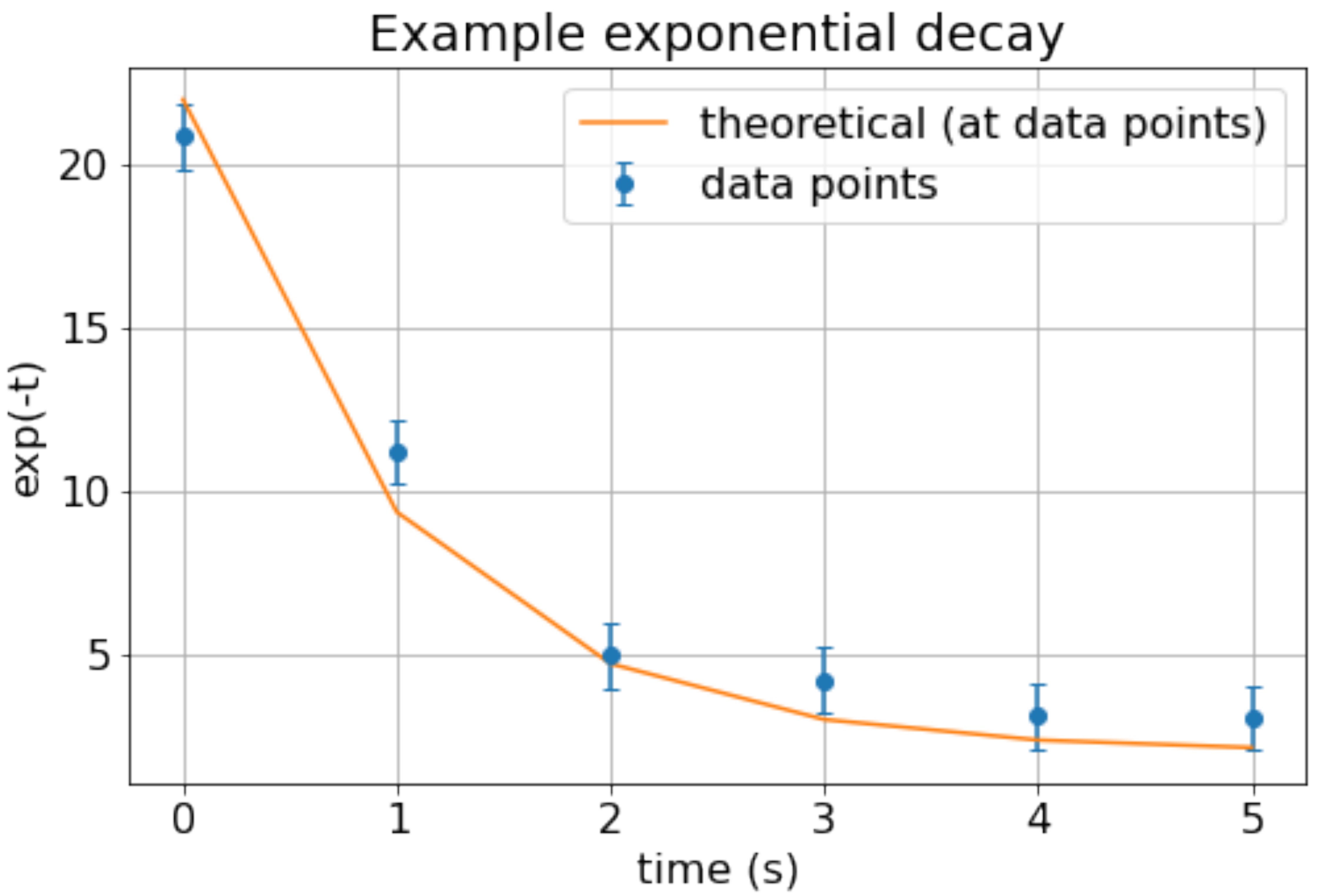


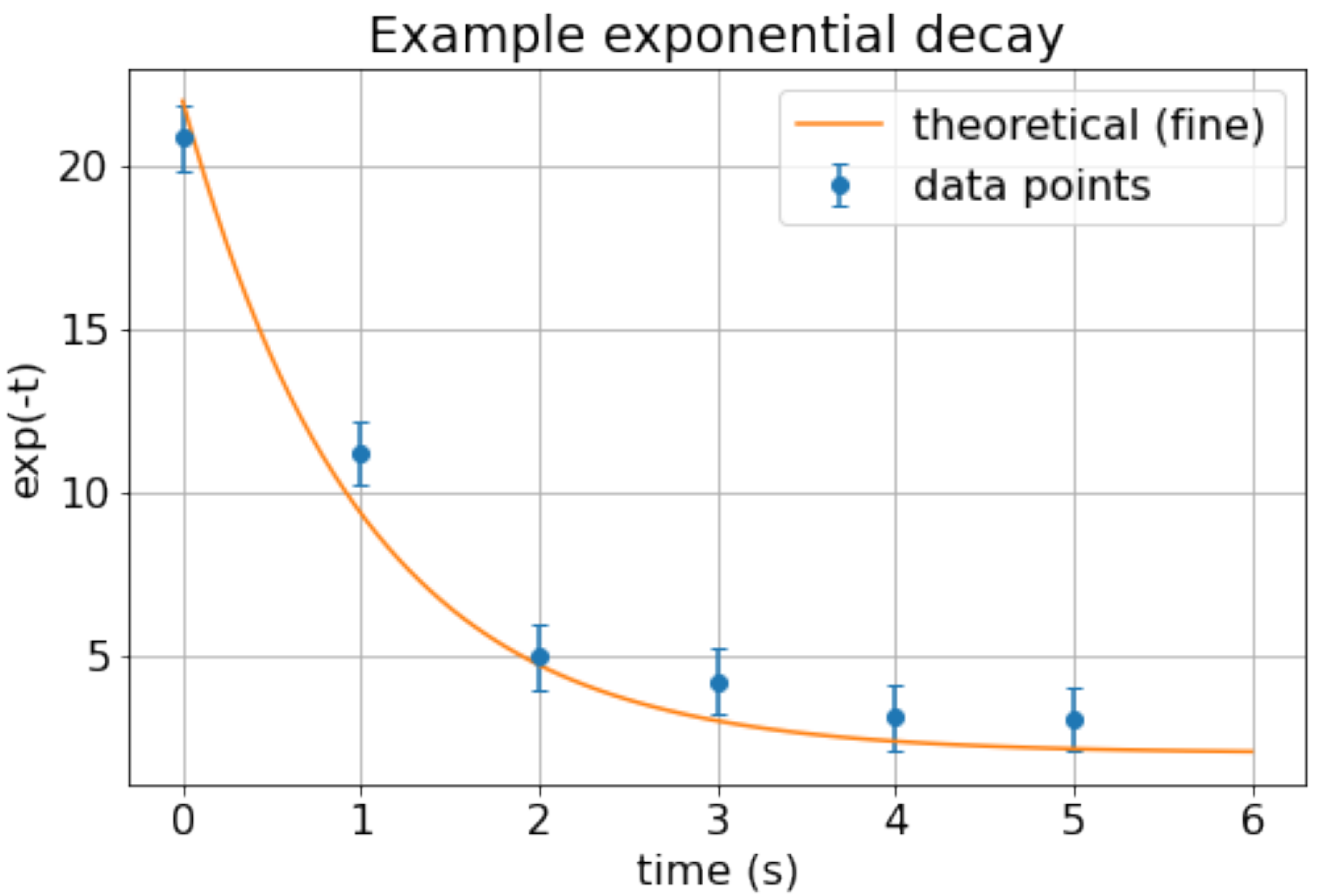
Matplotlib pyplot (plt) - set sizes for notebook:

```
plt.rcParams['figure.figsize'] = (6,4)
plt.rcParams['font.size'] = 12
plt.rcParams['savefig.bbox'] = 'tight'
```

Graphs and Presentation of Data

- Theoretical curves (incl. best-fit) should be overlaid as continuous curves, possibly with a finer sampling than the data points.
- If more than one set of data points/curves use legends to distinguish.





Measurement Errors

- Statistical (or Random) Errors:
  - we cannot measure any physical quantity with infinite precision - there is always some uncertainty on a measurement.
  - natural fluctuations & environment may cause uncertainty beyond the limit of the instrument.
  - when an experiment is repeated several times we find that we do not get the exact same answer each time but that the values fluctuate about some mean
  - We assume random errors average to 0 over many repeated measurements.

- Systematic and Non-Statistical Errors are not random fluctuations but additional uncertainties due to incomplete/imperfect/incorrect knowledge of experiment/calibration etc.
  - Not easy to detect and correct.
  - Generally lumped together into the term "Systematics"

- In general when we quote errors on an experiment they are the Statistical Errors.
  - Non-Statistical/Systematic Errors may be quoted in addition: , e,g. $x = 1.0 \pm 0.1_{stat} \pm 0.2_{sys}$

Measurement Uncertainty

- Measurement Errors:
  - Type A: derive from statistical analysis of repeated measurements
  - Type B: non-statistical, using other information, e.g. from instrument specification.
- Instrumentation Precision:
  - Analogue: The statistical error associated with analogue instrumentation is often due to how well one can read the scale, and it is often up to the experimenter to estimate.
  - Digital Meter: Guide: the precision of a digital meter is limited to the last digit.
    - e.g. repeated measurements of a voltage with a digital multimeter gives 8.41 V. We would thus quote the voltage as 8.41 ± 0.01 V.
  - ADC: resolution limited by number of bits, i.e. range divided by 2^N where N is the number of bits. ADCs round down!
  - Other: instrumentation documentation
- Non-instrumental:
  - noise or other errors may be greater than instrument precision
  - e.g. measure repeated time intervals with a precision stop watch.

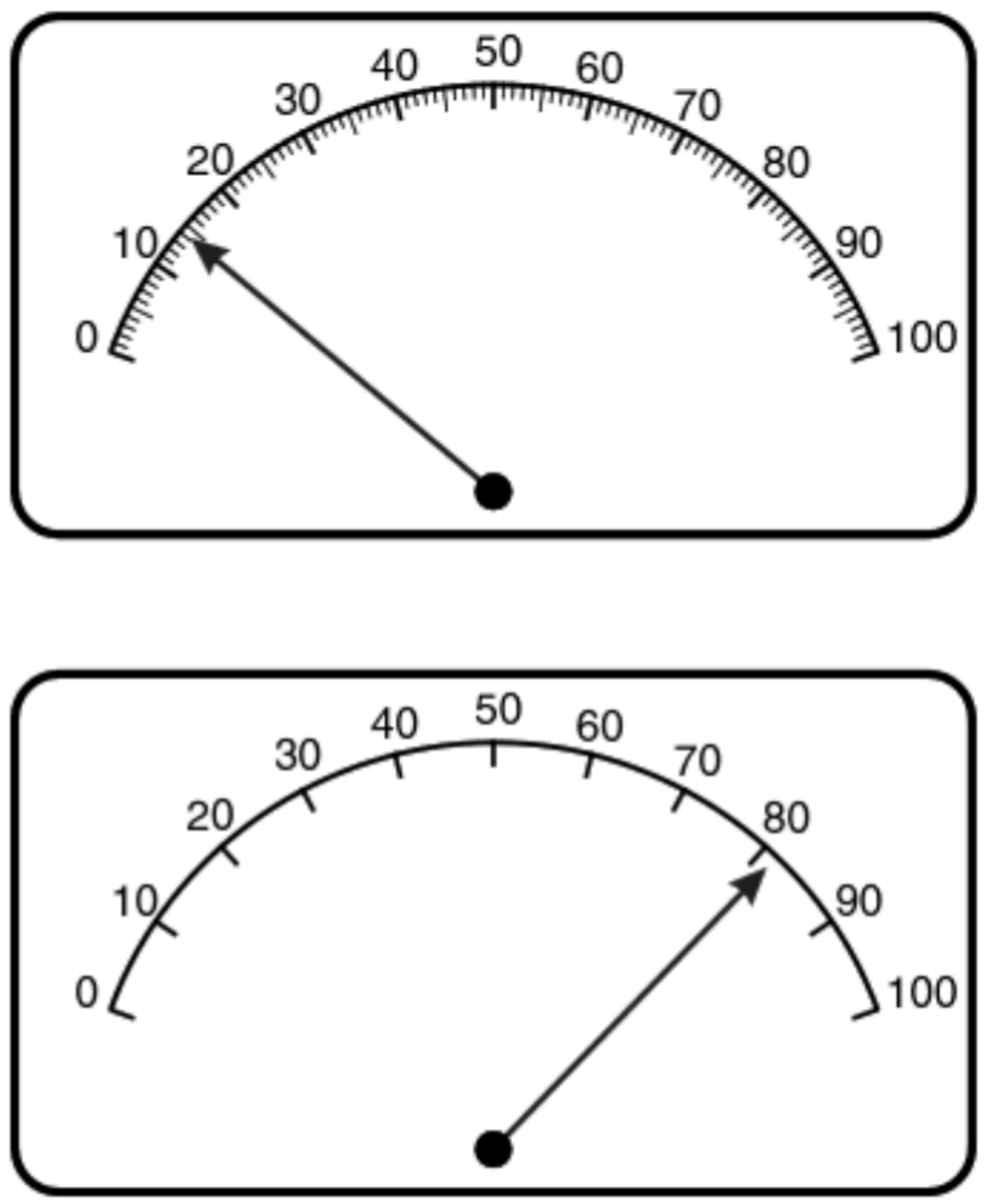


Fig. 1.4 The upper part of the figure displays a situation where estimating the uncertainty to be half a division is appropriate; in contrast in the lower part of the figure the uncertainty in the measurement is substantially smaller than half a division.

(from Hughes & Hase book)

Mean and Standard Deviation

For data, the mean, μ, and (sample) standard deviation, σ, can be calculated from the data as:

$$\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

and

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

(if deriving the mean and the variance from the data)

Probability Distributions

- In Physics experiments/observations the following probability distributions are commonly met:
  - Binomial
  - Poisson
  - Gaussian (also called Normal)

Binomial Distribution

- The Binomial Distribution describes the probability of observing k successes out of n tries where the probability of success is p:

$$P_B(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- The mean and standard deviation are given by (without proof):

$$\mu = np$$

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

- Notes:
  - If $np \gg 1$ the Binomial distribution is approximated by the Normal distribution.
  - As n becomes large and $p \to 0$ $(np = \mu)$ the Binomial distribution is approximated by the Poisson distribution.

Binomial Distribution Examples

- In a certain Physics course 7.3% of students failed and 92.3% passed, averaged over many years. What is (a) the expected number of failures in a particular class of 32 students, drawn from the same population? (b) the probability that five or more of the 32 students fail?

- If I toss a coin 12 times and get 11 heads do I have significant evidence that the coin is unfair? (Note: significant evidence is defined as $<5\%$ compatibility level and highly significant as $<1\%$ level)

- A hospital admits four patients suffering from a disease which has a mortality rate of 80%. Find the probabilities of (a) none of the patients surviving, (b) exactly one survives, (c) two or more survive.

- Of a certain type of seed, 25% normally germinate. To test a new germination stimulant, 100 seeds are treated with the stimulant and planted. If 32 of them germinate, can you conclude (at the 5% confidence level) that the stimulant helps?

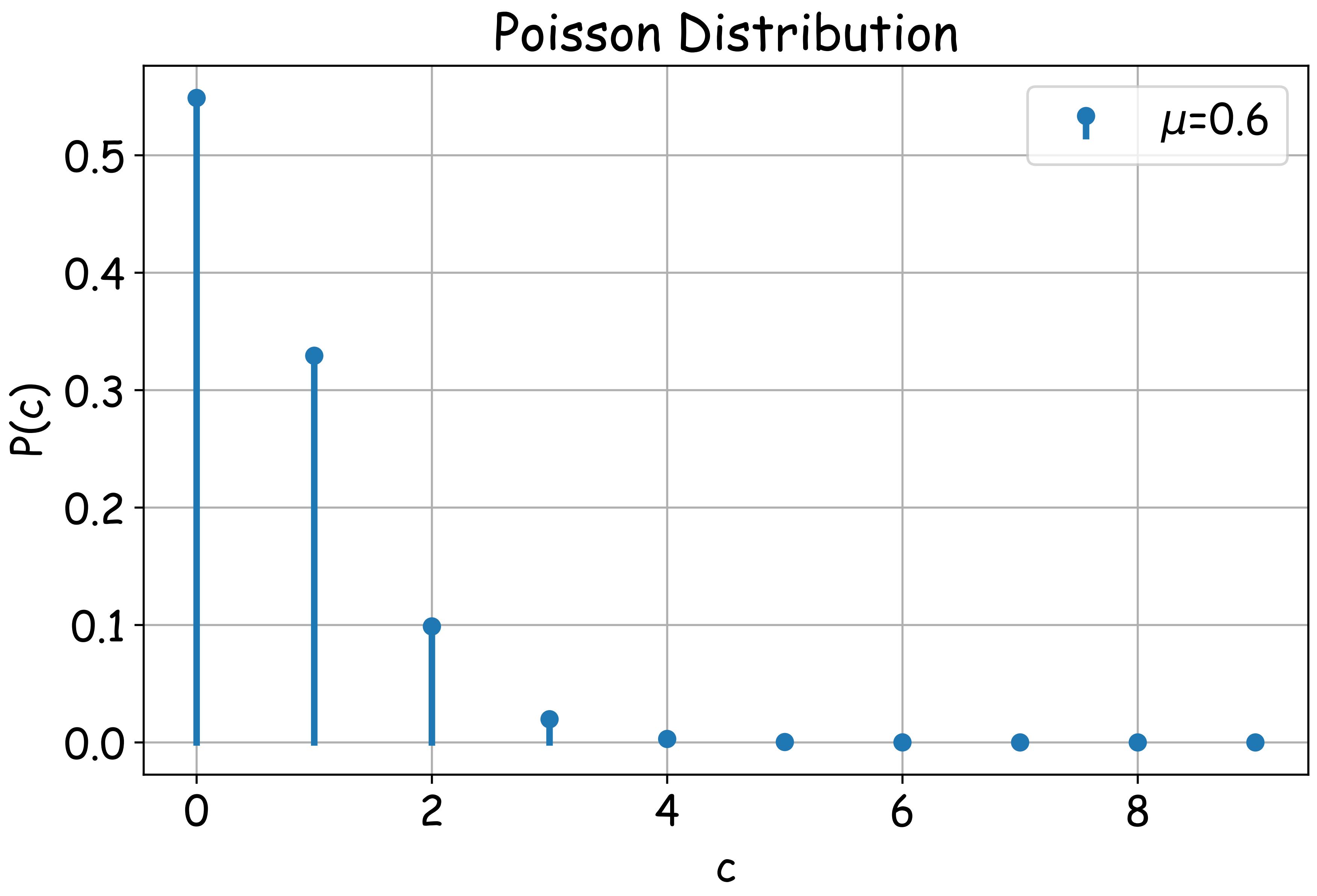
Poisson Distribution

- Poisson (discrete distribution)
  - When one counts the number of random events in an interval (time, area, volume, etc) and repeats the experiment under identical conditions then one does not always get the same result.
  - The distribution of counted events follows a Poisson Distribution:

$$P(n) = \frac{\mu^n e^{-\mu}}{n!}$$

  - where,
    - $P(n)$ is the probability of obtaining n events in a given interval
    - μ is the mean of the distribution.
    - the standard deviation has value $\sqrt{\mu}$
  - The Poisson distribution is the limiting case of the Binomial distribution as n becomes large and $p \to 0$ $(np = \mu)$, and is applicable to many experiments involving counting events such as radioactive decay, photons etc. (In most cases p & n may be unknown/unknowable and only μ measureable)

Poisson Distribution



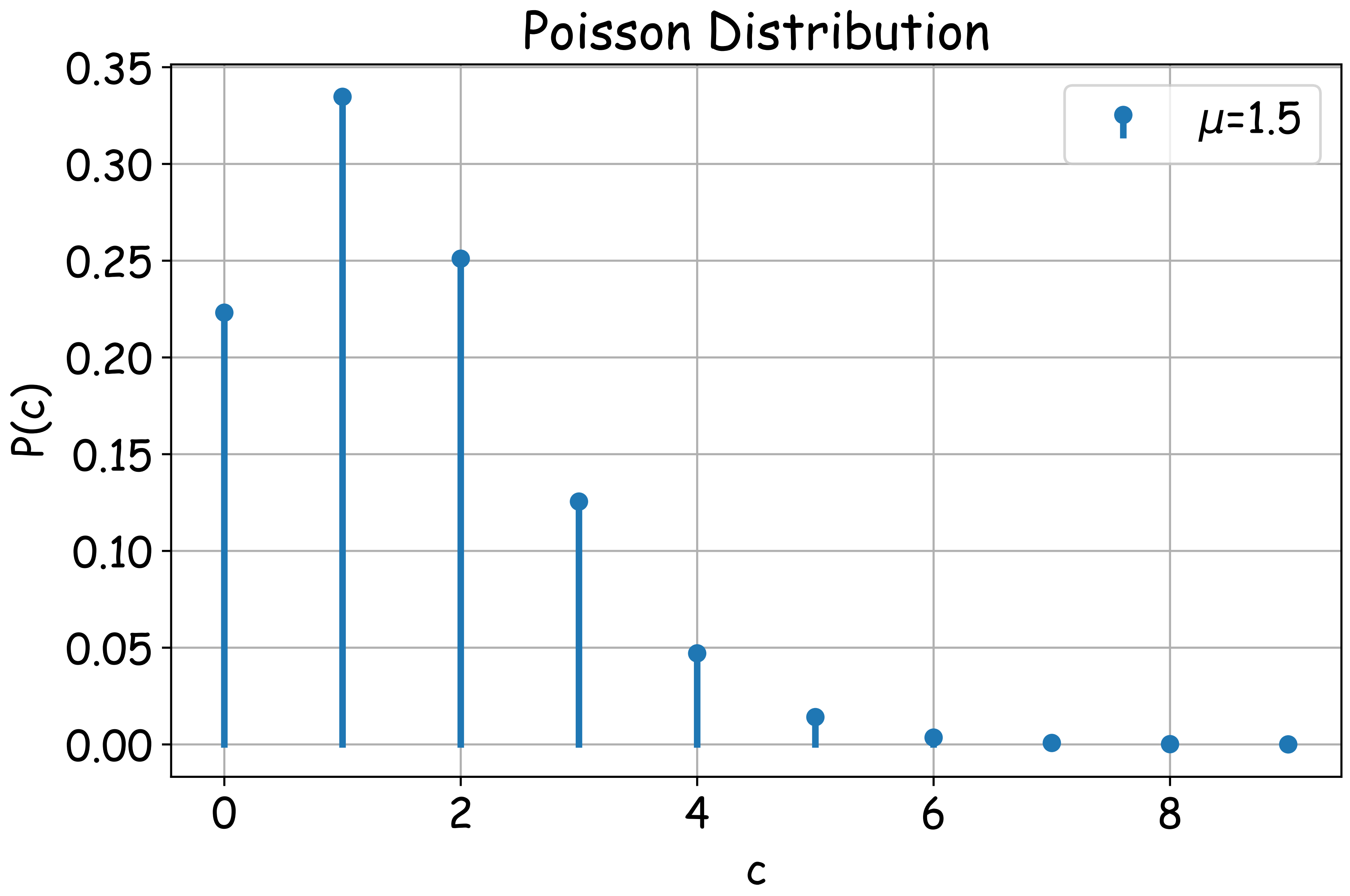
Asymmetric!

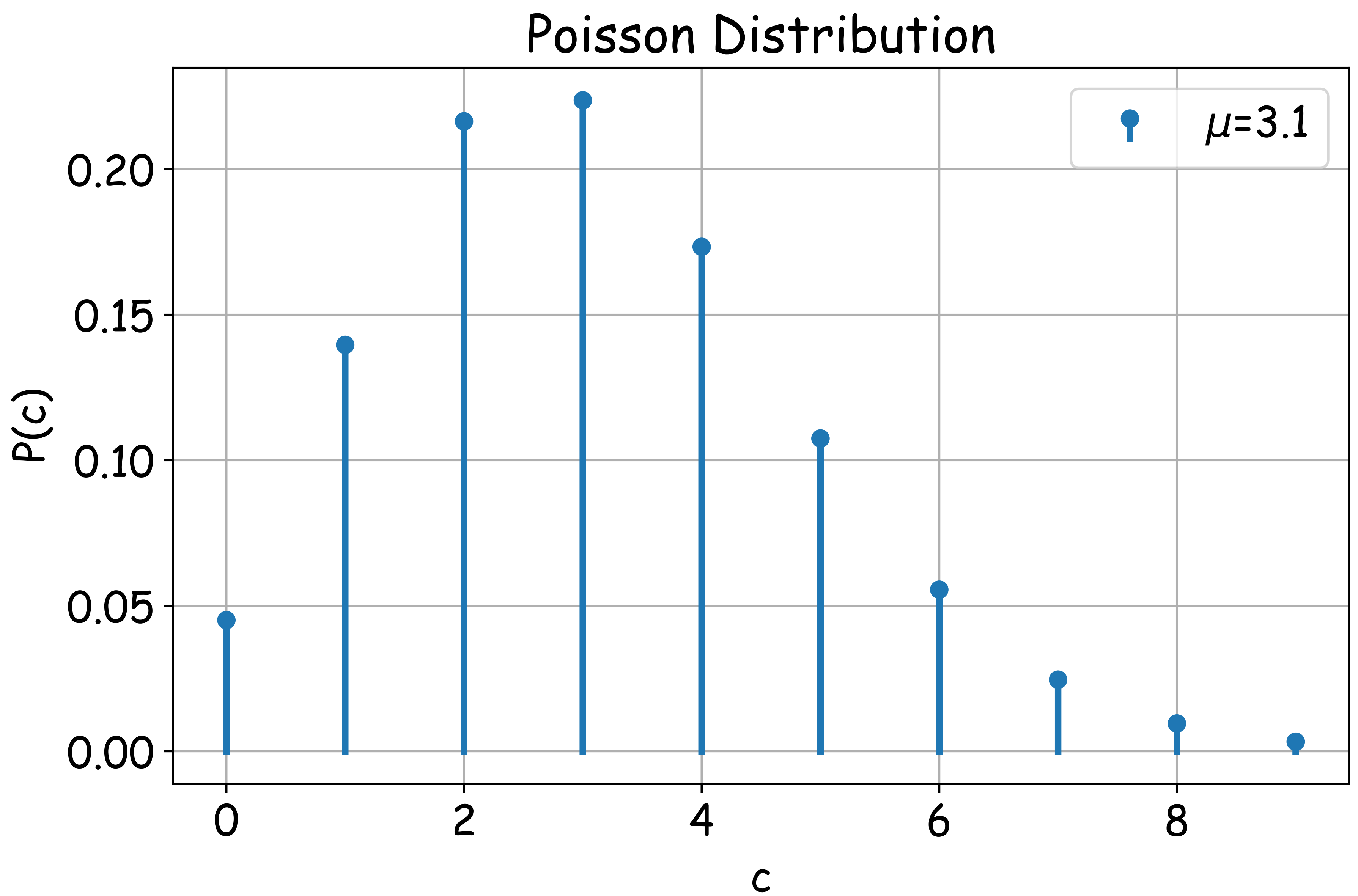
$\mu = 0.6$

$\sigma = \sqrt{0.6} = 0.775$

It does not make sense to quote: 0.600 ± 0.775

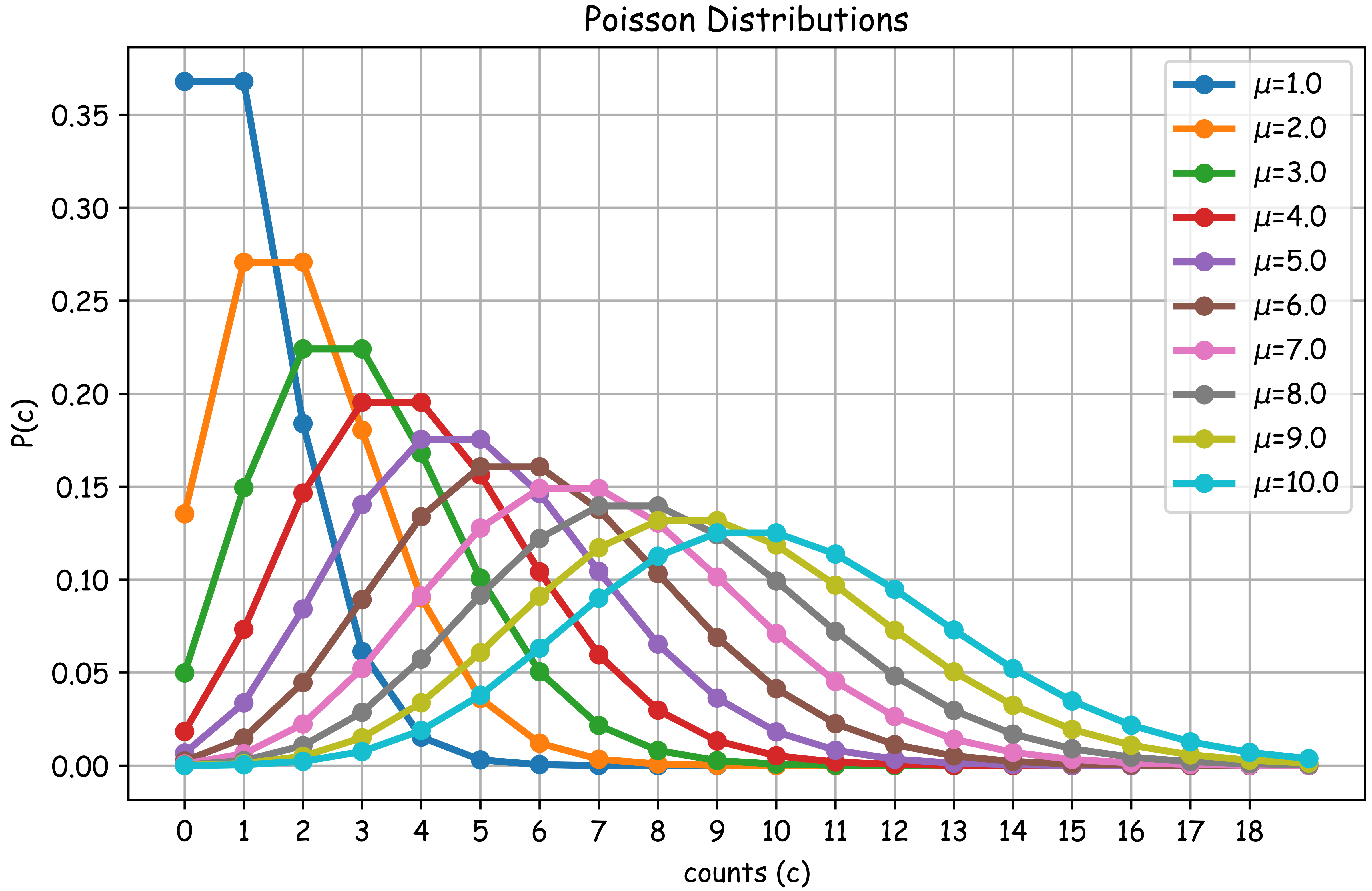
We need asymmetric error bars!





Propagation of Errors

- For small means the distribution is asymmetric.
- For means ≳10 the distribution is nearly symmetric and is approximately described by a Gaussian distribution of mean μ and standard deviation $\sqrt{\mu}$.



- For dealing with errors (especially propagation) in counting experiments we generally want enough counts for the Poisson distribution to be in Gaussian regime.

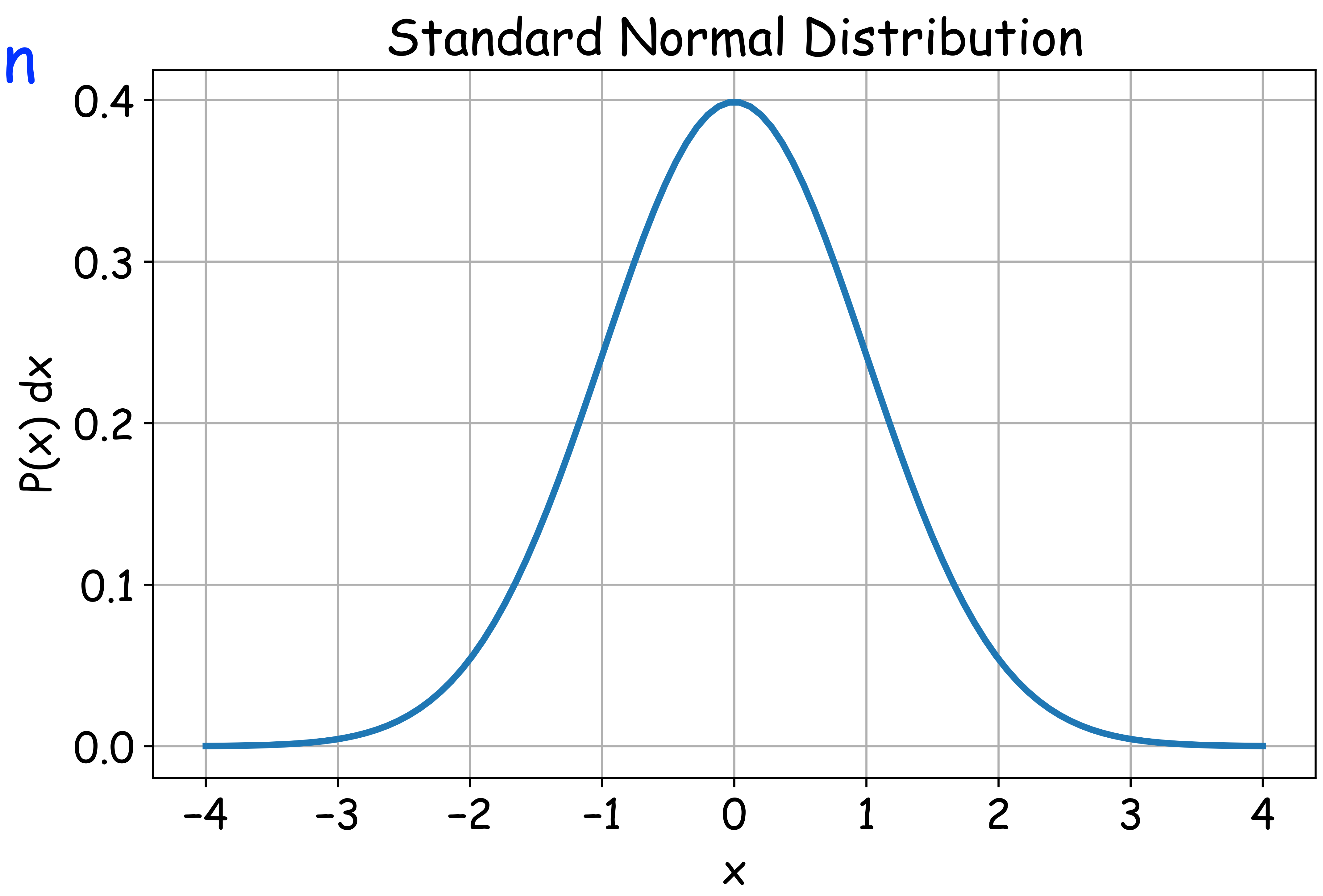
The Gaussian (Normal) Distribution

- Statistical Errors tend to follow a Gaussian Distribution

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

where:

- $P(x)\,dx$ is the probability of obtaining a value between x and $x+dx$
- μ is the mean (centre) of the distribution
- σ is the width of the distribution
- normalised: area = 1.



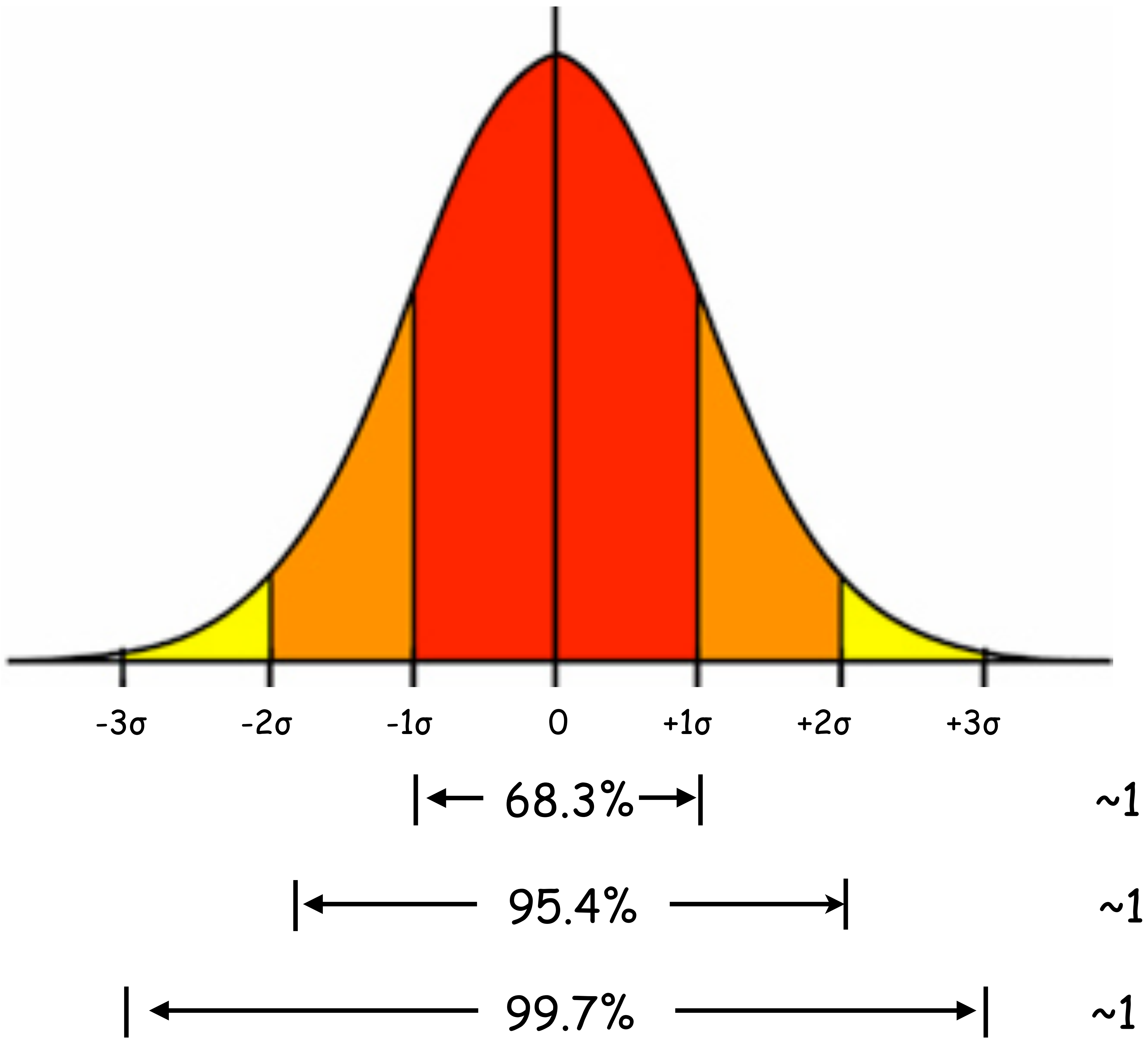
Standard Normal: $\mu=0$, $\sigma=1$

The Normal distribution is the limiting case of the Binomial distribution when $np \geq 5$ and $n(1-p) \geq 5$.

The Normal distribution is the limiting case of the Poisson distribution when $\mu \geq 35$ (note: rough > 10).

The Gaussian (Normal) Distribution

- If we know the mean and standard deviation for a set of measurements (or technique), what is the probability that a measurement will fall in a given range?



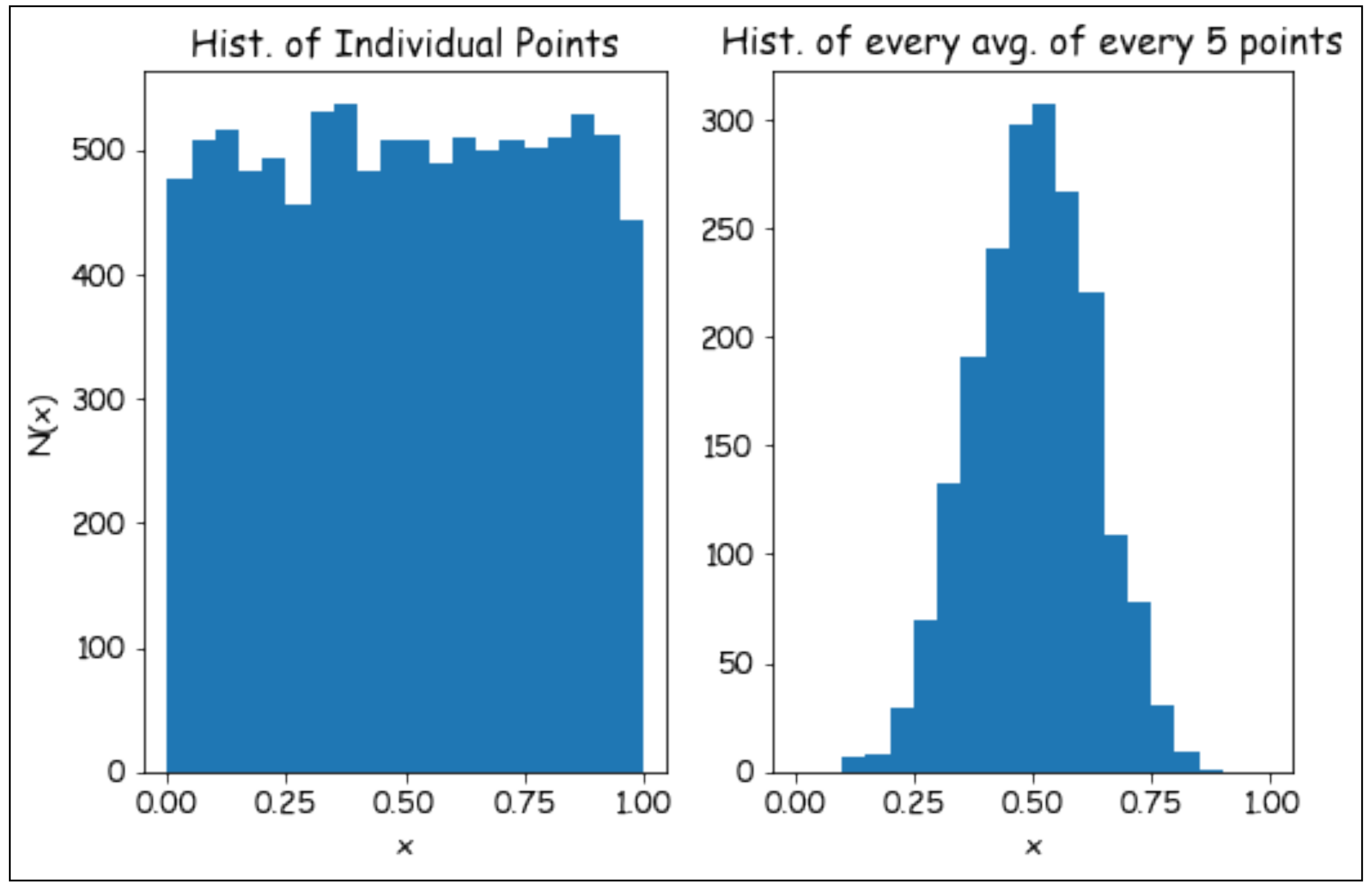
~1 in 3 outside of range

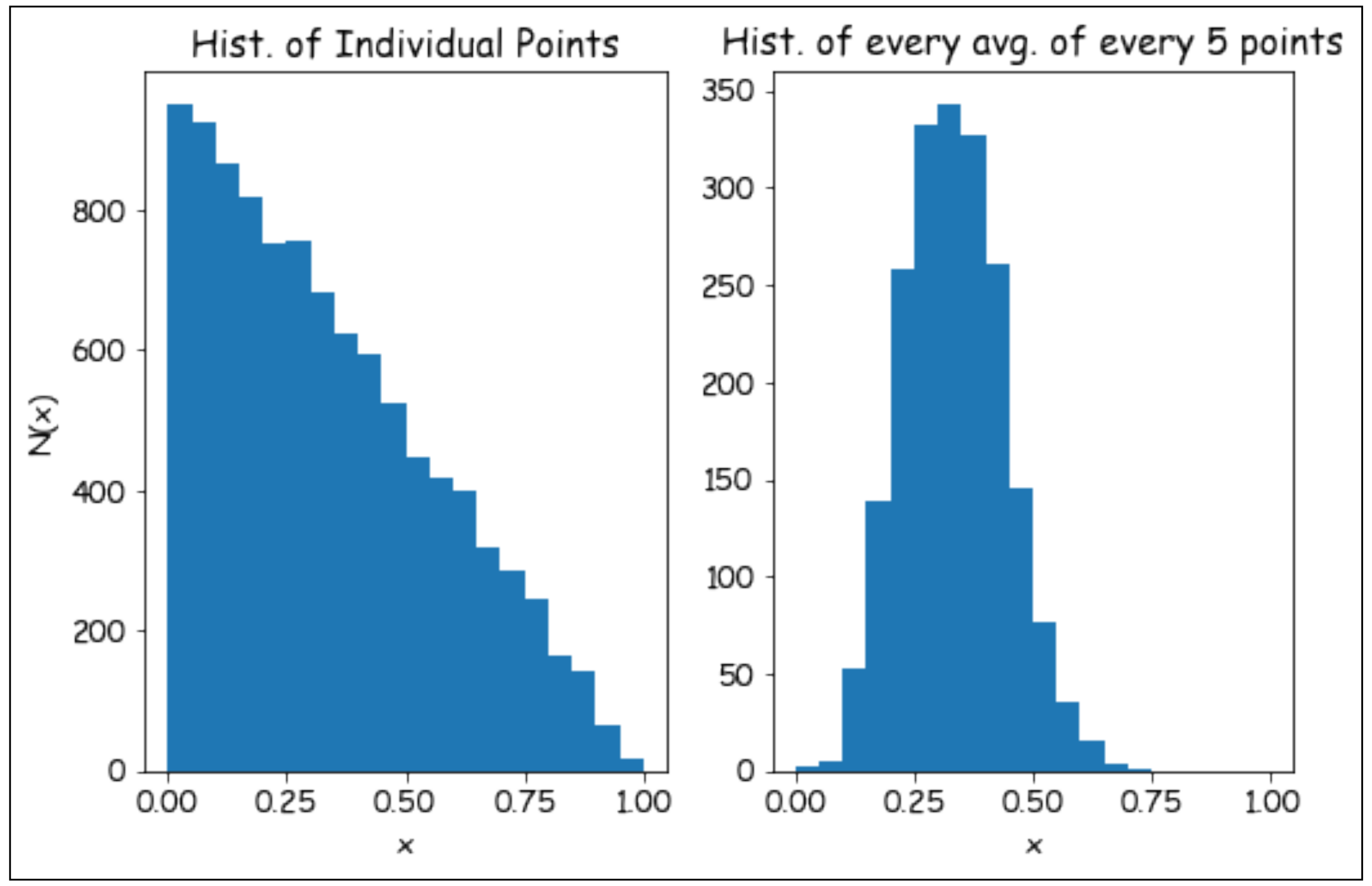
~1 in 20 outside of range

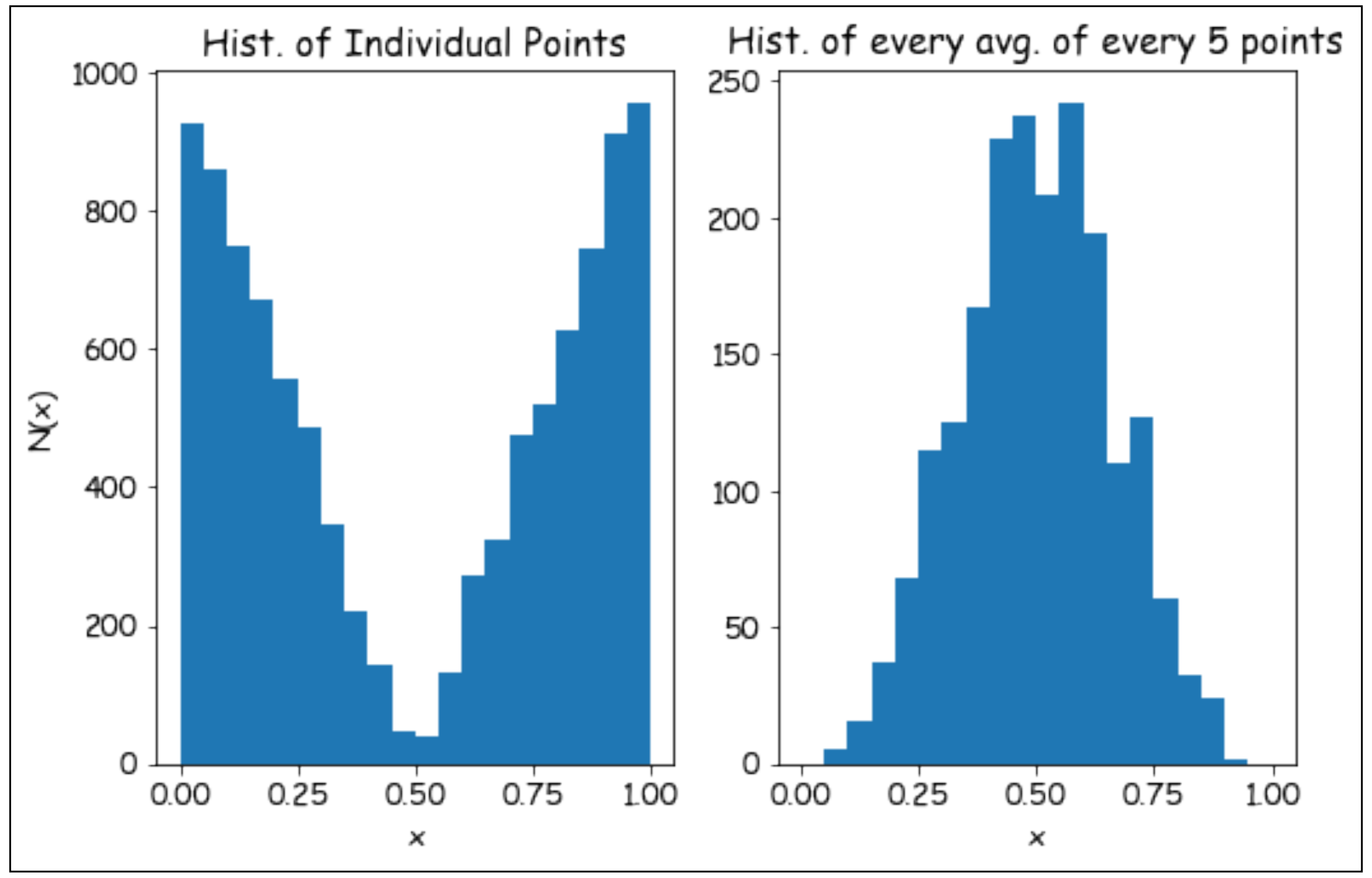
~1 in 400 outside of range

Central Limit Theorem

- Why is the normal distribution so prevalent?
  - Ans: The Central Limit Theorem
    - Irrespective of the parent distribution of some variable, the distribution of the mean of that variable tends towards a normal distribution with the same mean, as the number of samples becomes large (not many needed for many 'reasonable functions'!)







Propagation of Errors

- Assume we have made some measurements (e.g. length and width) and we want to combine the measurements using some formulae to calculate a some property (for example, area).
- How do we estimate the uncertainty on the final result given we know the uncertainties on the initial measurements? (i.e. what is $A \pm dA$?)
- Use Propagation of Errors Formula (without proof) for Gaussian errors:
  - Say x is calculated using values with uncertainties u, v, $\ldots$ (i.e. $x = f(u, v, \ldots)$)
  - Then, for uncorrelated measurement fluctuations:

$$\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v}\right)^2 + \ldots$$

  - If the measurement fluctuations are correlated then we have to include an additional term called the **covariance**.

Propagation of Errors: some common formulae

$x = av \pm bv$	$\sigma_x^2 = a^2\sigma_u^2 + b^2\sigma_v^2 \pm \cancel{2ab}\sigma_{uv}^2$
$x = \pm auv$	$\dfrac{\sigma_x^2}{x^2} = \dfrac{\sigma_u^2}{u^2} + \dfrac{\sigma_v^2}{v^2} + \cancel{\dfrac{2\sigma_{uv}}{uv}}$
$x = \pm \dfrac{au}{v}$	$\dfrac{\sigma_x^2}{x^2} = \dfrac{\sigma_u^2}{u^2} + \dfrac{\sigma_v^2}{v^2} - \cancel{\dfrac{2\sigma_{uv}}{uv}}$
$x = au^{\pm b}$	$\dfrac{\sigma_x}{x} = \pm b\dfrac{\sigma_u}{u}$
$x = ae^{\pm bu}$	$\dfrac{\sigma_x}{x} = \pm b\sigma_u$
$x = a^{\pm bu}$	$\dfrac{\sigma_x}{x} = \pm(b\ln a)\sigma_u$
$x = a\ln(\pm bu)$	$\sigma_x = a\dfrac{\sigma_u}{u}$

Numerical Propagation of Errors

- Say we have a function $x = f(u, v)$ with errors on u (σ_u) and v (σ_v).
- To propagate the error excluding covariance :

$$\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v}\right)^2 + \ldots$$

- For complicated functions it can be tedious to calculate all of the derivatives.
- Alternatively, one can use a computer to numerically propagate errors through any formula:

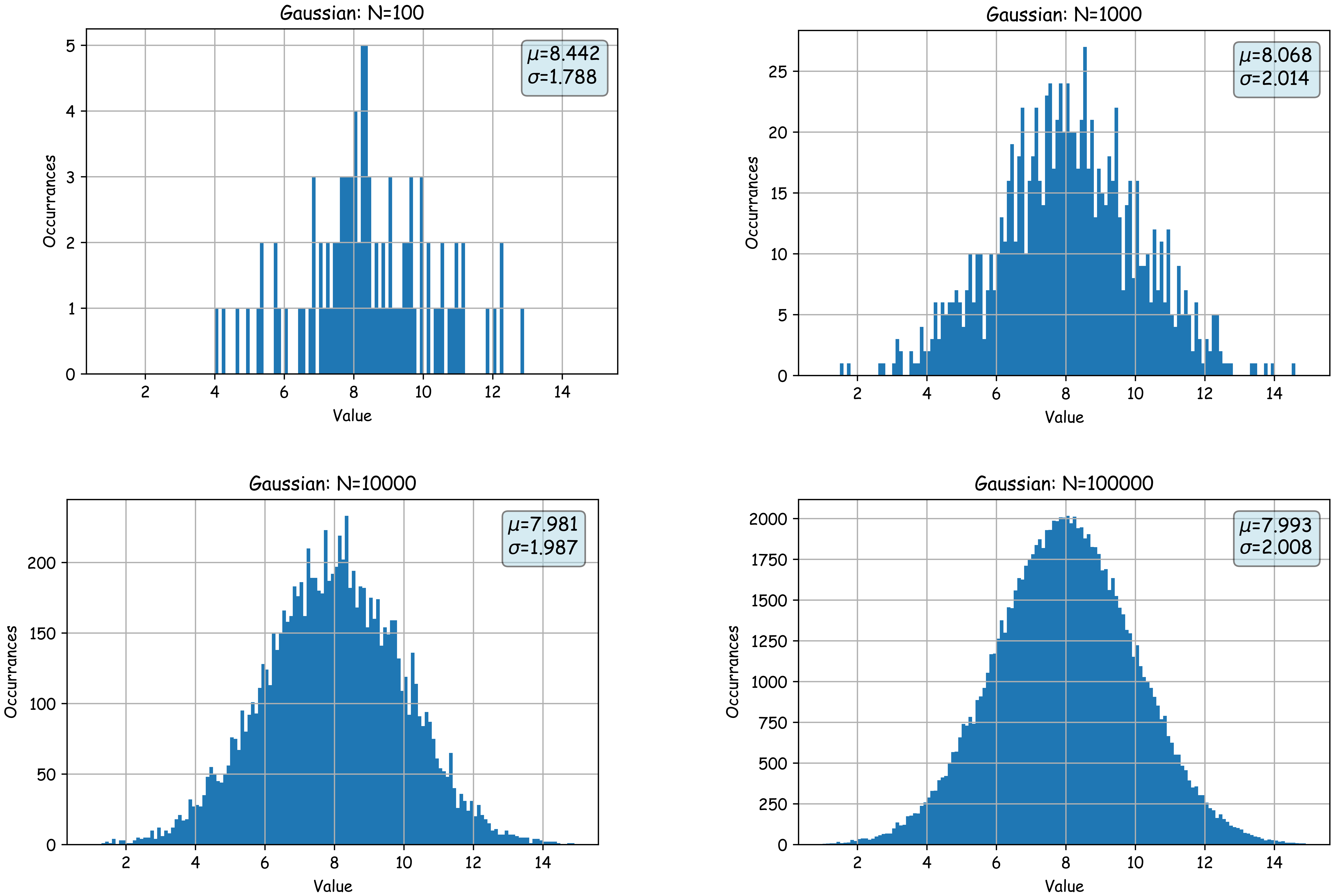
$$dx_u = f(u + \sigma_u, v) - f(u, v)$$

$$dx_v = f(u, v + \sigma_v) - f(u, v)$$

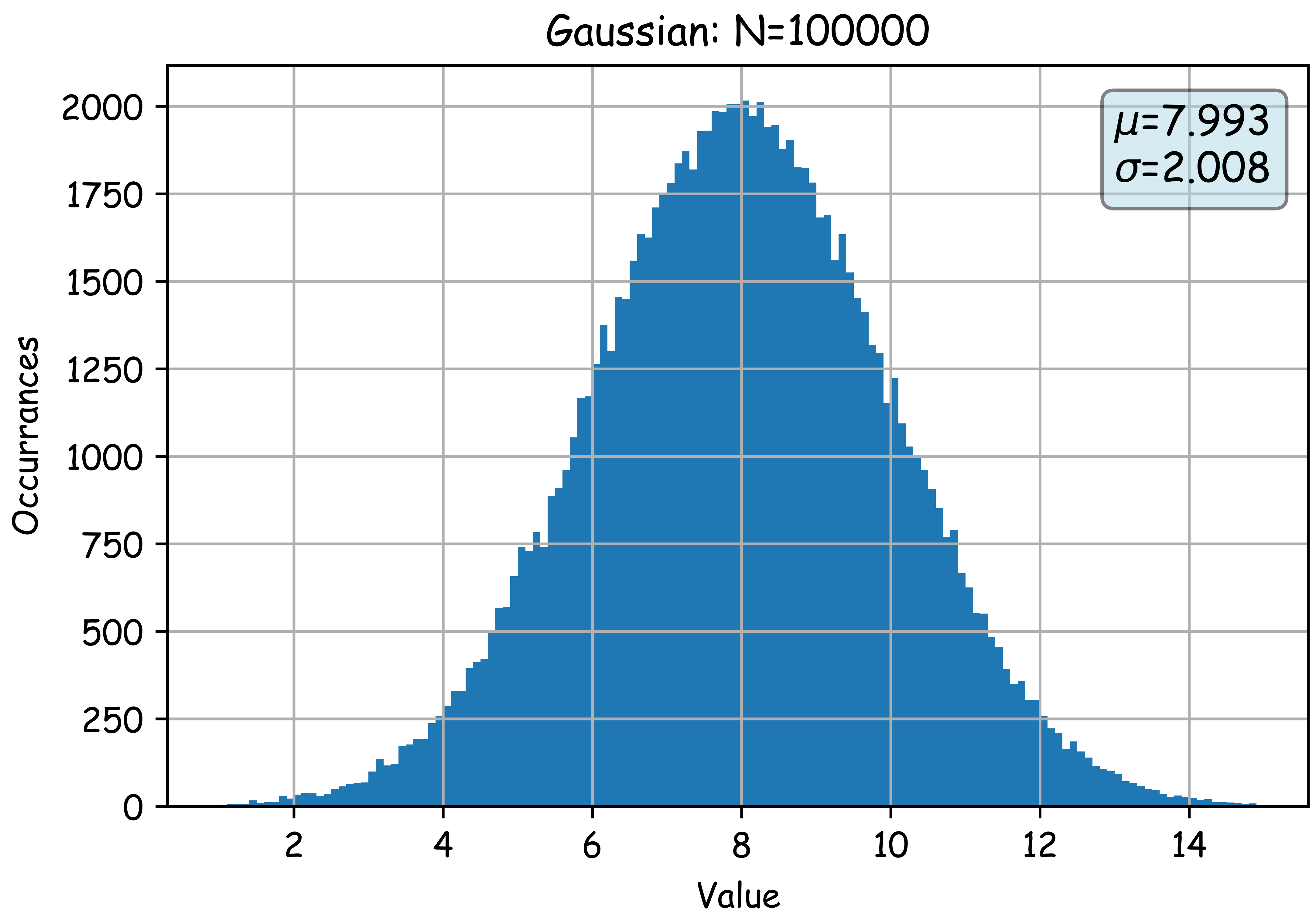
$$\sigma_u = \sqrt{dx_u^2 + dx_v^2}$$ w/o covariance

The Mean and Its Uncertainty

- The more measurements of a quantity we make the more precisely we can characterise the distribution



The Mean and Its Uncertainty



- What does this distribution tell us?
  - The probability of getting a given value in a single measurement.
- What is the uncertainty on the mean?
  - We know the mean much more accurately than to within $\pm 1\sigma$ of the parent distribution!
- Using Propagation of Errors it is possible to show that the uncertainty on the mean is:

$$\sigma_\mu = \frac{\sigma}{\sqrt{N}}$$

If N data points are averaged to get the mean, the error on the mean is not the standard error/deviation σ of the distribution but $\sigma/\sqrt{N}$ ("the standard error on the mean")

The Weighted Mean and its Uncertainty

- If we have a set of measurements taken with different uncertainties (e.g. we improve the technique or apparatus part of the way through), then we can combine the data using the following formulae:

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

"Weighted Mean"

$$\sigma_\mu^2 = \frac{1}{\sum(1/\sigma_i^2)}$$

"Error on the Weighted Mean"

Note: the values being combined must be compatible with each other!

Quoting Errors

- Please read Sections 2.8 & 2.9 of Hughes & Hase book!

- General guidance:
  - The best estimate of a parameter is the mean.
  - The error is the standard error on the mean.
  - Perform calculations using all significant figures.
  - Errors are generally only quoted to 1 significant place (unless: lots and lots of data has been used to determine the error or the first significant figure is a "1" - then use two)
  - Match the number of places in the mean (or parameter being quoted after error propagation) to the error.
    - e.g. $(96.8645 \pm 0.2701) \times 10^3\ \Omega$ should be quoted as $(96.9 \pm 0.3) \times 10^3\ \Omega$
    - e.g. $(96.8645 \pm 0.1435) \times 10^3\ \Omega$ should be quoted as $(96.86 \pm 0.14) \times 10^3\ \Omega$

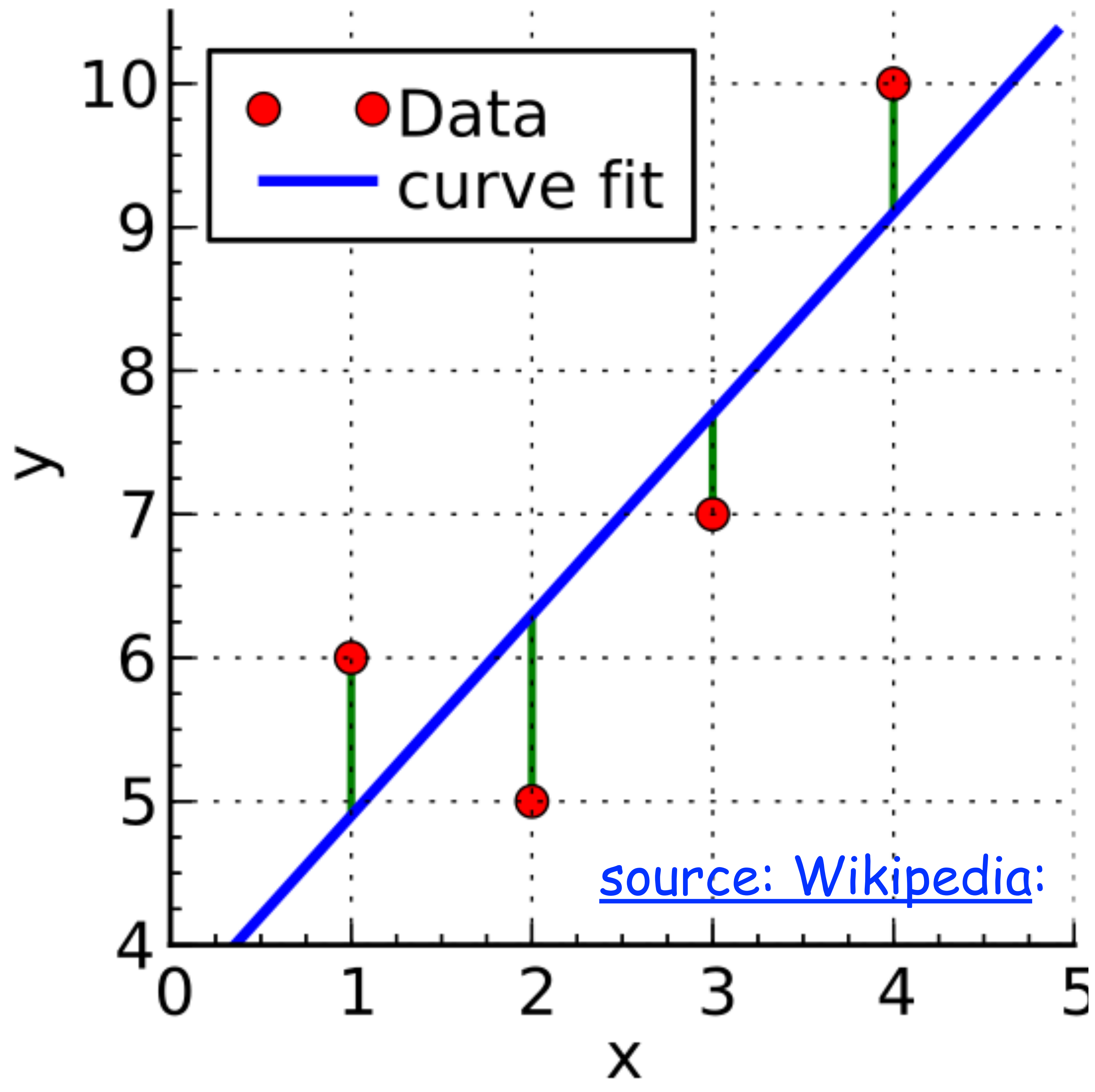
Curve Fitting: least-squares and χ^2 fitting

- The Method of Least Squares involves adjusting the parameters of a function so that the sum of deviation of each data point squared is minimised,
  - e.g. fitting a straight line to data we must find m and c which minimise:

$$S = \sum_i \left[y_i - f(x_i)\right]^2 = \sum_i \left[y_i - (mx_i + c)\right]^2$$

- We can also include the errors on the data points to 'weight' the points by their errors (Chi-squared (χ^2) minimisation.):

$$S = \sum_i \left[\frac{y_i - f(x_i)}{\sigma_i}\right]^2 = \sum_i \left[\frac{y_i - (mx_i + c)}{\sigma_i}\right]^2$$



- Software packages such as `scipy.minimise.curve_fit()` perform least squared and χ^2 minimisations.

Using curve_fit()

- See: https://github.com/UCD-Physics/Python-HowTos/blob/main/Curve_fit.ipynb
-

χ^2 and Goodness of Fit

- The χ^2 test is a test for Goodness of Fit.
- It can be used to compare data with theory (e.g. a theoretical curve is fit to experimental data and we make a statement about the probability that the theory and data agrees).
- It can also be used to compare two different data sets to see if they agree.
- Definition:

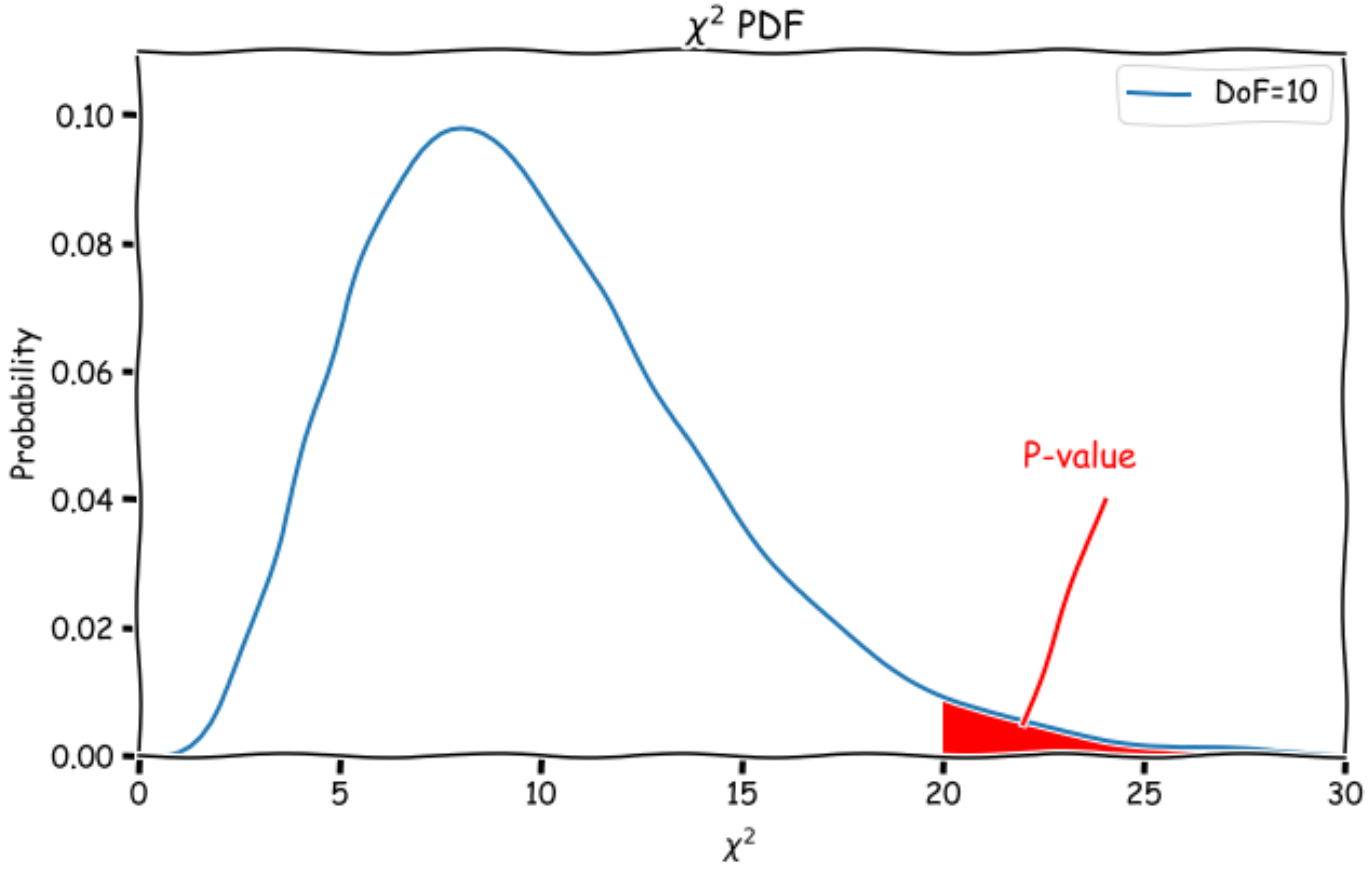
$$\chi^2 \equiv \sum_{i=1}^{N} \left(\frac{\text{measured}_i - \text{expected}_i}{\text{error on measured}_i} \right)^2$$

[* note different errors for counting experiments]

- If we have reasonably good agreement between data and theory then we would expect that we would get a contribution of ≈ 1 from each data point to sum.
- This is approximately true, in fact we get $\chi^2 \approx \upsilon = n - n_c$ where υ is called the "Number of Degrees of Freedom" and is equal to the number of data points, n, minus the number of constraints (free parameters in fit function), n_c, derived from the data.
- χ^2/υ is called the "Reduced Chi Squared" and for a good fit is ≈ 1.

χ^2 and Goodness of Fi

- χ^2 is a distribution with a unique curve for each number of degrees of freedom.
- After finding the optimum parameters by minimising χ^2 we can check to see if the χ^2 value is reasonable for the degrees of freedom in question:
  - the reduced χ^2 gives a good indication (but interpretation depends strongly on the DOF)
  - better: the P-value:
    - the probability of obtaining a value of $\chi^2 \geq \chi^2_{observed}$ by chance (for normally distributed errors if the model and data agree) is given by integrating the χ^2 distribution from the observed value to infinity:



- Look up in table, or
- use scipy.stats.chi2.sf()

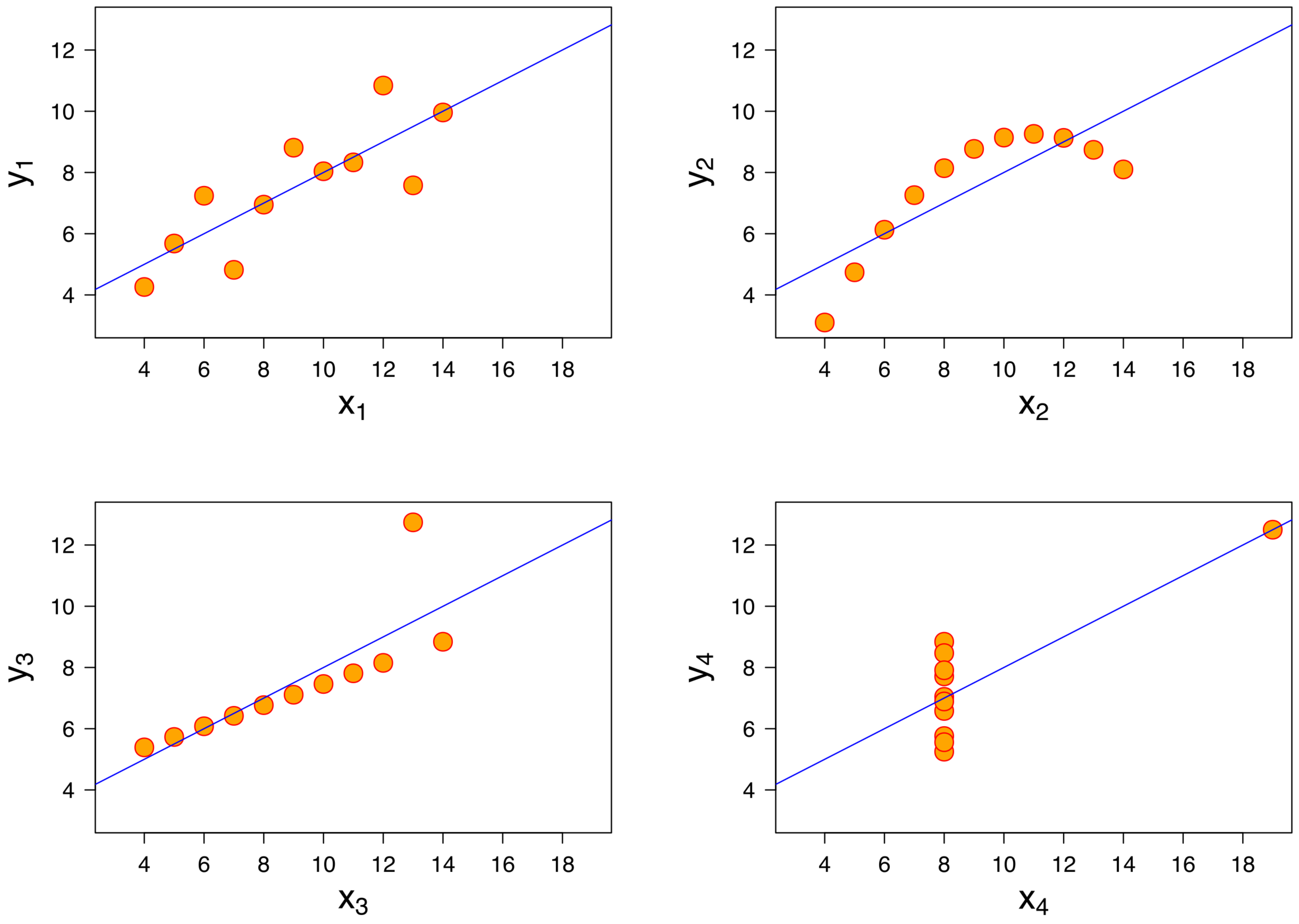
Agreement?

- General guidance:
  - Comparing Experimental Results with an accepted answer (Hughes & Hase. 3.3.4, p.28):
    - up to one standard error, they are in excellent agreement ;
    - between one and two standard errors, they are in reasonable agreement ;
    - more than three standard errors, they are in disagreement .
  - χ^2 test for goodness of fit (Hughes & Hase. 8.4, p.106):
    - very good agreement: $P(\chi^2_{min}; \nu) \sim 0.5$
      - If $P(\chi^2_{min}; \nu) \rightarrow 1$, uncertainties are too large or data is too perfect
    - The agreement is questionable if $P(\chi^2_{min}; \nu) \approx 10^{-3}$.
    - The agreement is bad if $P(\chi^2_{min}; \nu) \lesssim 10^{-4}$

Always visually inspect your data:

- Visual inspection can reveal trends to the eye that are not picked up by statistics! (especially outliers!)

- Example, <u>Anscombe's quartet</u>:
  - all four data sets have the same mean and variance for x & y, and the same linear fit parameters and coefficient of correlation



A Complete Example

- https://github.com/UCD-Physics/Python-HowTos/blob/main/Curve_fit.ipynb