# PHYS465: Statistical Data Analysis in Physics

## *Week 2: Hypothesis Testing, Common distributions, Model testing*

Dr. Mathew Smith

mat.smith@lancaster.ac.uk

Physics Building; C46

# *Week 12: Learning aims*

## *Today we will introduce*

- Probabilities and how they relate to measurements
- The Null Hypothesis
  - And how to test it
- How our experiment will change how the data is distributed
- How to test which distribution is correct
- How to select which model is most likely

Practical examples on Thursday!

# Revision: Probabilities

Given an experiment:

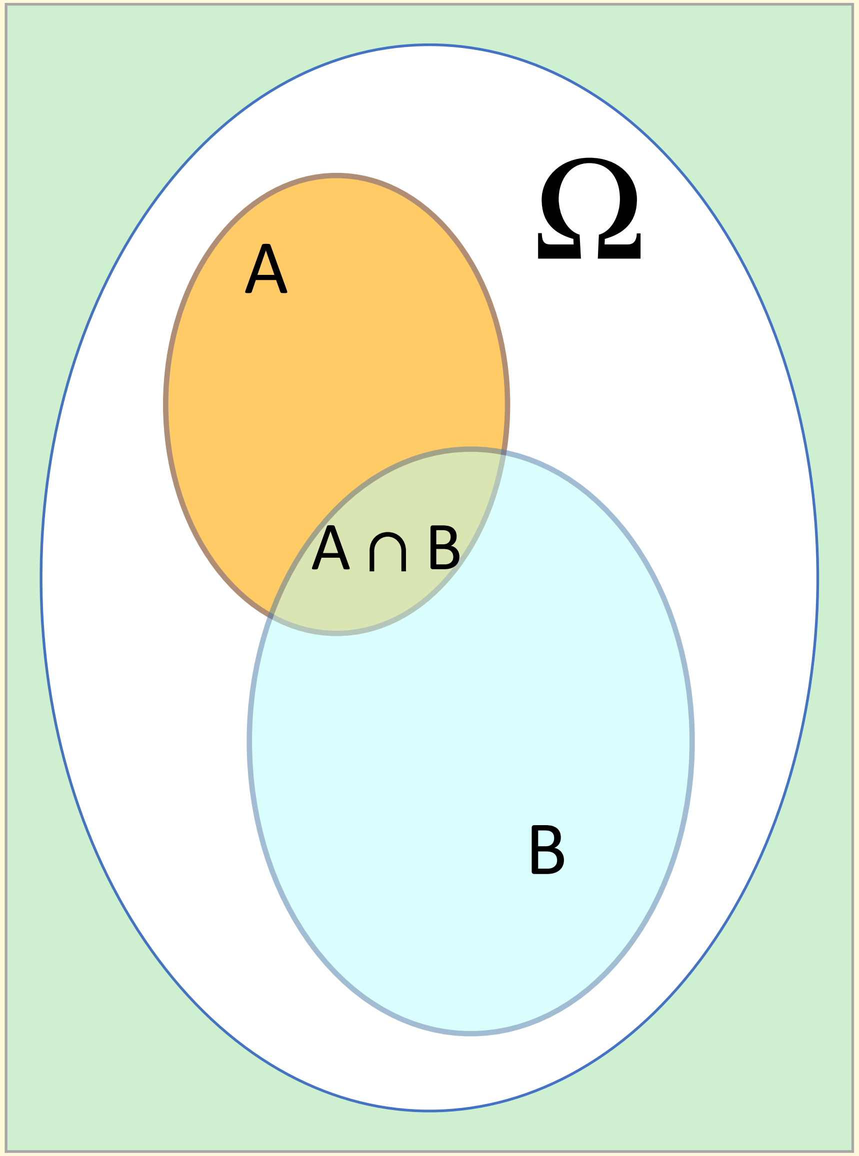
- $\Omega$ is the list of all possible outcomes

Axioms:

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$
- $P(A_1 \cup A_2 \cup A_3 \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$

Consequences:

- $P(A) + P(A^c) = 1$
- $P(A \cup B) = P(A) + P(B) + P(A \cap B)$
- $P(A \cap B) = P(A) \times P(B)$ if $A, B$ are independent



# Conditional Probabilities

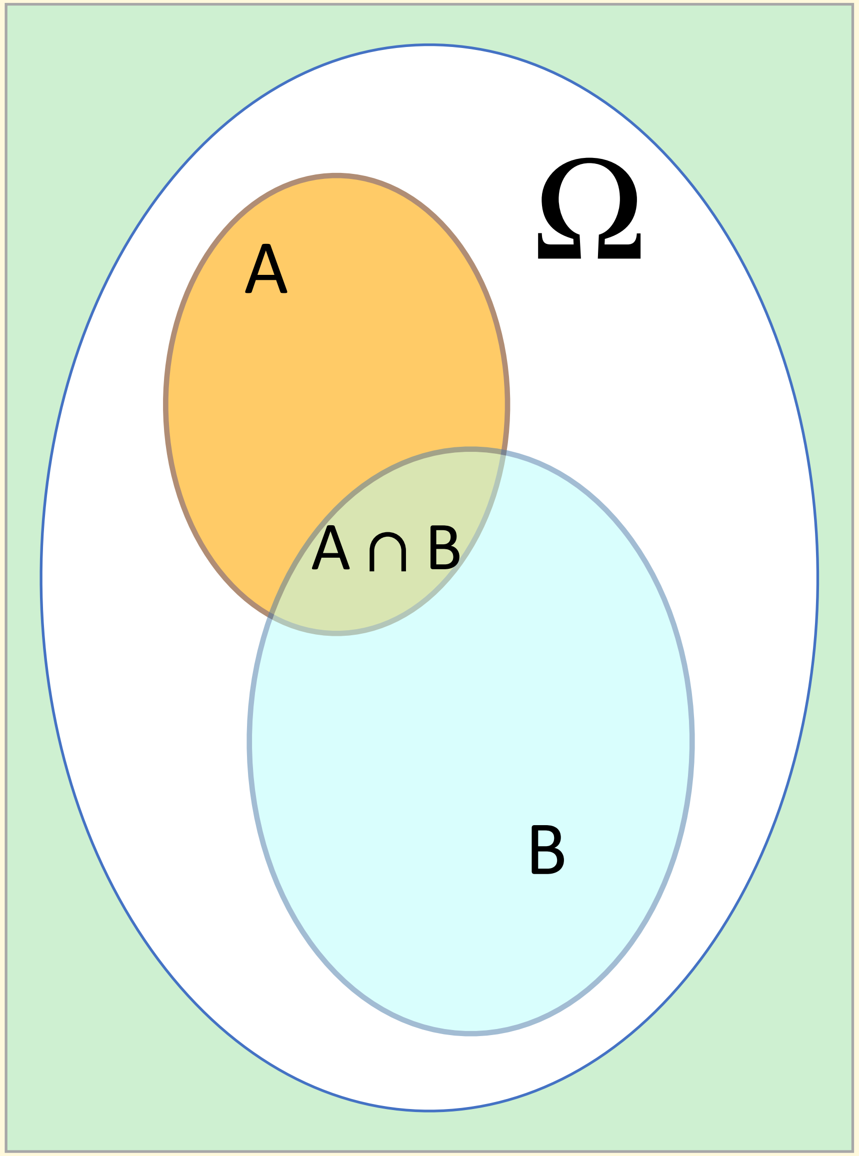
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

*The probability of $A$ given $B$ (an event or condition)*

when $\theta$ is a model, or parameter

$P(A \mid \theta)$ is called the ***likelihood:*** $\mathscr{L}(\theta \mid A)$

"the ***likelihood*** of obtaining the data 'A' given the model"



# Hypothesis testing

How good is the model?

**Fundamental question:** Is the model that we are testing a good fit to our measured data?

Is there a more likely alternative?

**Null Hypothesis:** Our starting assumption

Can be either parameter values or choice of model

**In practice:** Compares our assumption (or prediction) with experimental measurements
e.g. does the luminosity distribution of galaxies match a Schechter function?

**Outcome:** Make a statement on the probability of obtaining our result
(see confidence level slides).

# *Null* Hypothesis testing

e.g. Given a model $f(\theta)$, with parameter $\theta$, is the value of $\theta$ in a set of possible values $\Theta_0$?

$H_0$ : **The Null Hypothesis:** $\theta \in \Theta_0$

$H_1$ : An Alternative Hypothesis:

- $\theta \in \Theta_1$ where $\Theta_0 \cap \Theta_1 = 0$

$\alpha$ : threshold of rejection (e.g. 0.05)

With new observations $X$, such that $p(X, \theta)$

| | |
|---|---|
| $p(X) < \alpha$ | reject $H_0$ and accept $H_1$ |
| $p(X) > \alpha$ | no evidence to reject $H_0$ |

(N.B. This is not the same as accepting $H_0$)

| | | Outcome | |
|---|---|---|---|
| | | $H_0$ not rejected | $H_0$ rejected |
| Truth | $H_0$ is true | Probability of this: $1-\alpha$ | Type I error<br>Will happen $\alpha$ % of the time |
| | $H_1$ is true | Type II error<br>Happens $1-\beta$ % of the time | $P_{H_1}(H_0 \text{ is rejected}) = \beta$ |

# *Confidence Intervals*

*Given a probability threshold, what is the allowed range of the parameter:*

$$\hat{p} - \delta p \leq P_0 \leq \hat{p} + \delta p$$

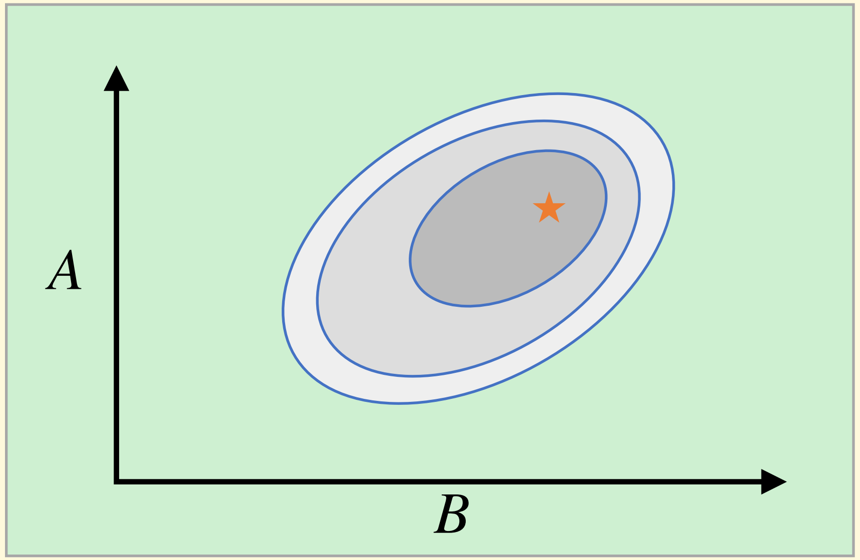
***Interpretation:*** if we repeat our experiment, how often will $P_0$ be within our quoted range

**For multiple variables $(A, B)$:**

An error ellipse containing $\delta p$

of the probability

Typical values are 68%, 90%, 95%.



# Key Assumption:

The probability, $p$, needed for **Hypothesis testing** and **confidence intervals** depend on how the data and model are distributed

Are they both drawn from a Gaussian distribution?

# *Key Distributions: Gaussian*

## Aka a normal

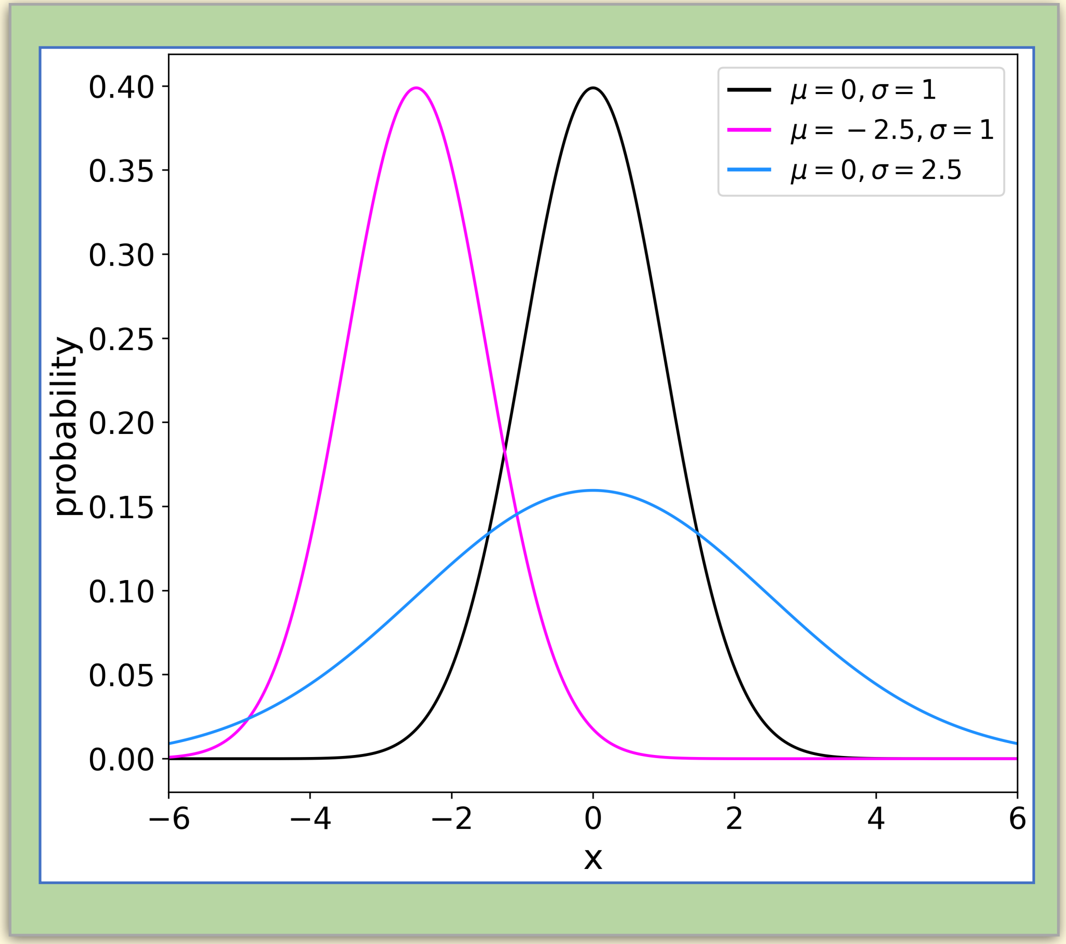
$$P(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\{-(x-\mu)^2/2\sigma^2\}$$

The limiting case ( $\to$ $\infty$) of most other distributions

***Expectation value:*** $\text{mean} = \mu$

***Variance:*** $\sigma^2 = \sigma^2$

$$P(-t\sigma < x_0 < t\sigma) = \frac{1}{\sqrt{2}} \int_{-t}^{t} e^{-z^2/2} dz$$




Common assumption across physics

# Key Distributions: Binomial

## Fixed number of outcomes

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

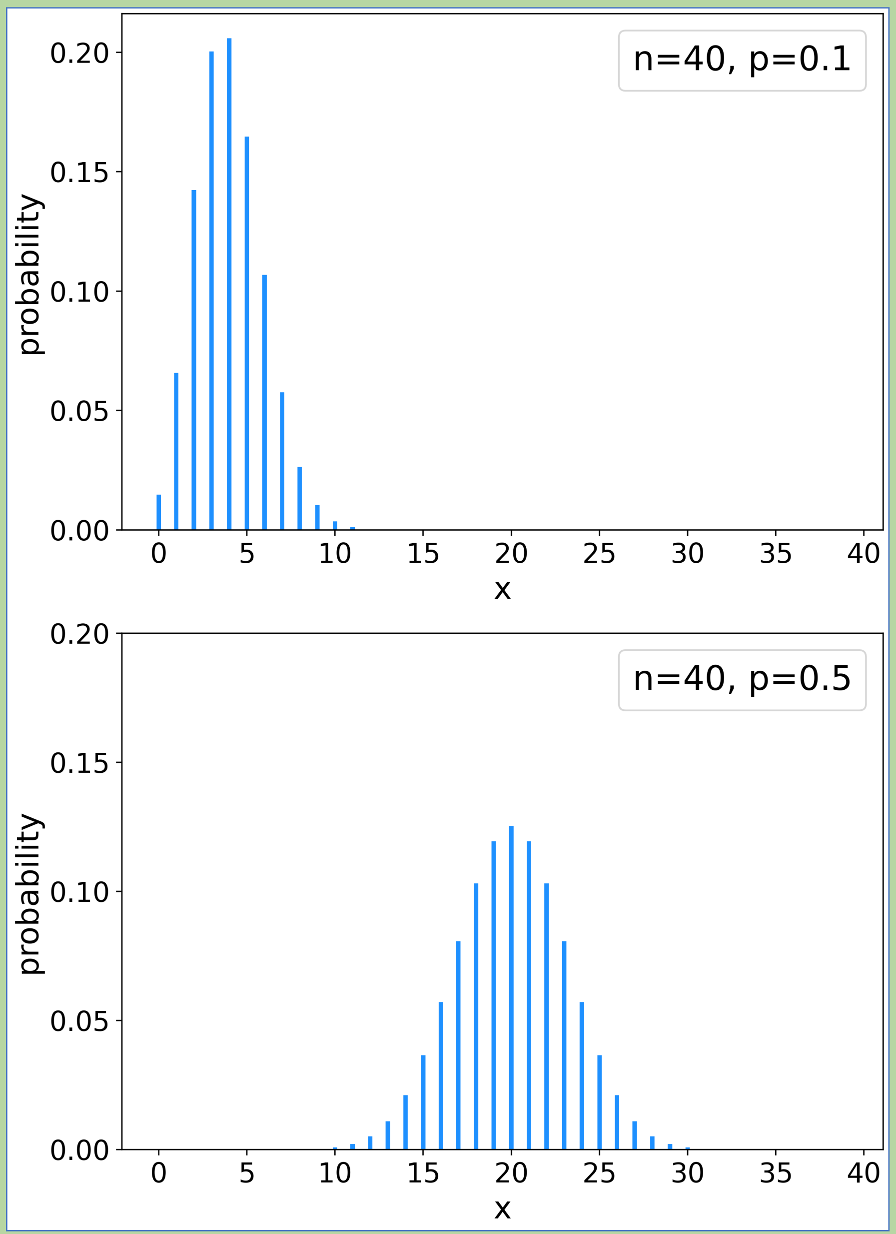
draw $n$ events from $k$ possibilities each with probability $p$

***Expectation value:*** $\text{mean} = np$

***Variance:*** $np(1-p)$

***=> Gaussian distribution as $N \to \infty$***

***Example: coin tossing***



*NB: Other common distributions include 'uniform', 'gamma', 'log-normal'…*

# Key Distributions: Poisson

## The counting distribution

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

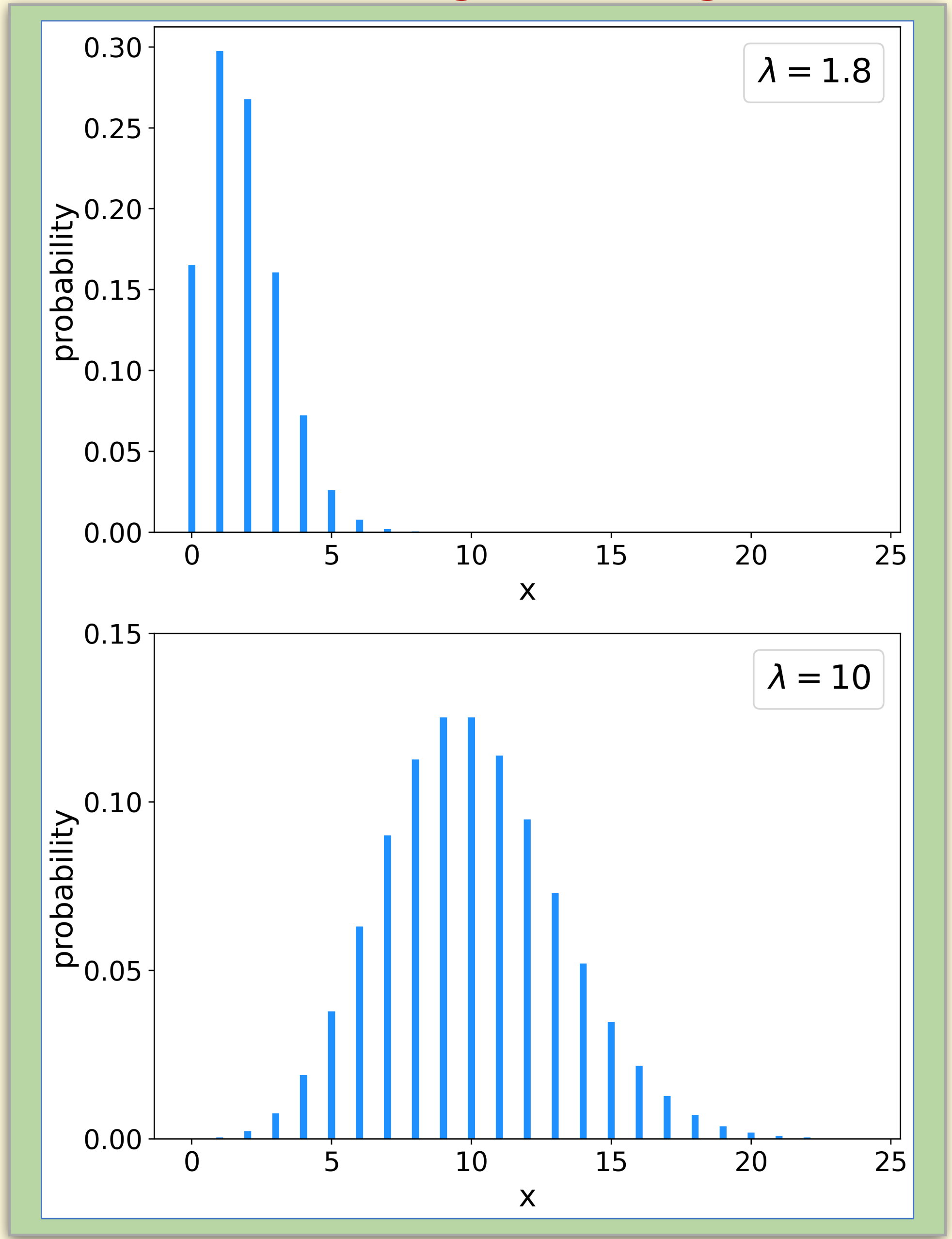
observe $k$ events drawn from $\lambda$ possibilities

*Expectation value:* $\text{mean} = \lambda$

*Variance:* $\sigma^2 = \lambda$

*=> Gaussian distribution as $k \to \infty$*

*Example: histogram (or photon) counting*



# Key Distributions: $\chi^2$

**Related to the Gamma distribution**

$$P \propto (\chi^2)^{\frac{\nu-2}{2}} \exp(-\chi^2/2)$$

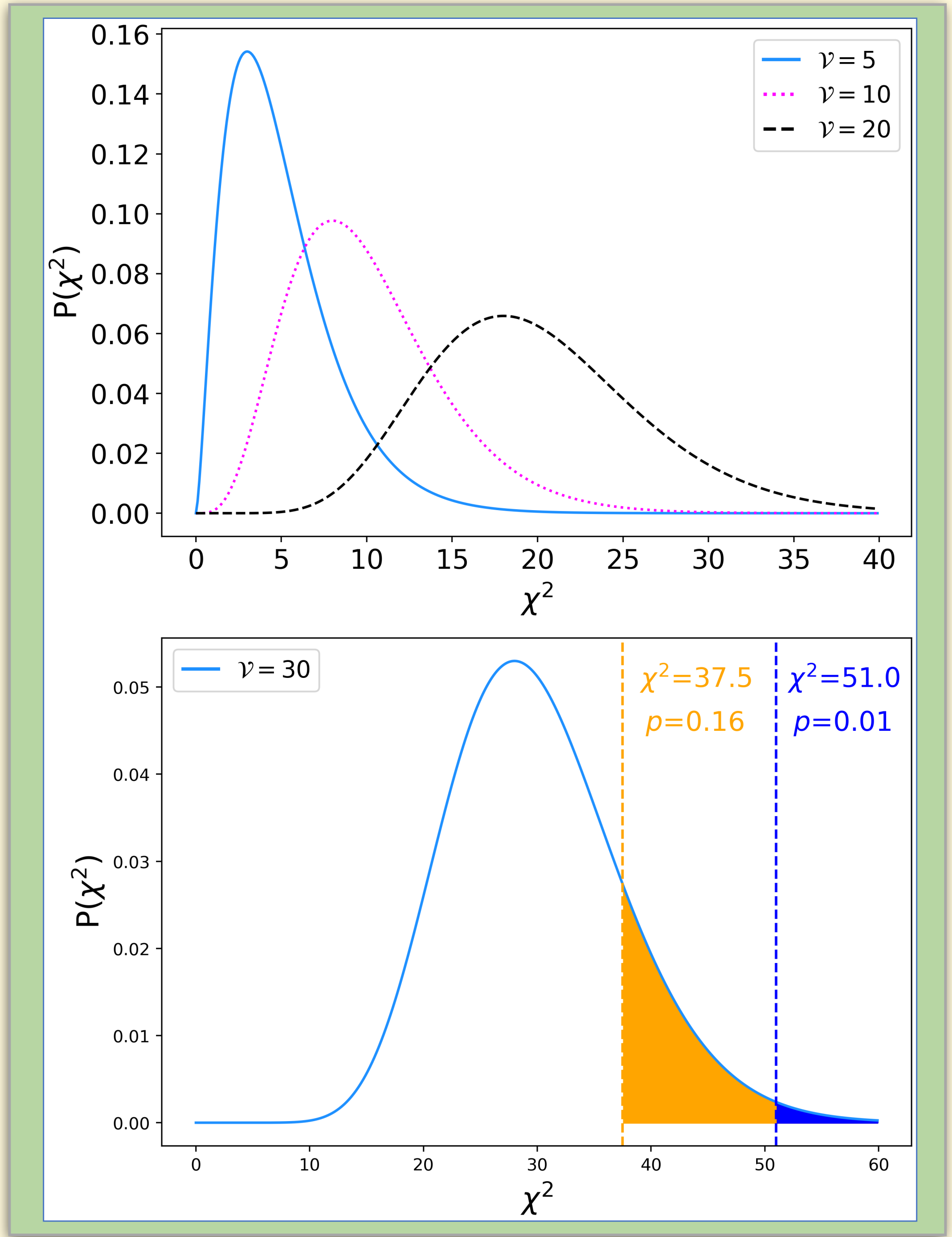
$\nu$ is the number of degrees of freedom

***Expectation value:*** $\mathrm{mean} = \nu = N - p$

***Variance:*** $\sigma^2 = 2\nu$

***i.e. if the model is correct then we expect*** $\chi^2 \sim \nu \pm \sqrt{(2\nu)}$

$\chi^2/\nu \sim 1$



# *Gaussian Distribution: probabilities*

Recall $$P(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\{-(x-\mu)^2/2\sigma^2\}$$

$$P(-t < x_0 < t) = \frac{1}{\sqrt{2}} \int_{-t}^{t} e^{-z^2/2} dz$$

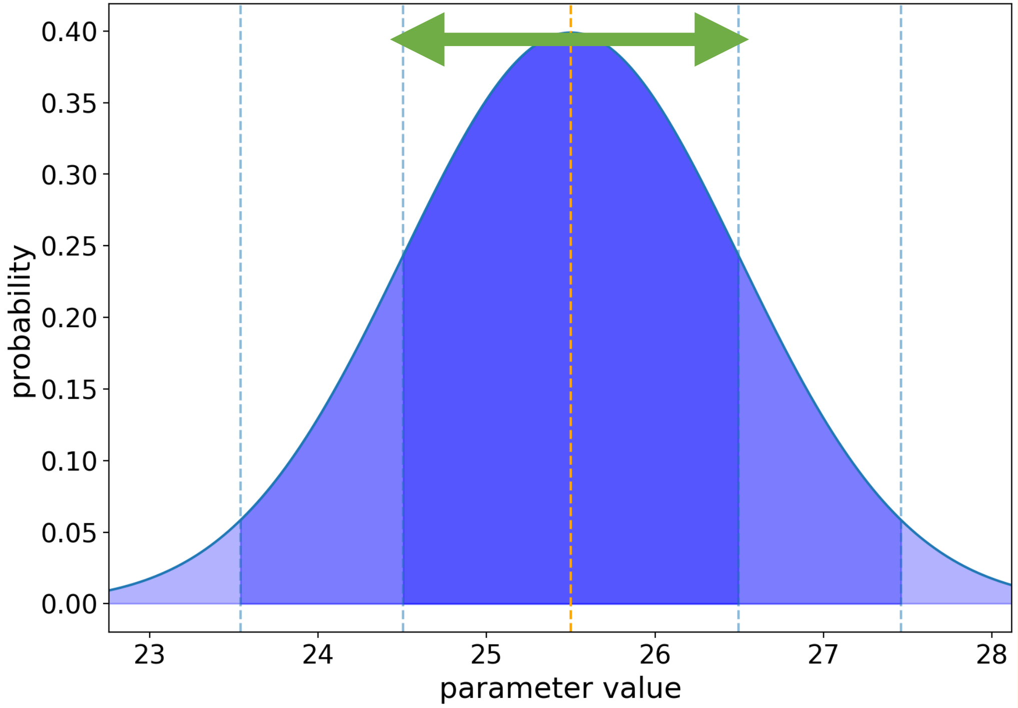
$1\sigma$ : contains **68.3%** of the probability

$2\sigma$ : contains **95.5%** of the probability

$3\sigma$ : contains **99.7%** of the probability

$5\sigma$ : contains **99.99994%** of the probability

area containing 68% of the probability



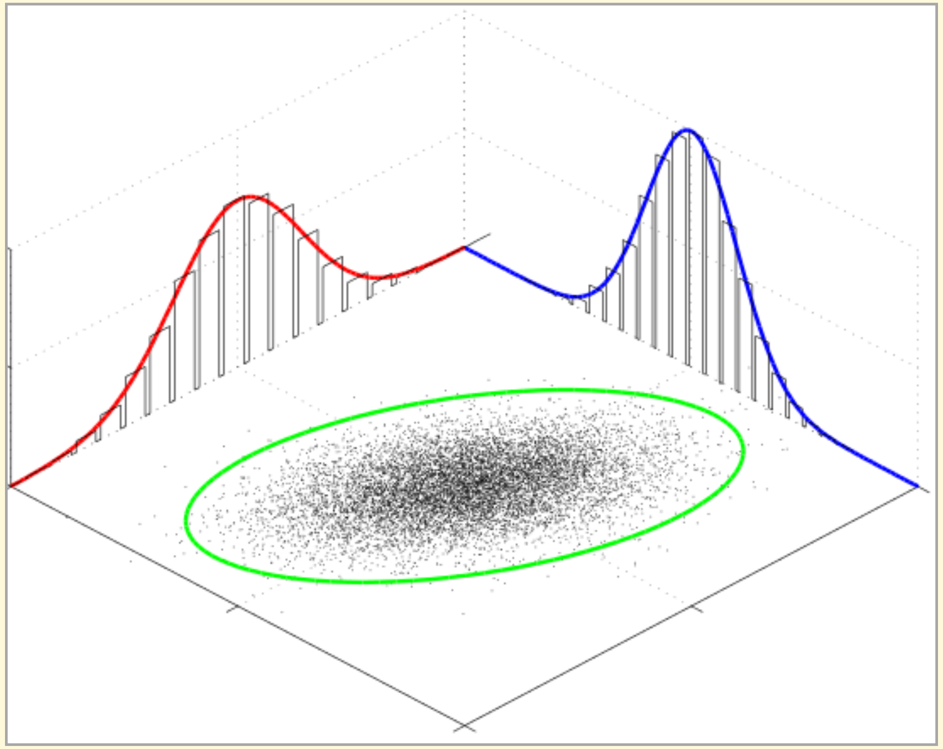

**WARNING**: This process knows nothing about physics: beware of unphysical regions

# Multiple parameters: *Marginalisation*

Distributions are often NOT the function of 1 parameter

$$p(x, y)$$

***NB: Coursework 1: is 'happiness' a function solely of GDP?***



The probability of $x$ is calculated by integrating over $y$

$$p(x) = \int p(x, y)\, dy$$

***NB: We wanted to measure*** $p(\textit{happiness} \mid \textit{GDP})$***, but we measured*** $p(\textit{happiness}, \textit{country} \mid \textit{GDP})$

$$p(\textit{happiness} \mid \textit{GDP}) = p(\textit{happiness}, \textit{UK} \mid \textit{GDP}) + p(\textit{happiness}, \textit{France} \mid \textit{GDP}) + p(\textit{happiness}, \textit{Spain} \mid \textit{GDP}) + \ldots$$

# An aside: *Covariance*

When two measurements are not independent, we call them covariant

For two connected variables ($z = x + y$):
$$\sigma_z = \sqrt{\sigma_x^2 + \sigma_y^2}$$

if $x$ is related to $y$

$$\sigma_z = \sqrt{\left(\frac{\partial z}{\partial x}\sigma_x\right)^2 + \left(\frac{\partial z}{\partial y}\sigma_y\right)^2 + 2\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}cov(x, y)}$$

$$cov(x, y) = \frac{1}{N-1}\sum(x - \hat{x})(y - \hat{y})$$

***NB: Parameter transformations (e.g. $y = \log(x)$) change the functional form of the distribution***

# An aside: *Covariance*

Often expressed in matrix form:

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

Variance in the diagonal terms; covariance in the off-diagonals

# Finding the Distribution

## Kolmogorov-Smirnov (KS) testing

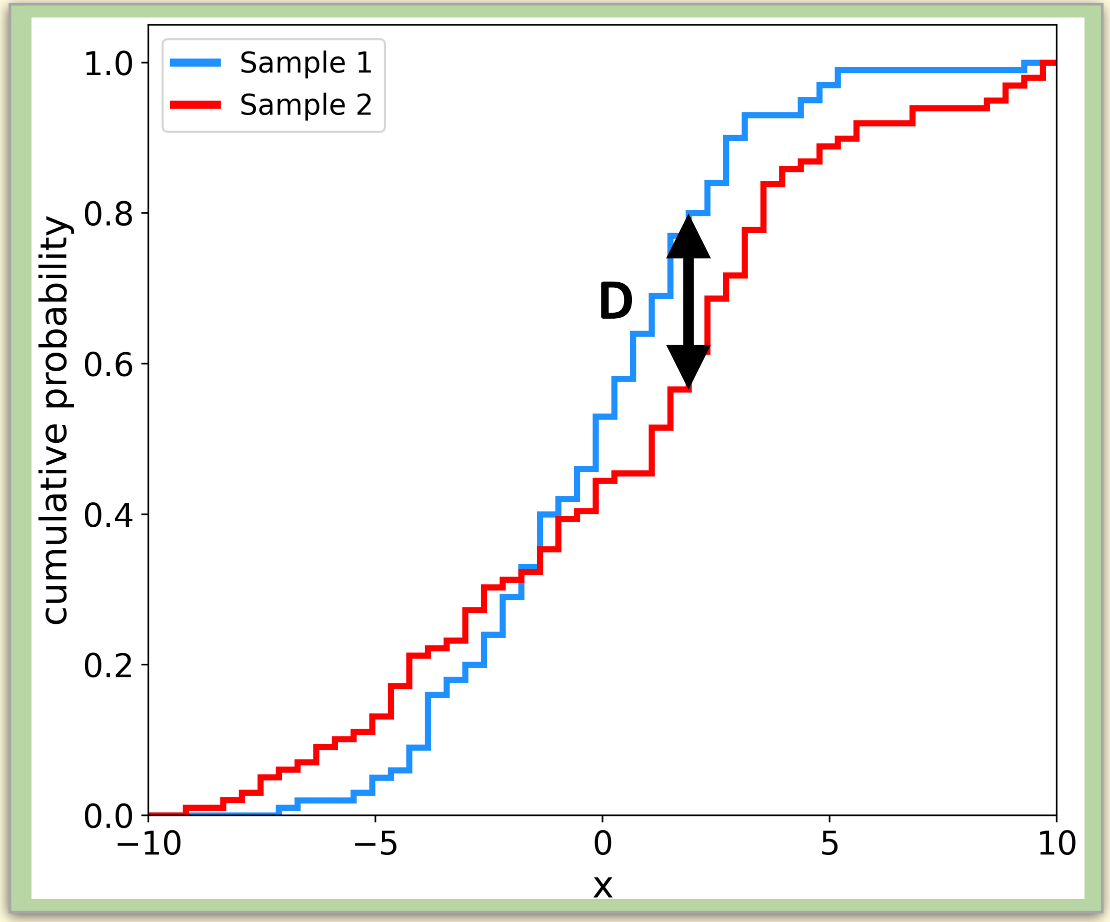
Null hypothesis: the **two distributions are the same**

The maximal distance between

Cumulative Density Functions (CDF)

$$D = \max_{-\infty<x<\infty} |S_N(x) - F(x)|$$

*Recall: a CDF is a rank-ordered PDF.*




No assumptions about the underlying distribution

***NB: See also Anderson-Darling test***

# Testing for Correlations

## Spearman's Rank (SR)

Testing whether **two variables are related**

Compares the ordered rankings (R) of two datasets

$$p_s = 1 - \frac{\sum_{i=1}^{N} [R(x_i) - R(y_i)]^2}{N(N^2 - 1)}$$

Output will be a value between -1 and 1.

Null Hypothesis: **the two variables are uncorrelated**

***NB: A non-parametric version of the Pearson test***

# Likelihoods

Given a set of measurements $x = (x_1, x_2, \ldots)$

the ***likelihood:***

$$\mathcal{L}(\theta\,|\,x) = \prod_{i=0}^{N} P(x_i\,|\,\theta)$$

| | |
|---|---|
| Best-fit value: | The value that maximises $\mathcal{L}$ |
| Plausible values: | Determined by the width of $\mathcal{L}$ |
| Uncertainty: | For a Gaussian distribution $1\sigma$ |

# Log Likelihoods

Given a set of measurements $x = (x_1, x_2, \dots)$

the ***log-likelihood:***

$$\ell(\theta) = \ln \mathscr{L}(\theta) = \sum_{i=0}^{N} P(x_i \,|\, \theta)$$

***Much easier to maximise***

| | |
|---|---|
| Best-fit value: | The value that maximises $\ell$ |
| Plausible values: | Determined by the width of $\ell$ |
| Uncertainty: | $\ell = \ell_{\text{max}} - 0.5$ |

***NB: Assumes Gaussianity***

# Log Likelihoods: the Gaussian case

Given a set of measurements $x = (x_1, x_2, \ldots)$

$$p(x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - \mu_i(\theta))^2}{2\sigma_i^2}\right)$$

$$\mathscr{L}(\theta\,|\,x) = p(x_1\,|\,\theta) * p(x_2\,|\,\theta) * \ldots$$

$$\mathscr{L}(\theta\,|\,x) = e^{-\chi^2/2} * \prod_{i=1}^{N} \frac{1}{\sigma_i} \times (2\pi)^{(-N/2)}$$

$$-2\ln \mathscr{L}(\theta\,|\,x) = \chi^2 + 2\sum_{i=1}^{N} \ln \sigma_i + N \ln(2\pi)$$

***Maximising $\mathscr{L}(\theta\,|\,x)$ is equivalent to minimising $\chi^2$***

# Confidence levels & $\chi^2$

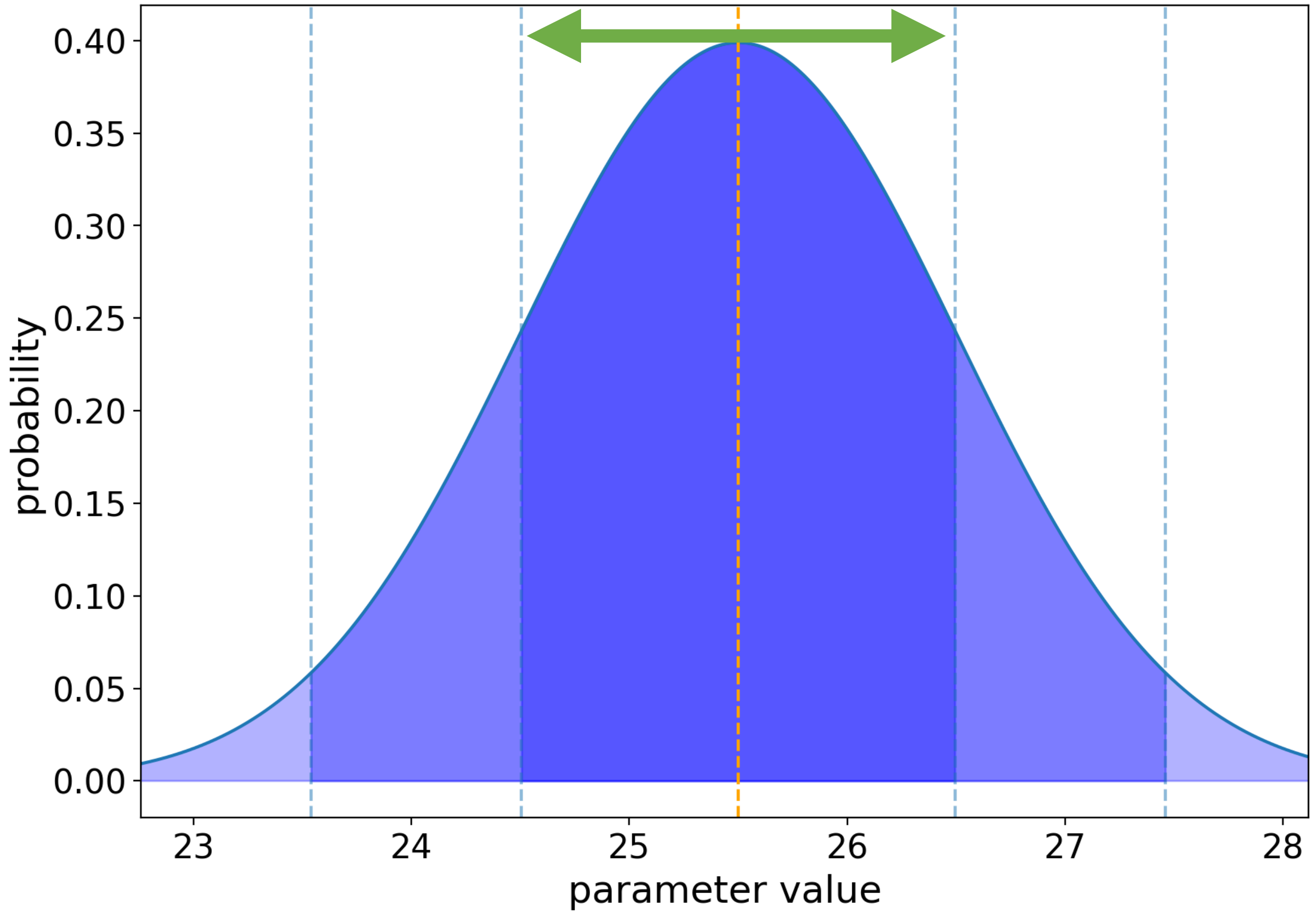
***If our best-fit value has*** $\chi^2 = \chi^2_{min}$:

Contours of equal probability are defined by

$$\chi^2 = \chi^2_{min} + \Delta$$

| CI | | $N_{params}$ 1 | 2 | 3 |
|---|---|---|---|---|
| 68% | $1\sigma$ | 1.0 | 2.30 | 3.53 |
| 95.4% | $2\sigma$ | 4.0 | 6.17 | 8.02 |
| 99.7% | $3\sigma$ | 9.0 | 11.8 | 14.2 |

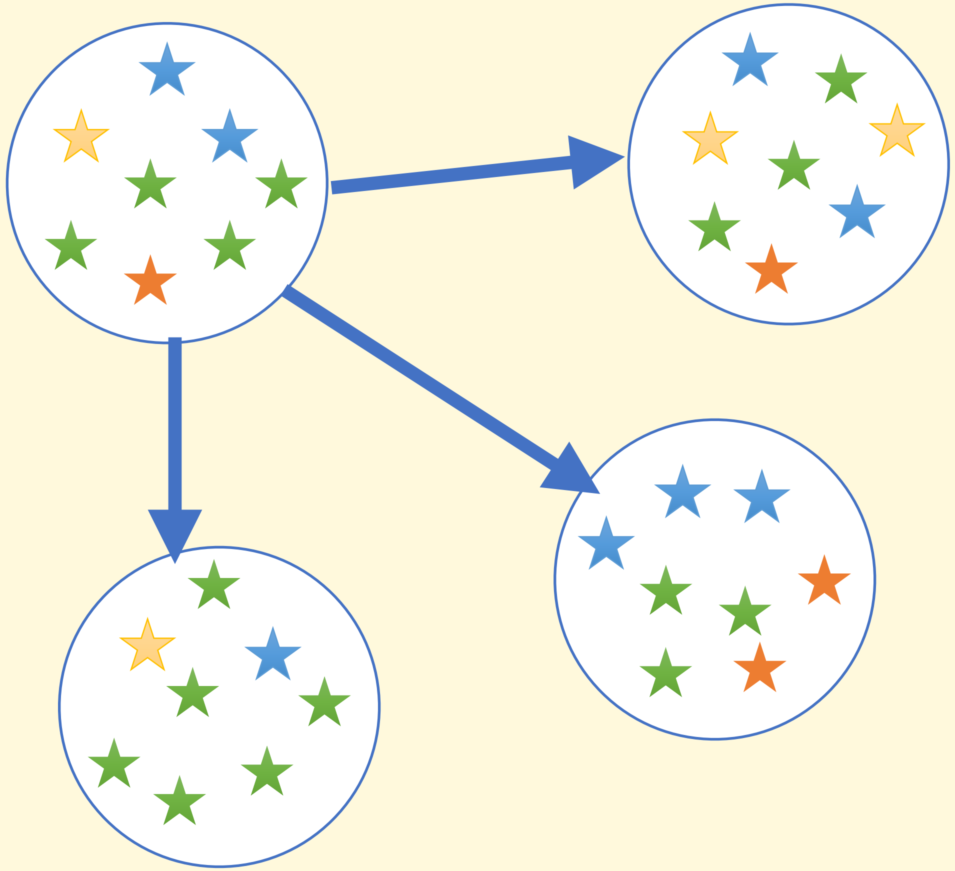
$\chi^2 < \chi^2_{\text{min}} + 1$




**WARNING**: This process knows nothing about physics: beware of unphysical regions

# $\chi^2$ : *estimating uncertainties : bootstrapping*

Resampling the data **with** replacement



(For a dataset with $N$ measurements)

For each of $k$ samples:

> Randomly select $N$ values

> Calculate the $\chi^2$

Calculate the mean, and variance on $\chi^2$

***Does not assume a functional form for the distribution***

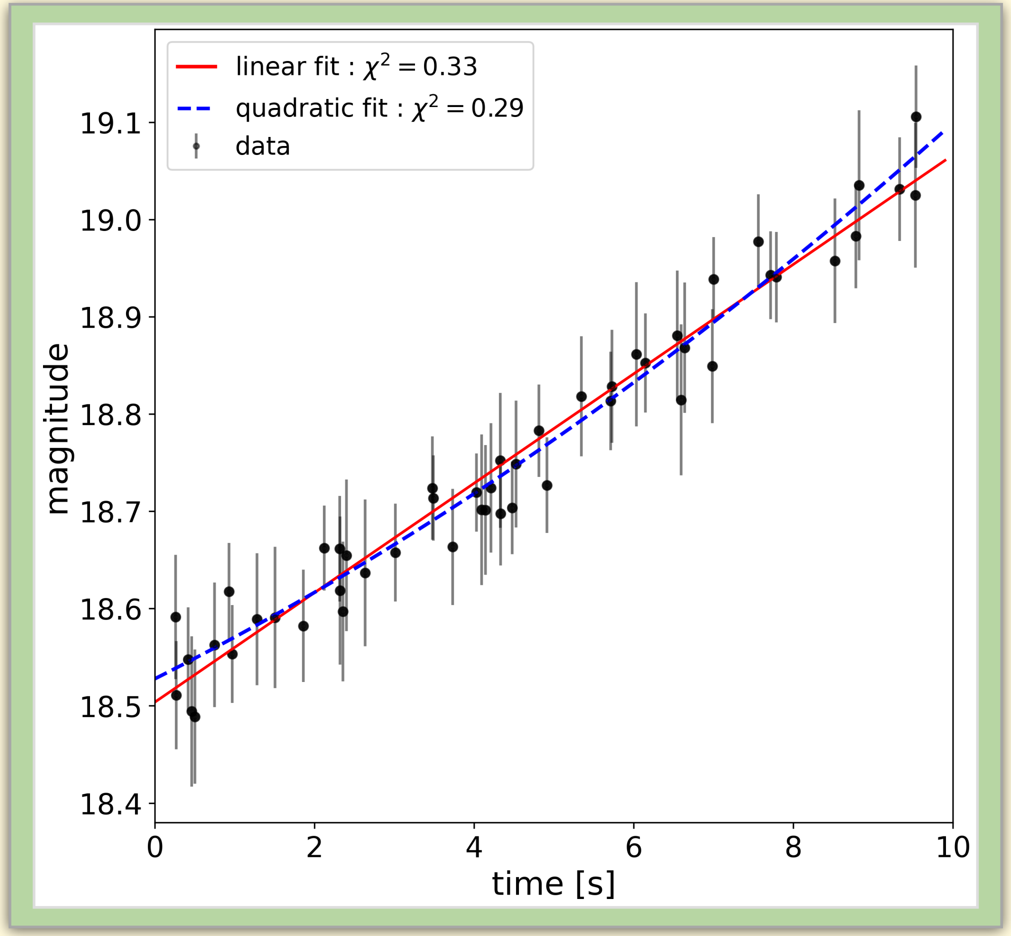
***NB: Can also sample from uncertainties : 'MonteCarlo'***

*See Week 14 for the optimum way to do this*

# Model Selection

Given two models $M_1, M_2$ how do we know which one to choose?

***Recall that for a good fit, we expect: min***$(\chi^2_{red}) \sim 1$




*Akaike Information Criteria (AIC)*

$$\mathrm{AIC} = -2\ln(\mathcal{L}(\theta)) + 2p$$

*Bayes Information Criteria (BIC)*

$$\mathrm{BIC} = -2\ln(\mathcal{L}(\theta)) + p\log N$$

$(p = N_{\mathrm{params}}\,;\, n = N_{\mathrm{data}})$

# *Week 12: Learning outcomes*

## *Today you have learnt*

- How to relate probabilities to measurements: the likelihood
- <u>How to define a null hypothesis</u>
  - How to use probabilities to test our hypothesis
- <u>Three key distributions in Physics</u>
- Key tests of whether our hypothesis (assumptions) are correct
  - Distributions: KS test
  - Correlations: Spearman’s Rank
- <u>Goodness-of-fit metrics: calculating likelihoods</u>
  - Selecting a model given a likelihood

Practical examples on Thursday!