Note - Machine Learning and Statistics

Lecture 1

Precise & accuracy

- **Precise** measurement: the spread of results is "**small**"
- **Accurate** measurement: result is in agreement with "**accepted**" value

Mean, standard deviation, standard error

- **Mean**: where is the measurement centred

$$\overline{x} = \frac{1}{N}(x_1 + x_2 + \cdots + x_n) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- **Standard deviation**: width of the distribution

$$\sigma_{n-1} = \sqrt{\frac{(d_1^2 + d_2^2 + \cdots + d_N^2)}{N-1}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} d_i^2}$$

  where $d_i = x_i - \overline{x}$

- **Standard error**: uncertainty in the location of the centre, α

$$\alpha = \frac{\sigma_{N-1}}{\sqrt{N}}$$

  We should quote our finding as $\overline{x} \pm \alpha$

> Consider an experiment with N number of data points collected:
>
> - the standard deviation is independent of N
>   标准差不受样本量 *N* 的影响, 因为它只关心数据的分散程度
> - the standard error improves with N
>   标准误差随样本量 *N* 的增加而减小, 意味着随着收集更多的数据, 均值估计会更加精确

Random errors, the normal distribution

- **Gaussian or Normal Distribution**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(x-\overline{x})^2}{2\sigma^2}]$$

> Facts: peak centred around mean, symmetric about mean,area under curve equals 1(normalised)

- The error in the error

$$\alpha^{\pm} = \text{error in the error} = \frac{1}{\sqrt{2N-2}}$$

> Bigger sample size:
>
> - lower error in the error
> - more confidence
> - can quote more significant figures
>
> Need $N = 50$ for the error to be known to 10% Need $N > 10k$ for the error to be known below 1%

Reporting results, Confidence limits and error bars

- The five golden rules for reporting results
  1. The best estimate for a parameter is the mean
  2. The error is the standard error in the mean
  3. Round up the error to the appropriate number of significant figures
  4. Match the number of decimal places in the mean to the standard error
  5. Include units
- What is the probability of the data to lie within some multiple of σ?"
  - We need to evaluate the **error function** of the Gaussian distribution(G)

$$Erf(x_1;\overline{x},\sigma) = \int_{-\infty}^{x_1} G(x;\overline{x},\sigma)$$

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

$$\text{CDF}(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$$

$$P(-\sigma \leq X \leq \sigma) = \text{erf}\left(\frac{\sigma}{\sqrt{2}}\right)$$

- The standard deviation is thus used to define a **confidence level** on the data
- If your result and the accepted value differ by:
  - Up to 1 standard error it is in **excellent agreement**
  - Between 1 and 2: **reasonable agreement**
  - More than 3 standard errors: **disagreement**

Poisson distribution

- The conditions under which a Poisson distribution holds are:
  - Events are rare
  - Events are independent
  - Tha average rate does not change with time

$$P(N;\overline{N}) = \frac{\exp(-\overline{N})\overline{N}^N}{N!}$$

> Mean = $\overline{N}$
> Standard deviation = $\sqrt{\overline{N}}$

Chauvenet's Criterion

Chauvenet's Criterion 是一种统计学方法，用于检测和判断数据集中是否存在离群值(outliers)。离群值是指与其他数据点偏差较大的数据点。Chauvenet 的准则基于标准正态分布，提供了一种方法来确定一个数据点是否与数据集的其他部分显著不同。

使用 Chauvenet's Criterion 的步骤：

1. **计算数据的均值(mean, μ)和标准差(standard deviation, σ)。**
2. **计算每个数据点与均值的偏差**：使用以下公式计算每个数据点与均值的标准化差值(即 Z 值)：

   $$Z = \frac{|x_i - \mu|}{\sigma}$$

   其中 x_i 是第 i 个数据点，μ 是均值，σ 是标准差。
3. **计算离群值的概率**：利用正态分布，Z 值代表的是数据点与均值的偏差程度。对于正态分布，计算该数据点的累计概率。这个概率表示数据点在多大程度上可以被视为异常值。
4. **判断数据点是否为离群值**：根据数据点的数量 N，Chauvenet's Criterion 提供了一个门槛。如果一个数据点的概率低于：

   $$P = \frac{1}{2N}$$

   则该数据点可以被视为离群值并被拒绝(REJECT)，否则接受(ACCEPT)。

Lecture 2

Error propagation

Objective:

1. Understanding how to propagate the error is **a vital part of data analysis and reduction**.
   了解如何传播错误是**数据分析和减少错误的重要部分。**
2. Understanding which factors contribute to the limiting error is **a vital part of experimental design**.
   了解哪些因素导致了限制误差是**实验设计的重要部分。**

$$\alpha_{\text{speed}} \neq \alpha_{\text{distance}} + \alpha_{time}$$

$$\alpha_{\text{speed}} \approx \text{speed}\sqrt{(\frac{\alpha_d}{d})^2 + (\frac{\alpha_t}{t})^2}$$

Single-variable functions 单变量误差传播

Functional approach

$$\overline{Z} \pm \alpha_Z = f(\overline{A} + \alpha_A)$$
$$\overline{Z} = f(\overline{A})$$
$$\overline{Z} \mp \alpha_Z = f(\overline{A} - \alpha_A)$$

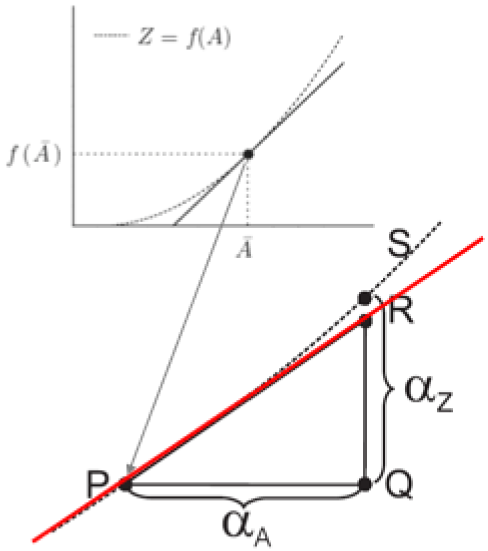
- Valid for every single-varible function

$$\alpha_Z = |f(\overline{A} + \alpha_A) - f(\overline{A})|$$

Calculus-based approach

For $Z = f(A)$:

$$P = (\overline{A}, f(A))$$
$$Q = (\overline{A} + \alpha_A, f(\overline{A}))$$
$$R = (\overline{A} + \alpha_A, f(\overline{A}) + \frac{df}{dA} \times \alpha_A)$$
$$S = (\overline{A} + \alpha_A, f(\overline{A} + \alpha_A))$$
$$f(\overline{A}) + \frac{df(A)}{dA}\alpha_A = f(\overline{A} + \alpha_A)$$



$$\alpha_Z = |\frac{dZ}{dA}|\alpha_A$$

Multi-variable functions 多变量函数的误差传播

Functional approach

Consider a function of two variables, $Z = f(A, B)$ The error of Z has 2 components:

- Change in Z when A is varied and B is constant

$$\alpha_Z^A = f(\overline{A} + \alpha_A, \overline{B}) - f(\overline{A}, \overline{B})$$

- Change in Z when B is varied and A is constant

$$\alpha_Z^B = f(\overline{A}, \overline{B} + \alpha_B) - f(\overline{A}, \overline{B})$$

The total error on Z is obtained via Pythagoras, adding in quadrature:

$$(\alpha_Z)^2 = (\alpha_Z^A)^2 + (\alpha_Z^B)^2 + (\alpha_Z^C)^2 + ...$$

> $$\begin{gathered}(\alpha Z)^2 = \left[f(\bar{A} + \alpha_A, \bar{B}, \bar{C}, \ldots) - f(\bar{A}, \bar{B}, \bar{C}, \ldots)\right]^2 \\ + \left[f(\bar{A}, \bar{\bar{B}} + \alpha_B, \bar{C}, \ldots) - f(\bar{A}, \bar{B}, \bar{C}, \ldots)\right]^2 \\ + \left[f(\bar{A}, \bar{B}, \bar{C} + \alpha_C, \ldots) - f(\bar{A}, \bar{B}, \bar{C}, \ldots)\right]^2 \\ + \ldots\end{gathered}$$

- Calculus approximation

$$(\alpha_Z)^2 = (\frac{\partial Z}{\partial A})^2(\alpha_A)^2 + (\frac{\partial Z}{\partial B})^2(\alpha_B)^2 + (\frac{\partial Z}{\partial C})^2(\alpha_C)^2 + ...$$

Least Squares Method 最小二乘法

The importance of residuals 残差的重要性

残差=实际值−模型预测值

$$R_i = y_i - y(x_i)$$

> 最佳拟合直线应该使所有残差尽可能小。

The goodness-of-fit parameter 拟合优度参数

Determining the optimal values of parameters for a function is called **regression analysis**.
确定函数参数的最优值称为回归分析。

The **best-fit straight line** is the one that is close to as many data points as possible -> residuals will be small
最佳拟合直线是尽可能接近更多数据点的直线 -> 残差会很小

This is quantified by the **goodness-of-fit parameter**, X^2:

- When X^2 is minimised, the probability that we obtain our original set of measurements from the best-fit straight line, is maximised.
  当 X^2 最小化时，我们从最佳拟合直线获得原始测量值集合的概率最大化。

$$X^2 = \sum_i \frac{(y_i - y(x_i))^2}{\alpha_i^2}$$

> What we have just done is called **method of least squares** (= minimising the sum of the squares of the residuals)
> 我们刚刚做的叫做最小二乘法（=最小化残差平方和）
> In the case of the straight line, the best values of slope and intercept are those that minimise the summed differences squared
> 对于直线，斜率和截距的最佳值是最小化平方和的值
> This is derived from what is called **maximum likelihood** together with the **central limit theorem** (assumption: the parent distribution from which we draw the yi values is Gaussian width width given by the standard error, αi)
> 这是从所谓的最大似然和中心极限定理中得出的（假设：我们从中得出 yi 值的父分布是高斯

> 宽度，由标准误差 αi 给出） The y-coordinate we get from the best-fit line equation is the **most probable** value of the parent distribution
> 我们从最佳拟合线方程中得到的 y 坐标是父分布的最可能值
> The probability of obtaining our measurement values from the best-fit line is maximised when χ2 is minimised
> 当 χ2 最小化时，从最佳拟合线获得测量值的概率最大化

- X^2 for data with Poisson errors

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

> i: number of counts(bins)
> O_i: observed number of occurrences(in the i-th bin)
> E_j: expected number of occurrences(in the i-th bin)

> If we have a goot fit, $\alpha_i = \sqrt{E_i} \approx \sqrt{O_i}$

Minimisation 最小化

The best values for **slope** and **intercept** are those that minimise the squares of the differences summed for all data points
斜率和截距的最佳值是最小化所有数据点之差的平方和

$$S = \sum_i (y_i - y(x_i))^2 = \sum_i R_i^2$$
$$S = \sum_i (y_i - mx_i - c)^2$$

$$\frac{\partial S}{\partial m} = -2\sum_i (x_i[y_i - mx_i - c]) = 0$$
$$\frac{\partial S}{\partial c} = -2\sum_i (y_i - mx_i - c) = 0$$

S is a minimum when $\frac{\partial S}{\partial m} = \frac{\partial S}{\partial c} = 0$

The required values of the slope (m) and the intercept (c) are obtained from the two simultaneous equations
斜率 (m) 和截距 (c) 的所需值可通过两个联立方程得出

$$m\sum_i x_i^2 + c\sum_i x_i = \sum_i x_i y_i$$

$$m\sum_i x_i + cN = \sum_i y_i$$

- The solutions for gradient, intercept, and their uncertainties reduce to simple analytic expressions
  截距 c 的公式

$$c = \frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{\Delta}, \alpha_c = \alpha_{CU}\sqrt{\frac{\sum_i x_i^2}{\Delta}}$$

斜率 m 的公式：

$$m = \frac{N\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\Delta}, \alpha_m = \alpha_{CU}\sqrt{\frac{N}{\Delta}}$$

> 共同不确定性:
> Common uncertainty: $\alpha_{CU} = \sqrt{\frac{1}{N-2}\sum_i (y_i - mx_i - c)^2}$

总不确定性参数 Δ

$$\Delta = N\sum_i x_i^2 - (\sum_i x_i)^2$$

non-uniform error bars

Need to perform a **weighted** least-squares fit to take them into account
需要执行**加权**最小二乘法才能将它们考虑在内

$$R_i = \frac{y_i - y(x_i)}{\alpha_i}$$

The sum of the squares of the normalised residuals is called X^2

This is now a weighted fit, and we need to take this into account in the analytic expressions for m, c, α_c, with $w_i = \alpha_i^{-2}$

> Points with small errors are more important!

Lecture 3

Least-squares fit to an arbitrary function

- Arbitrary non-linear function with N parameters

$$y(x) = f(x; a_1, a_2, ..., a_N)$$

- Procedure:
    1. for each value of the independent variable, x_i, calculate $y(x_i)$ using an estimated set of values for the parameters
    2. for each value of the independent variables, calculate the square of the normalised residual, $[{(y_i - y(x_i))}/{\alpha i}]^2$
    3. calculate χ^2(sum the square of the normalised residuals)
    4. minimise χ^2 by optimising the fit parameters

Residuals for a least-squares fit to an arbitrary function

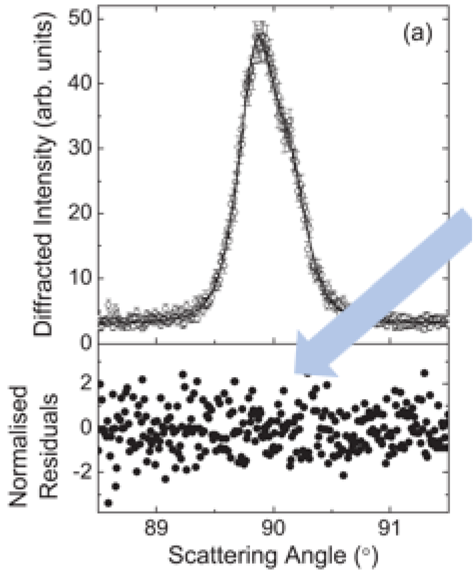
- Voltage across an inductor as a function of time - $V(t; a_1, a_2, ..., a_N)$ is a non-linear function
    - $V(t; V_{bat}, V_0, T, \phi, \tau)$

  - Use model and reasonable estimates of parameters to calculate values of the voltage for a range of times
  - Calculate χ^2
  - Minimise χ^2 by varying all 5 parameters to give best fit

$$V(t) = V_{bgt} + V_0 \cos(2\pi \frac{t}{T} + \phi) \exp(\frac{-t}{\tau})$$

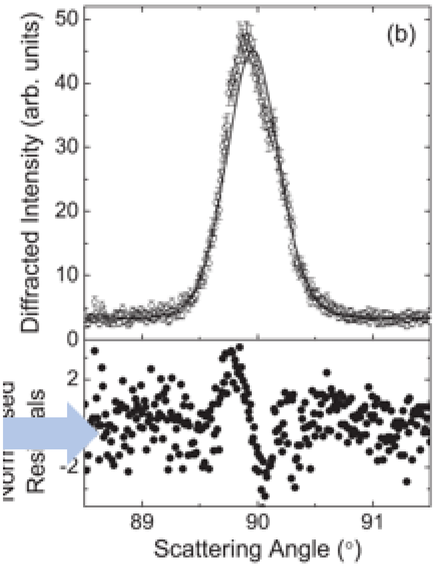
Data and weighted least-squares-fit for X-ray diffraction of copper

- Fit data with a double-peak model:



  - residuals randomly distributed -> good fit

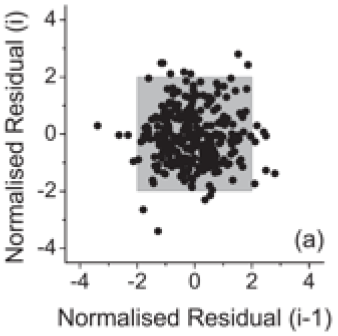
- Fit data with a single-peak model:



  - residuals show structure -> bad fit

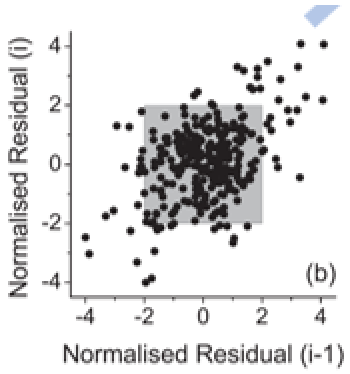
- To better visualise structure in residuals: make a lag plot -> normalised residuals R_i vs lagged residuals R_{i-k}(k is usually 1)

- Good fit



  - Random parttern
  - at least 91% of data points in a 2D box of ± 2 limits

- Bad fit:



  - non-random pattern
  - < 91% of data points in a 2D box of ± 2 limits

Durbin-Watson statistic

The degree of correlation in the lag plot can be reduced to a single numerical value by evaluating the **Durbin-Waston statistic, D**

$$D = \frac{\sum_{i=2}^{N} [R_i - R_{i-1}]^2}{\sum_{i=1}^{2} [R_i]^2}$$

> What does it mean?
>
> - 0 < D < 4
> - D = 0: systematically correlated residuals
> - D = 2: randomly distributed residuals with Gaussian distribution
> - D = 4: systematically anticorrelated residuals

Calculating the error in a least-squares-fit

the error surface

- Remember how χ^2 evolved with the gradient m of a straight line?

- More generic:
  - Function is more complex
  - Non-uniform error bars
  - Goodness-of-fit remains χ^2, but now it evolves over **surface** defined by the fit parameters
- Shape of error surface is important:
  - Many(few) contours on axis of a parameter = high(low) sensitivity of fit to that parameter
  - Tilted ellipses $\rightarrow$ correlation between uncertainties of parameter
  - No tilt $\rightarrow$ no correlation between α_A, α_B
- Investigate shape of error surface via Taylor expansion of χ^2
- Taylor expansion: behavior in close vicinty to a value

$$f(x, a+\Delta a) = f(x,a) + \Delta a \frac{\partial f}{\partial a} + \frac{1}{2}\frac{\partial^2 f}{\partial a^2}(\Delta a)^2 + \dots$$

$$\chi^2(\overline{a}_j + \Delta a_j) = \chi^2(\overline{a}_j) + \frac{1}{2}\frac{\partial^2 \chi^2}{\partial a_j^2}$$

If Δa_j (the deviation from the best-fit value) is similar to uncertainty

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

χ^2 increases by 1
This effectively corresponds to the 1σ contour

Can now express the standard error in terms of the curvature of the error surface

$$\alpha_j = \sqrt{\frac{2}{(\frac{\partial^2\chi^2}{\partial a_j^2})}}$$

Curvature matrix for straight line fit

- Let's stay with straight line fit and introduce the concept of curvature matrix
- We will see later that the uncertainties for N fit parameters can be calculated from the inverse of the curvature matrix, i.e. the error matrix
- For a straight line fit, the χ^2 surface is perfectly parabolic with respect to both variabels, such that:
  - There is only 1 minimum
  - Finding the minimum is easy
  - The curvature matrix has analytic results that let you calculate errors easily$A=\begin{bmatrix}A_{cc}&&A_{cm}\A_{mc}&&A_{mm}\end{bmatrix}$

$$A_{cc} = \sum_i \frac{1}{\alpha_i^2} \quad A_{cm} = A_{mc} = \sum_i \frac{x_i}{\alpha_i^2} \quad A_{mm} = \sum_i \frac{x_i^2}{\alpha_i^2}$$

The χ^2 surface for an arbitrary function

Arbitrary function: the χ^2 surface can be very complicated!

- Multiple local minima
- Need an initial guess for the best-fit parameters(to avoid being trapped in a local minimum)
- The $\Delta\chi^2$ contours can be asymmetric(might not be able limits)
- The elements of the curvature matrix might not have analytic results $\rightarrow$ need numerical techniques to calculate them
- Newton-Raphson
- Technique that numerically solves $f(x) = 0$

  1. First (approximate) solution is x_1
  2. Let $f(x)$ crosses zero at $x_1 + h, f(x_1 + h) = 0$
  3. If h is small, can use Taylor expansion to find second approximation for zero crossing point x_2, $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$
  4. Repeat this to get successively closer approximation,

  $$x_{s+1} = x_s - \frac{f(x_s)}{f'(x_s)}$$

  5. Procedure ends when the zero-crossing point is found(within given tolerance)

- Grid search
- The goodness-of-fit parameter is minimized by changing each parameter in turn, based on input step sizes

  1. Gradient is keptt the same, intercept is increased. If χ^2 increases, the direction of motion across surface is reversed
  2. Parameters are changed in turn until convergence
     Inefficient method, because parameters are changed sequentially

How do fitting programs minimise?

Iterative approaches

- Gradient descent
  - Change all parameters simultaneously, with vector directed towards minimum
    1. Vector $\nabla\chi^2$ is along direction along which χ^2 varies most rapidly
    2. Take steps along this steepest descent util convergence, with the gradient being

    $$\left(\nabla\chi^2\right)_j = \frac{\partial\chi^2}{\partial a_j} \approx \frac{\chi^2\left(a_j + \delta a_j\right) - \chi^2\left(a_j\right)}{\delta a_j}$$

    3. Update rule: $a_{s+1} = a_s - B\nabla\chi^2(a_s)$

    > B is scaling factor

- Second-order expansion

  - An excellent approximation of the local minimum is a second-order expansion in the parameters about the minimum, i.e. perform a Taylor expansion of χ^2 about the set of parameters a_s

$$\chi^2\left(\mathrm{a}_s+\mathrm{h}\right) \approx \chi^2\left(\mathrm{a}_s\right)+\mathrm{g}_s^{\mathrm{T}} \mathrm{h}+\frac{1}{2} \mathrm{h}^{\mathrm{T}} \mathrm{H}_s \mathrm{h}$$

$$\mathfrak{g}_s=\nabla \chi^2\left(\mathfrak{a}_s\right)=\left[\frac{\partial \chi^2}{\partial a_1}, \cdots, \frac{\partial \chi^2}{\partial a_{\mathcal{N}}}\right]^{\mathrm{T}}$$

- Marquardt-Levenburg method

- Combines best features of gradient and expansion approaches

  - Uses method of steepest descent to progress towards minimum when initial guess was far from optimum value
  - Smooth transition to expansion method when goodness-of-fit parameter reduces, and surface becomes parabolic(no multiple local minima)

$$\mathrm{a}_{s+1}=\mathrm{a}_s-\left(\mathrm{H}_s+\lambda \operatorname{diag}\left[\mathrm{H}_s\right]\right)^{-1} \mathrm{g}_s$$

Covariance(error) matrix and uncertainties in fit parameters

- Simply: the curvature matrix, A, is one half of the Hessian matrix

  - It is also an NxN matrix
  - It has components $A_{jk}=\frac{1}{2} \frac{\partial x^2}{\partial a_j \partial a_k}$
  - Off-diagonal terms are related to degree of correlation of parameter uncertainties(they describe the curvature of the surface, remember!)

- Inverse of curvature matrix = covariance(error) matrix

$$[C]=[A]^{-1}$$

And finally: the uncertainty a_j of the parameter a_i is $\alpha_j=\sqrt{C_{jj}}$, so $a_j+a_j=a_j \pm \sqrt{C_{jj}}$

Correlation between uncertainties of fit parameters

- The off-diagonal elements of the covariance matrix are the correlation coefficients

- It's easier to use a dimensionless measure for the correlation matrix:

  - Diagonal elements are all 1
  - Correlation coefficients $\rho_{AB}=\frac{c_{AB}}{\sqrt{c_{AA} c_{BB}}}$

$$\begin{gathered}-1 \leq \rho \leq 1 \\ 0 \text{ when A, B uncorrelated}\end{gathered}$$

- What is the voltage and its error for $f=75 Hz$?

  - Without correlation: $\alpha_V^2=f^2 \alpha_m^2+\alpha_c^2=f^2 C_{22}+C_{11}$
  - With correlation: $\alpha_V^2=f^2 C_{22}+C_{11}+2 f C_{12}$

Lecture 4

Null hypothesis

Null hypothesis: "My data can be described by a Gaussian"

Hypothesis testing: "What is the probability that my data is described by a Gaussian, like I thought?"

Testing the quality of a fit = Testing the null hypothesis

A common statistic used to test the null hypothesis is the χ^2 statistic, out old friend

Degrees of freedom

I f we have measured N independent data points and we are fitting a model with N parameters,

then the number of degrees of freedom, v, is defined as

$$v = N - N$$

= number of independent pieces of information used to estimate a parameter

The last deviation is not free, it is set by the previous ones such that their sum equals 0. To calculate the standard deviation, we thus have one less free variabels. $v = 3 - 1 = 2$

So what is better: to have many df? Or to have just a few df?

Complex situation might require assumptions to make the problem easier, which reduces the df.

In data analysis:
The more data points that are unconstrained, the more robust a statistical estimate of parameters such as the mean, variance and χ^2 become

$$v = N - N$$

Each parameter of a parent distribution estimated from a sample distribution (v_i) reduces the $df(v)$ by 1

The χ^2 probability distribution function

χ^2 is a random variable(it depends on a variety of input parameters, e.g., the input data, the chosen model, the uncertainties, etc)
-> it has a normalised probability distribution function(PDF)

$$X(\chi^2; v) = \frac{(\chi^2)^{(\frac{v}{2}-1)} \exp[-\chi^2/2]}{2^{v/2}\Gamma(v/2)}$$

Facts:

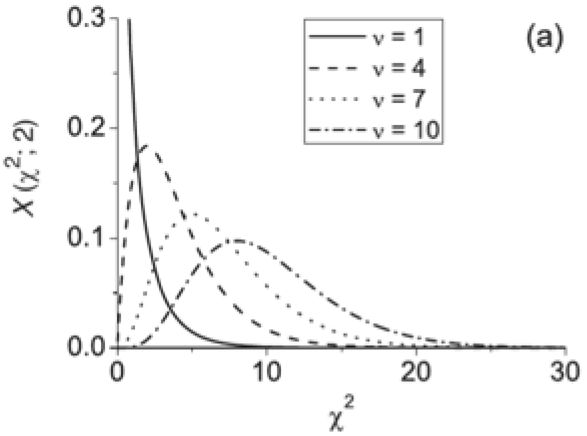
- $X(\chi^2, v)$ is asymmetric(median $\neq$ mode)

- It has a mean(expectation value) of v, and a standard deviation, $\sigma_{\chi^2} = \sqrt{2v}$
- P of obtaining a value of χ^2 between χ^2_{min} and ∞ is the cumulative probability function

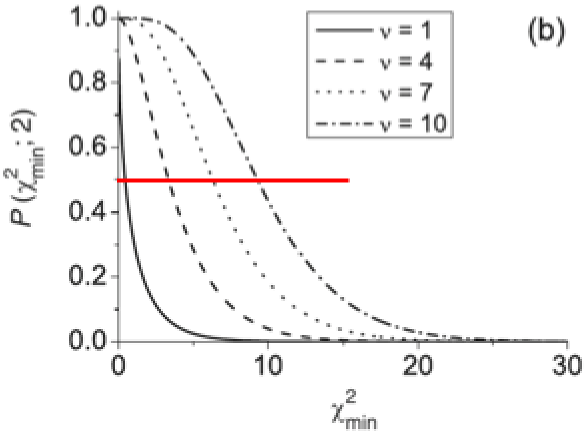
$$P\left(\chi^2_{\min} \leq \chi^2 \leq \infty; v\right) = \int_{\chi^2_{\min}}^{\infty} X\left(\chi^2; v\right) \mathrm{d}\chi^2$$

Using χ^2 as a hypothesis test

Normalized PDF, asymmetric for low ν

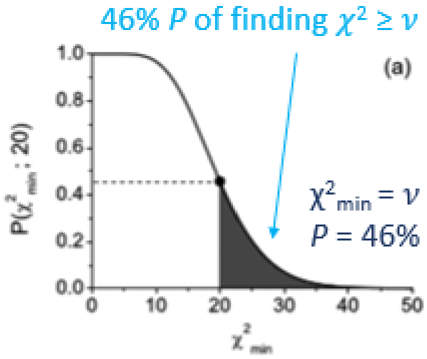


Corresponding CDF



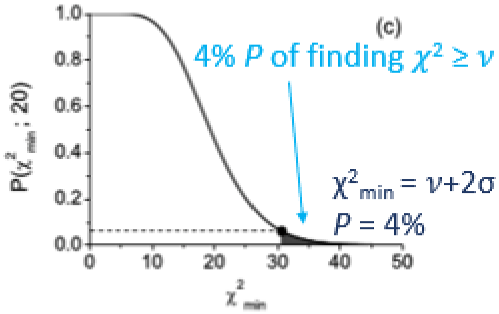
Expectation if the proposed model is in good agreement with data: χ^2_{min} is close to the mean of the χ^2 distribution

For many v, distribution is more symmetric -> mean, median, mode similar, $P(\chi^2_{min} \approx \nu; \nu) \approx 0.5$



If $\chi^2_{\text{min}} \gg v \rightarrow$ probability is small

- is null hypothesis wrong?
- are uncertainties incorrect?



If $\chi^2_{\text{min}} < v \rightarrow$ probability tends towards 1

- Not an indication of improved fit
- Likely, uncertainties are overestimated, which results in unrealistically small χ^2 values

> - For a reasonable fit, the value of $P(\chi^2_{\text{min}}; \nu) \approx 0.5$.
> - If $P(\chi^2_{\text{min}}; \nu) \rightarrow 1$, check your calculations for the uncertainties in the measurements, α_i.
> - The null hypothesis is generally **not rejected** if the value of χ^2_{min} is within $\pm 2\sigma$ of the mean, ν, i.e., in the range
>
> $$\nu - 2\sqrt{2\nu} \leq \chi^2_{\text{min}} \leq \nu + 2\sqrt{2\nu}.$$
>
> - The null hypothesis is **questioned** if $P(\chi^2_{\text{min}}; \nu) \approx 10^{-3}$ or $P(\chi^2_{\text{min}}; \nu) > 0.5$.
> - The null hypothesis is **rejected** if $P(\chi^2_{\text{min}}; \nu) < 10^{-4}$.

The reduced χ^2 statistic

There's a fast way of telling if a null hypothesis should be rejected: reduced χ^2

$$\chi^2_v = \chi^2_{\text{min}}/v$$

> - For a reasonable fit, the value of $\chi^2_\nu \approx 1$.
> - If $\chi^2_\nu \ll 1$, check your calculations for the uncertainties in the measurements, α_i.
> - The null hypothesis is **questioned** if $\chi^2_\nu > 2$ for $\nu \approx 10$.

> - The null hypothesis is **questioned** if $\chi^2_\nu > 1.5$ if ν is in the approximate range $50 \leq \nu \leq 100$.

Brief recap

- **Null hypothesis** = sample distribution is well modeled by the proposed parent distribution ("my model describes the data well")
- χ^2 **test**: find the value of χ^2_{min}, then determine the number of degrees of freedom (df)
- Two numbers to test the validity of the null hypothesis:
    1. Reduced χ^2
    2. Probability P of obtaining χ^2_{min} ≥ to fit value given ν
- Reject the null hypothesis if $\chi^2_\nu > 3$

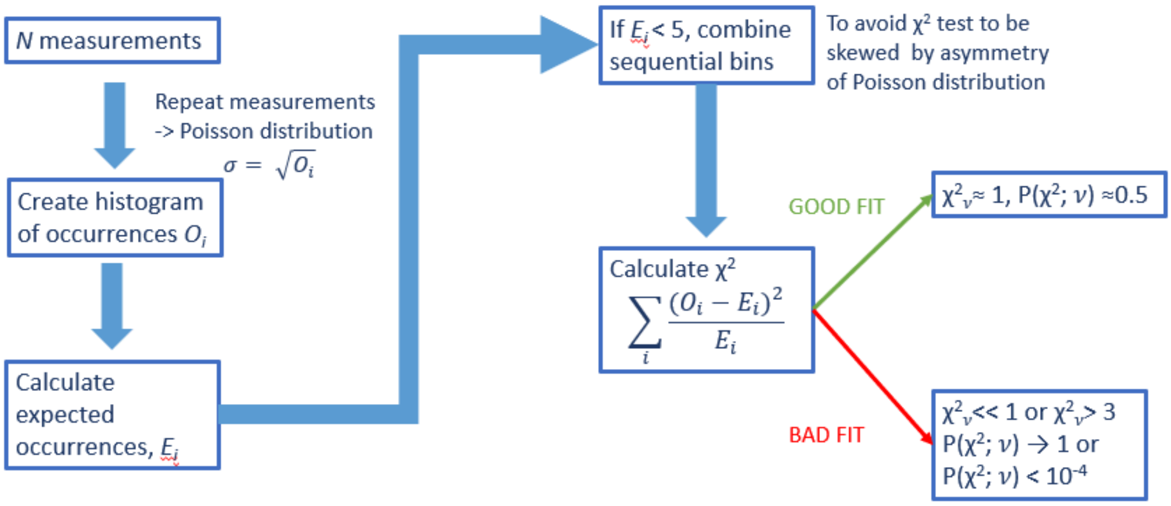
Null hypothesis rejected: what should I do?

- A) Try different models
- B) Gather more data
- C) Do more analysis, e.g., look at residuals

What makes a good fit?

- Two-thirds of the data points should be within one standard error of the theoretical model.
- $\chi^2_\nu \approx 1$.
- $P(\chi^2_{\text{min}};\nu) \approx 0.5$.
- A visual inspection of the residuals shows no structure.
- A test of the autocorrelation of the normalized residuals yields $D \approx 2$.
- The histogram of the normalized residuals should be Gaussian, centered on zero, with a standard deviation of 1.

Testing distributions using χ^2



N measurements
Repeat measurements -> Poisson distribution
$\sigma = \sqrt{O_i}$
Create histogram of occurrences O_i
Calculate expected occurrences, E_i
If E_i < 5, combine sequential bins
To avoid χ^2 test to be skewed by asymmetry of Poisson distribution
Calculate χ^2
$\sum_i \frac{(O_i - E_i)^2}{E_i}$
GOOD FIT
$\chi^2_\nu \approx 1$, $P(\chi^2;\nu) \approx 0.5$
BAD FIT
$\chi^2_\nu << 1$ or $\chi^2_\nu > 3$
$P(\chi^2;\nu) \rightarrow 1$ or
$P(\chi^2;\nu) < 10^{-4}$

Anscombe's quartet

4 data sets:

- Described by same statistic
- Very different distributions
- Illustrate the danger of not inspecting your plots

Benford's Law(aka the first digit law)

The "first digit phenomenon" can be described by the following probability:

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right)$$

"In many real-life numerical sets of data, the leading digit is likely to be small" (wiki).

- 1 appears about 30% of the time.
- 9 occurs about 5% of the time.
- Works best if data spans many orders of magnitude.