STATISTICAL METHODS FOR THE PHYSICAL SCIENCES

Week 4: Significance testing, estimation, maximum likelihood

Significance testing and test statistics

- How do we know whether a hypothesis (e.g. a given model) is a good match to our data?
- To test this, we need to compare a ***test statistic*** obtained from our best-fitting model parameters or assumptions, with the distribution expected if the hypothesis is actually a true representation of the data.
- Recap: a ***statistic*** is a single number that is calculated from random data, e.g. the mean, variance, t-statistic.
- We call a statistic a test statistic when we use it to test the ***significance*** of our hypothesis.
- We usually frame the test in terms of a ***null hypothesis***, usually denoted H_0, the falsification of which is 'interesting' to us.
- The significance gives us a probability that the data are satisfactorily explained by the null hypothesis.
- In some sense it corresponds to our level of surprise that we should get the data we have, assuming that we initially believed the null hypothesis.

Thought experiment: red and green sweets

- Our null hypothesis is that the bag of sweets contains equal numbers of red and green sweets.
- We are allowed to draw 10 sweets from the bag with replacement.
- We draw 8 green sweets. Should we be surprised?
- Our null hypothesis implies that, if *X* is the number of green sweets drawn:

$$X \sim \text{Binom}(n = 10, \theta = 0.5)$$

- We can use the number of green sweets drawn as our test statistic.
- What is the probability that we draw *8 or more green sweets?* (because we would be even more surprised if we drew 9 or 10 sweets!):

$$\Pr(x \geq 8 | \theta = 0.5) = 0.0547$$

- What about 9, or even 10 green sweets?

$$\Pr(x \geq 9 | \theta = 0.5) = 0.011 \qquad \Pr(x = 10 | \theta = 0.5) = 0.00098$$

- If we draw 10 green sweets we are either very lucky, or our null hypothesis is wrong!

Significance testing or 'goodness-of-fit test'

1. We first define our null hypothesis H_0
2. To perform a significance test we must define our test statistic $T(\mathbf{x})$ which is a function of the data and whose *sampling distribution* can be calculated for a given H_0.
3. We calculate the observed value of the test statistic: $T_{\rm obs} = T(\mathbf{x}_{\rm obs})$
4. We then calculate $p = \Pr(T \geq T_{\rm obs}|H_0)$ using $p(T|H_0)$:

$$p = \Pr(T \geq T_{\rm obs}|H_0) = \int_{T_{\rm obs}}^{\infty} p(T|H_0)dT$$

5. If H_0 is true, p is distributed uniformly between [0,1] (it is related to the CDF of T!).

In some cases, we may want to consider `2-sided' tests, where ***unusually high or low*** values of T may be considered significant

The meaning of *p* values and sigmas

- *p* is often called the ***observed significance*** or the ***confidence level*** (i.e. in the null hypothesis).
- The question of what *p* is considered acceptable to reject a model depends on the model in question (e.g. how big a deal is rejection likely to be?), e.g.:

$$p \leq .05, \quad p \leq .01, \quad p \leq 0.001$$

These probabilities correspond to the probability of a normally distributed variable deviating from the mean by respectively:

$$\sim 2\sigma, \quad \sim 2.5\sigma, \quad \sim 3.3\sigma$$

- It is quite common in the physical sciences to quote *p* values directly in terms of normal distribution 'sigmas' (even if the test statistic itself isn't normally distributed!), which map directly on to a probability:

$$\begin{aligned} 1\sigma &\rightarrow p = 0.317 \\ 2\sigma &\rightarrow p = 0.046 \\ 3\sigma &\rightarrow p = 0.0027 \\ 4\sigma &\rightarrow p = 3.18 \times 10^{-5} \\ 5\sigma &\rightarrow p = 6 \times 10^{-7} \end{aligned}$$

Parametric and non-parametric tests

- *Parametric* tests make implicit assumptions about the shape of a distribution used to describe the underlying population (e.g. a normal distribution), and about the form or parameters of the distribution (e.g. mean and variance).
  - Examples: the sample mean, t-test, Chi-squared test, Pearson's *r* correlation coefficient
  - Advantages: Allows more information to be obtained, e.g. relative difference in means, effective model tests etc.
  - Disadvantages: Strongly dependent on assumptions being true; sensitive to outliers

- *Non-parametric* tests make no or few assumptions about the underlying population distribution and its parameters.
  - Examples: the sample median, K-S test, Spearman's rho correlation co-efficient
  - Advantages: Minimal assumptions, not very sensitive to outliers
  - Disadvantages: less statistically powerful than parametric when data are normally distributed (harder to get a significant result from a real effect); results are hard to interpret (e.g. 'relative rankings' rather than actual quantities)

Parametric: Student's t test

- Student's *t*-statistic for comparison of data with n values with a precisely known mean:

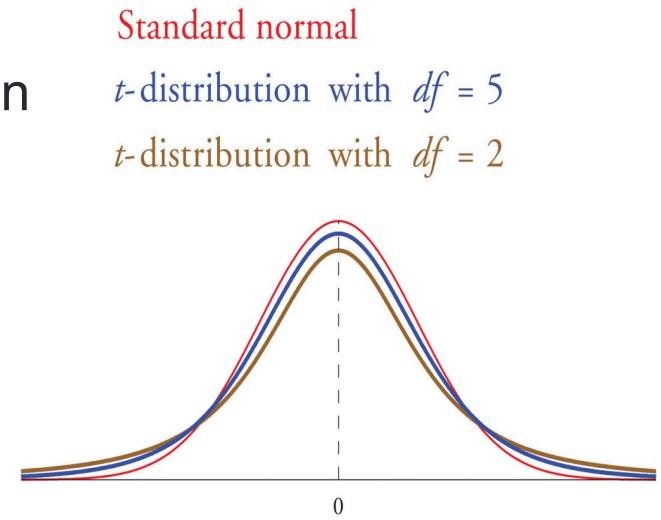
$$t = \frac{observed\ difference\ in\ means}{standard\ error} = \frac{\overline{x} - \mu}{\sqrt{s_x^2/n}}$$

where $\overline{x}$ and s_x^2 are the mean and variance of the data (*sample mean and variance*) and μ is the known mean (*population mean*).

- If the sample mean is normally distributed about the known mean (i.e. the null hypothesis is that: $\mathrm{E}[\overline{x}] = \mu$), then:

$$p(t|H_0) = t_\nu \qquad (\text{where } \nu = n - 1)$$

- The *t*-test is a two-sided test: the value of *t* can be positive or negative, what counts for the significance is the absolute value.
- Versions of *t* also exist for comparing the means of two or more samples.



Parametric: Pearson's chi-squared test

- If the data are normally distributed about the expected values, then:

$$X^2_{\rm min} = \sum_{i=1}^{n} \frac{(observed-expected)^2}{variance}$$
$$= \sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{\sigma_i^2} = \sum_{i=1}^{n} \Delta_i^2$$

Karl Pearson
1857-1936

gives the squared weighted ('standardised) residuals.

- These are distributed so that, if H_0 is correct:

$$p(X^2_{\rm min}|H_0) \sim \chi^2(\nu)$$

where ν is the number of degrees of freedom. For *n* data points and *m* free parameters in the model: $\nu = n - m$

- This test is *one-sided* - unusually low values of chi-squared mean something else (usually that the size of your error bars is overestimated!). A good rule of thumb is that a good fit gives a chi-squared which is close to the d.o.f.,

  i.e. the ***reduced chi-squared*** $\chi^2_\nu/\nu \simeq 1$

Non-parametric: Kolomogorov-Smirnov (K-S) test

- *Empirical distribution function*: sample equivalent of the population cdf:

Heaviside step function: increments edf by 1 for each x_i

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} H(x - x_i)$$

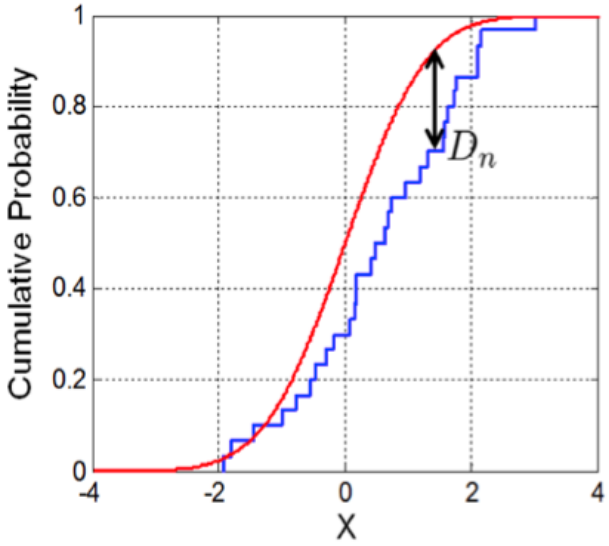
- Calculate maximal distance D_n between edf and cdf

$$D_n = \sup_x |F_n(x) - F(x)|$$

'supremum'

- If $F(x)$ is the true cdf, $D_n\sqrt{n}$

follows a **Kolmogorov distribution** for large n

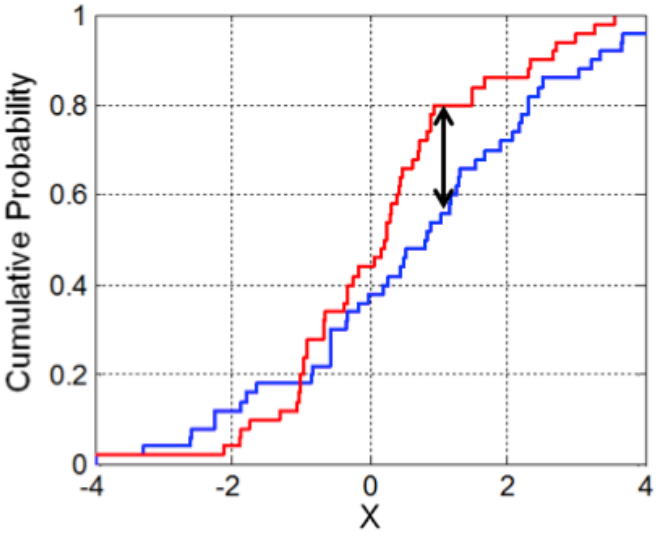


Non-parametric: 2-sample K-S test

What if we want to compare two samples of a population, to determine if they are consistent with being drawn from the same unknown underlying cdf?

Calculate maximal distance between two edfs:

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|$$



General K-S test caveats:

- K-S test of the cdf shape is not valid if the parameters of cdf tests are determined from the same data used to estimate the cdf parameters: the expected cdf must be known (e.g. either from the physics or a different data set)
- K-S test is sensitive to global differences between cdf and edf, or two edfs (e.g. which produce different sample means) but is less effective at uncovering small differences in distribution tails.

Adjusting significance by the number of trials

- When doing a statistical test it is important to be aware of how many trials we have carried out. Imagine we conduct *n* trials, e.g. we measure n samples of the same population.
- We detect in a single trial some effect (e.g. significant deviation of the mean from the expected value), with probability (*p*-value) P_1 that the effect is just due to chance.
- We should ask the question – assuming the effect is not real, what is the chance that I would ***not*** get a fake detection in n trials?

$$P_{\text{no-fake}} = (1 - P_1)^n$$

- So we actually have a probability that the effect is fake across all trials of:

$$P_{\text{all}} = 1 - (1 - P_1)^n$$

Fitting models to data

- Frequently in scientific data analysis we have to fit a model to our data.
- E.g. a *physical model* to relate response variables (i.e. observed data) to explanatory variables (which we control).
- We also need a *statistical model* to account for random error in our data, either observational or because the data themselves sample an intrinsically random process.
- The model predictions depend on the values of the model parameters:
  - *Simple hypothesis*: no unknown terms
  - *Composite hypothesis*: one or more unknown terms known as the model's *free parameters*
- How do we estimate the best-fitting values of these unknown terms?
- The technique to do this is called ***maximum likelihood estimation***

Case study: Rutherford & Geiger data: single data point

- The model is that there is a constant average rate of scintillations, and that the observed data follow a Poisson distribution.
- One free parameter, the expected rate *λ*.
- The ***likelihood function*** of an observed variable $x = x_{\rm obs}$ is a probability distribution function with *x* fixed. E.g. for the Poisson case, the likelihood for $x_{\rm obs} = 3$ is:

$$p(x|\lambda) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} \longrightarrow l(\lambda) = \frac{\lambda^3 \mathrm{e}^{-\lambda}}{3!}$$

Since the data are known we call this a likelihood not probability function, i.e. when the pdf/pmf formula is used as a function of the 2nd (conditional) argument not the first!

- The best estimate of the parameter is that which maximises the likelihood – i.e. it is the *mode* of the likelihood function.

$$\frac{\partial l(\lambda)}{\partial \lambda} = \left(x_{\rm obs}\lambda^{x_{\rm obs}-1} - \lambda^{x_{\rm obs}}\right)\frac{\mathrm{e}^{-\lambda}}{x_{\rm obs}!} = 0$$

i.e. in this case our ***maximum likelihood estimate*** (***MLE***) of *λ* (denoted by a 'hat') is: $\hat{\lambda} = x_{\rm obs}$

also should check that: $\frac{\partial^2 l(\lambda)}{\partial \lambda^2} < 0$

(for several maxima we take the largest value of *l*)

Case study: Rutherford & Geiger data: many data points 1

- Now consider a set of measurements from the Rutherford & Geiger data:

$$x_i (i = 1, 2, \ldots, n)$$

- We can write the data as a vector (using bold type to denote a vector quantity):

$$\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$$

- Now we can use the multiplication rule to find the probability, assuming that the measurements are independent (this is clearly okay for Poisson data!):

$$\begin{aligned} p(\mathbf{x}|\lambda) &= p(x_1, x_2, \ldots, x_n|\lambda) \\ &= p(x_1|\lambda) \times p(x_2|\lambda) \times \ldots \times p(x_n|\lambda) \\ &= \prod_{i=1}^{n} p(x_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} \mathrm{e}^{-\lambda}}{x_i!} \end{aligned}$$

- The likelihood function is this function evaluated for λ given the data:

$$\mathbf{x} = \mathbf{x}_{\mathrm{obs}}$$

Case study: Rutherford & Geiger data: many data points 2

- It is generally easier to find the mode of $l(\lambda)$ by taking the logarithm of the likelihood (here it makes sense to use natural log):

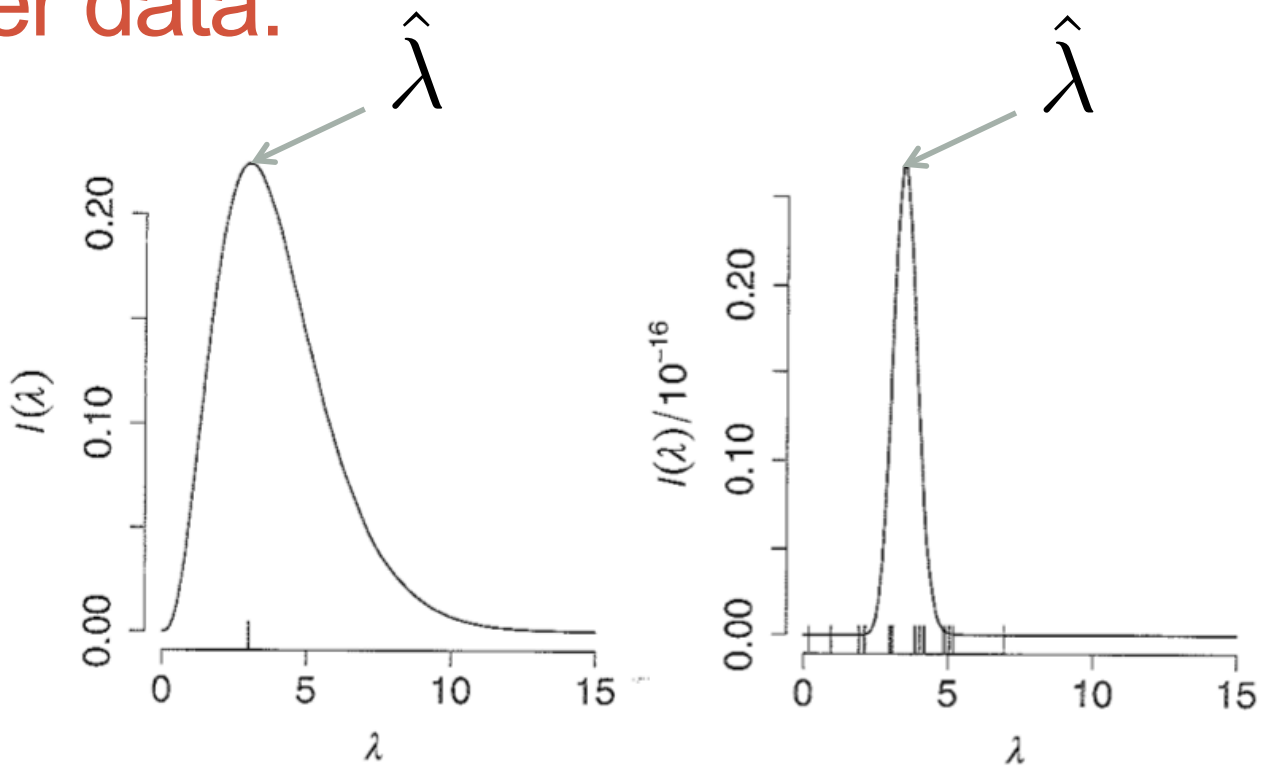
$$\begin{aligned} L(\lambda) &= \ln\left[l(\lambda)\right] = \ln[p(\mathbf{x}|\lambda)] \\ &= \ln\left(\prod_{i=1}^{n} \frac{\lambda^{x_i}\mathrm{e}^{-\lambda}}{x_i!}\right) = \sum_{i=1}^{n} \ln\left(\frac{\lambda^{x_i}\mathrm{e}^{-\lambda}}{x_i!}\right) \\ &= \sum_{i=1}^{n} \left[x_i \ln(\lambda) - \lambda - \ln(x_i!)\right] \\ &= \left(\sum_{i=1}^{n} x_i\right) \ln(\lambda) - n\lambda - \sum_{i=1}^{n} \ln(x_i!) \end{aligned}$$

  - Thus we have:

$$\left.\frac{\partial L(\lambda)}{\partial \lambda}\right|_{\lambda=\hat{\lambda}} = \left(\sum_{i=1}^{n} x_i\right) \frac{1}{\hat{\lambda}} - n = 0 \Rightarrow \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_{i,\mathrm{obs}} = \overline{x}$$

i.e. the MLE of λ for Poisson-distributed data is the mean observed value!

Maximum likelihood with Rutherford & Geiger data:



Likelihood functions for 1 and 20 measurements of counts/interval

Note that the amplitudes of the likelihood function can become very small for many measurements – even for continuous data, one particular sequence of values can be very improbable. But what counts is the position of 'maximum probability' for such values…

General maximum likelihood estimation

- Consider a *physical model* relating a response variable *y* to some explanatory variable *x*. We have *n* measurements of *y* for corresponding values of *x*: $x_i = x_1, x_2, \ldots, x_n$
- We can write both sets of values as vectors, $\mathbf{x}$ and $\mathbf{y}$
- The model is not completely specified, some parameters are unknown. The *M* model parameters can also be described by a vector:

$$\boldsymbol{\theta} = \theta_1, \theta_2, \ldots, \theta_M$$

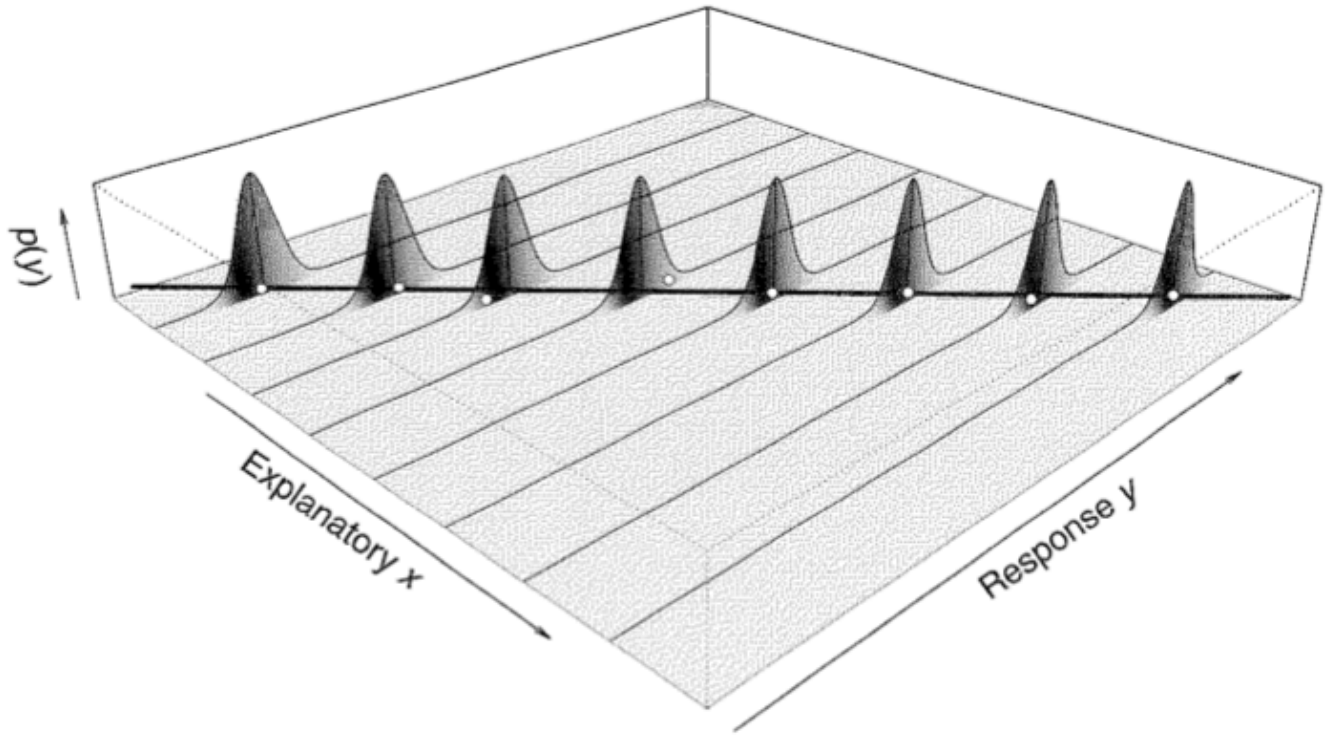
- The model describes our expectation value of *y* for a given *x* and the model parameters *:* $\mathrm{E}[y] = f(x, \boldsymbol{\theta})$
- The *statistical model* gives us the probability distribution of $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$
- The likelihood function is:

$$l(\boldsymbol{\theta}) = p(y_1, \ldots, y_n|\mathbf{x}, \boldsymbol{\theta}) = p(y_1|x_1, \boldsymbol{\theta}) \times \ldots \times p(y_n|x_n, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta})$$

- And the ***log-likelihood function*** is:

$$L(\boldsymbol{\theta}) = \ln l(\boldsymbol{\theta}) = \ln \left(\prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta}) \right) = \sum_{i=1}^{n} \ln \left[p(y_i|x_i, \boldsymbol{\theta}) \right]$$

Visualising MLE



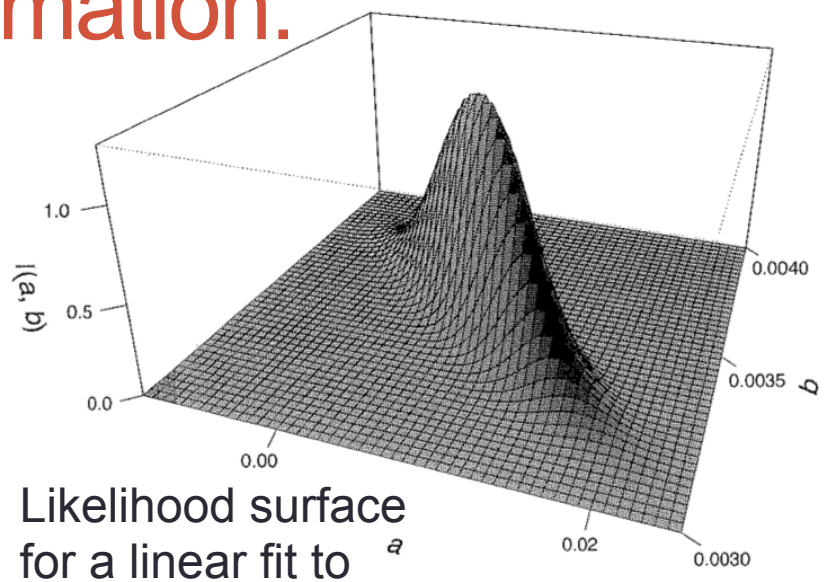
- The ‘physical’ model is shown as a thick black line
- The statistical model is superimposed as individual pdfs (in this case, they are normally distributed)

Maximum likelihood estimation: considerations

- The partial differentials of the likelihood w.r.t. each parameter are known as the ***scores*** (in vector calculus this quantity is known as the *Jacobian*):

$$U(\boldsymbol{\theta}) = \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_1}, \cdots, \frac{\partial L(\boldsymbol{\theta})}{\partial \theta_M} \right)$$

- i.e. $U(\boldsymbol{\theta}) = \nabla L$
- The MLE corresponds to the point where the scores are zero: $U(\hat{\boldsymbol{\theta}}) = (0, \ldots, 0) = \mathbf{0}$



Likelihood surface for a linear fit to Reynolds' data

- Maximisation can be done with a variety of computational methods.
- But sometimes the ML surface is too complex (too many parameters, too complex a physical/statistical model) – maximum can be found with 'brute force': try values of parameters and map out the ***likelihood surface*** (can be done more efficiently with ***Markov Chain Monte Carlo***)
- If data values are not statistically independent, covariance can be used to account for this in the ML estimation process (not covered in this course).

Weighted least squares

- Now consider the case where the response variable *y* is normally distributed about an expectation value, which is some function of the response variable *x* and the model parameters:

mean is: $\mu_i = \mathrm{E}[y_i] = f(x_i, \boldsymbol{\theta})$ and variance is: σ_i^2

- So the probability is:

$$p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(y_i - \mu_i)^2}{2\sigma_i^2}\right]$$

- And log-likelihood:

$$L(\boldsymbol{\theta}) = \ln\left[p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma^2})\right] = -\frac{1}{2}\sum_{i=1}^{n} \ln[2\pi\sigma_i^2] - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2}$$

- We can define a new statistic, $X^2(\boldsymbol{\theta})$

$$X^2(\boldsymbol{\theta}) = -2L(\boldsymbol{\theta}) + const = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2}$$

From weighted least squares to chi-squared minimisation

$$X^2(\boldsymbol{\theta}) = -2L(\boldsymbol{\theta}) + const = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2}$$

- ***Minimising*** $X^2(\boldsymbol{\theta})$ is equivalent to maximising $L(\boldsymbol{\theta})$ or $l(\boldsymbol{\theta})$
- It is the sum of squared residuals, i.e squared data-model variations, weighted by the precisions of each measurement ($1/\sigma^2$).
- Because the weighted residuals are distributed as standard normals, the distribution of $X^2(\boldsymbol{\theta})$ is a chi-squared distribution, also written as $\chi^2(\boldsymbol{\theta})$, with degrees of freedom $v = n - m$, where m is the number of free parameters in the model.
- If the precision on each measurement is identical, we have:

$$X^2(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} [y_i - \mu_i(\boldsymbol{\theta})]^2$$

Thus, minimising $X^2(\boldsymbol{\theta})$ is equivalent to minimising the SSE (see week 1 lecture notes and chapter 3 of Vaughan).

Chi-squared fitting and goodness-of-fit

- Formally, minimising weighted least-squares (also known ‘colloquiually’ as *chi-squared fitting*) can be done in the same way as maximising log-likelihood.
- ***If*** the best-fitting model is a true description of the data (with normally-distributed errors) ***then*** the corresponding *p*-value, also called a goodness-of-fit, can be determined using Pearson’s chi-squared test with the obtained $X^2(\boldsymbol{\theta})$ and degrees of freedom.
- But remember that very high *p*-values are also suspicious! They could indicate that the error bars on the data are too large.