STATISTICAL METHODS FOR THE PHYSICAL SCIENCES

Week 5: Hypothesis testing and confidence intervals (part 1)

Hypothesis testing: fixed-level tests and decisions

- A hypothesis test is like a fixed-level significance test: we pre-specify the significance *α* at which we will reject our null hypothesis. Rather than thinking about *p*-values, we are interested in a test-statistic *T* for which we can define a critical value, e.g. for a 1-sided test:

$$\alpha = \int_{T_{\mathrm{crit}}}^{\infty} p(T|H_0)dT$$

- The procedure is:

1. Define a null hypothesis (H_0) and an alternative hypothesis (H_1).
2. Define a test-statistic whose sampling distribution can be calculated assuming each hypothesis and is different under each hypothesis.
3. Choose a significance level *α.*
4. Calculate the critical value of the test statistic T_{crit}
5. Calculate the observed value of the test statistic T_{obs}
6. Reject the null hypothesis if $T_{\mathrm{obs}} \geq T_{\mathrm{crit}}$ (and accept the alternative if viable), otherwise accept the null hypothesis.

Type I and type II errors

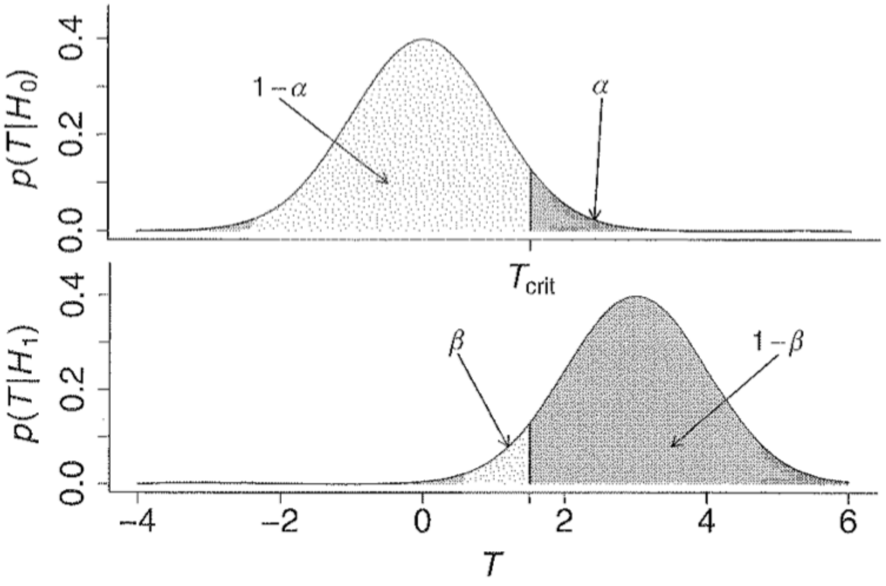
- A type I error is when we reject the null hypothesis when it is in fact true (i.e. a significant but wrong result, a false positive).
- A type II error is when we accept the null hypothesis when it is in fact false (i.e. a non-significant but wrong result, a false negative).
- Smaller values of *α* give a lower risk of making a type I error, but more risk of a type II error.
- If we assume the alternative hypothesis, H_1, is true, we can define the probability of a type II error thus:

$$\beta = \int_{-\infty}^{T_{\rm crit}} p(T|H_1)dT = 1 - \int_{T_{\rm crit}}^{\infty} p(T|H_1)dT$$

- 1-*β* is called the ***power*** of the test (it gives the probability of correctly selecting the alternative hypothesis).
- One should choose a test statistic for which the distributions $p(T|H_1)$ and $p(T|H_0)$ are very different, to maximise the chance of differentiating between the hypotheses.

Hypothesis testing example: upper limits

- What if we want to set an upper limit on a non-detection of something, against some background? (e.g. an astronomical source or a hypothetical particle)
- Here the relevant quantity is *β* – what is the probability that the source/ particle is there in the data but we have missed it? (false negative: type II error)
- We can choose *β* according to how stringent we want to be, e.g. 3-sigma upper limit, 5-sigma upper limit…



- E.g. our test statistic *T* may be our photon counts in our detector
- H_0 is the hypothesis that there is only background
- H_1 is the hypothesis that there is an 'interesting' signal (e.g. source, or particle)

Interpreting test results 1

- Statistical significance does not mean practical significance:
    - With good enough data, any hypothesis can be ruled out.
    - It doesn’t mean the hypothesis is not useful/valid within certain limits.
    - E.g. Mercury’s orbital precession vs. Newtonian gravity
- Systematic errors:
    - With very high-quality data or poorly calibrated experiment, systematics can produce a significant result even when the null hypothesis is valid.
    - A corollary: there is no point in improving the statistical quality of data beyond the point where systematics dominate the error.
- Lack of significance does not mean the null hypothesis is true (absence of evidence is not evidence of absence):
    - The data may not be sensitive enough for distinguishing a wide range of null hypotheses, especially if the data set or difference in hypotheses predictions is small.
    - Choose your statistical test to suit your null hypothesis.
- If a null hypothesis test is significant, it does not mean the alternative is favoured:
    - Maybe the null hypothesis is an inappropriate choice, e.g. analogous to a ‘straw man’ argument).

Interpreting test results 2

- Beware of overfitting data!
  - Occam's razor, keep things simple while consistent with the data (Einstein: "make things as simple as possible, but not simpler").
  - Don't add free parameters to a model when the data are consistent (i.e. a *reasonable p*-value) with fewer parameters.
- Beware of searches for significance:
  - Be wary of data trawls and publication bias, or trying many different statistical tests on the same data until a 'significant result' emerges.
- Do not test a hypothesis using the same data that first suggested the hypothesis!
  - 'Texas sharpshooter's fallacy'
  - Subconscious data trawl – the brain is very good at spotting patterns, but there are many possible types of pattern
  - An a priori expectation is required, either pre-specify the hypothesis you will test before looking at the data, or use the data only for exploratory analysis, in order to set a hypothesis to test with new data.

(Log-)likelihood: reminder

- 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:

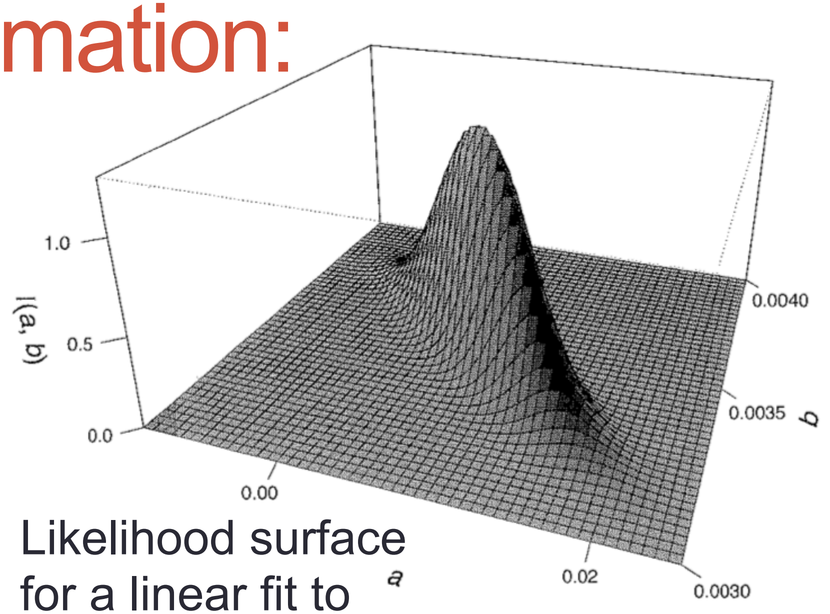
$$L(\boldsymbol{\theta}) = \ln l(\boldsymbol{\theta}) = \ln \left(\prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta}) \right) = \sum_{i=1}^{n} \ln \left[p(y_i|x_i, \boldsymbol{\theta}) \right]$$

Maximum likelihood estimation: reminder

- The partial differentials of the likelihood w.r.t. each parameter are known as the ***scores***:

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

- 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

Confidence intervals on MLEs

- Maximum likelihood estimates (MLEs) of model parameters are themselves statistics, since they are calculated from random data.
- Hence they have a probability distribution and an intrinsic uncertainty.
- We can express this uncertainty in our estimate of a parameter in terms of a *confidence interval* (or in simple-stats language, an error bar).
- It is important to remember that it is our MLE of θ, i.e. $\hat{\theta}$ which has a probability distribution, not θ itself, which has a 'true' value which we are trying to find out. Thus a confidence interval represents our best estimate that, within a certain *confidence level*, our interval contains the true value of θ.
- For one parameter θ and random data $\mathbf{x}$ with log-likelihood function $L(\theta) = \ln(\theta) = \ln\left[p(\mathbf{x}|\theta)\right]$, the variance of the estimator $\hat{\theta}$ can be approximated by:

$$V[\hat{\theta}] \approx -\left(\frac{\partial^2 L}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\hat{\theta}}$$

One-parameter example: Poisson-distributed data

Recall that for: $X \sim \mathrm{Pois}(\lambda) \qquad \mathrm{E}[X] = \lambda \qquad \mathrm{V}[X] = \lambda$

We already found previously that:

$$L(\lambda) = \left(\sum_{i=1}^{n} x_i\right) \ln(\lambda) - n\lambda - \sum_{i=1}^{n} \ln(x_i!)$$

$$\frac{\partial L(\lambda)}{\partial \lambda} = \left(\sum_{i=1}^{n} x_i\right) \frac{1}{\lambda} - n$$

Thus:

$$V[\hat{\lambda}] \approx = \left(-\frac{\sum_{i=1}^{n} x_i}{\lambda^2}\right)^{-1}\Bigg|_{\lambda=\hat{\lambda}} = \frac{\lambda^2}{\sum_{i=1}^{n} x_i}\Bigg|_{\lambda=\hat{\lambda}} = \frac{\hat{\lambda}^2}{\sum_{i=1}^{n} x_i} = \frac{\hat{\lambda}^2}{n\hat{\lambda}} = \frac{\hat{\lambda}}{n}$$

Thus the variance on our estimate of λ (the mean rate) is equal to the variance of *X* divided by *n*, i.e. the standard error:

$$SE_{\overline{x}} = \sqrt{\frac{s_x^2}{n}}$$

‘Graphical method’: relation to $L_{\max}$

- We can expand, as a Taylor series, the likelihood function for the true value of the parameter in terms of our estimate and its deviation from the true value:

$$L(\theta) = L(\hat{\theta}) + \left[\frac{\partial L}{\partial \theta}\right]_{\theta=\hat{\theta}} \left(\theta - \hat{\theta}\right) + \frac{1}{2}\left[\frac{\partial^2 L}{\partial \theta^2}\right]_{\theta=\hat{\theta}} \left(\theta - \hat{\theta}\right)^2 + \cdots$$

- From our definition of the MLE, we have $L(\hat{\theta}) = L_{\max}$ and $\left[\frac{\partial L}{\partial \theta}\right]_{\theta=\hat{\theta}} = 0$ so that:

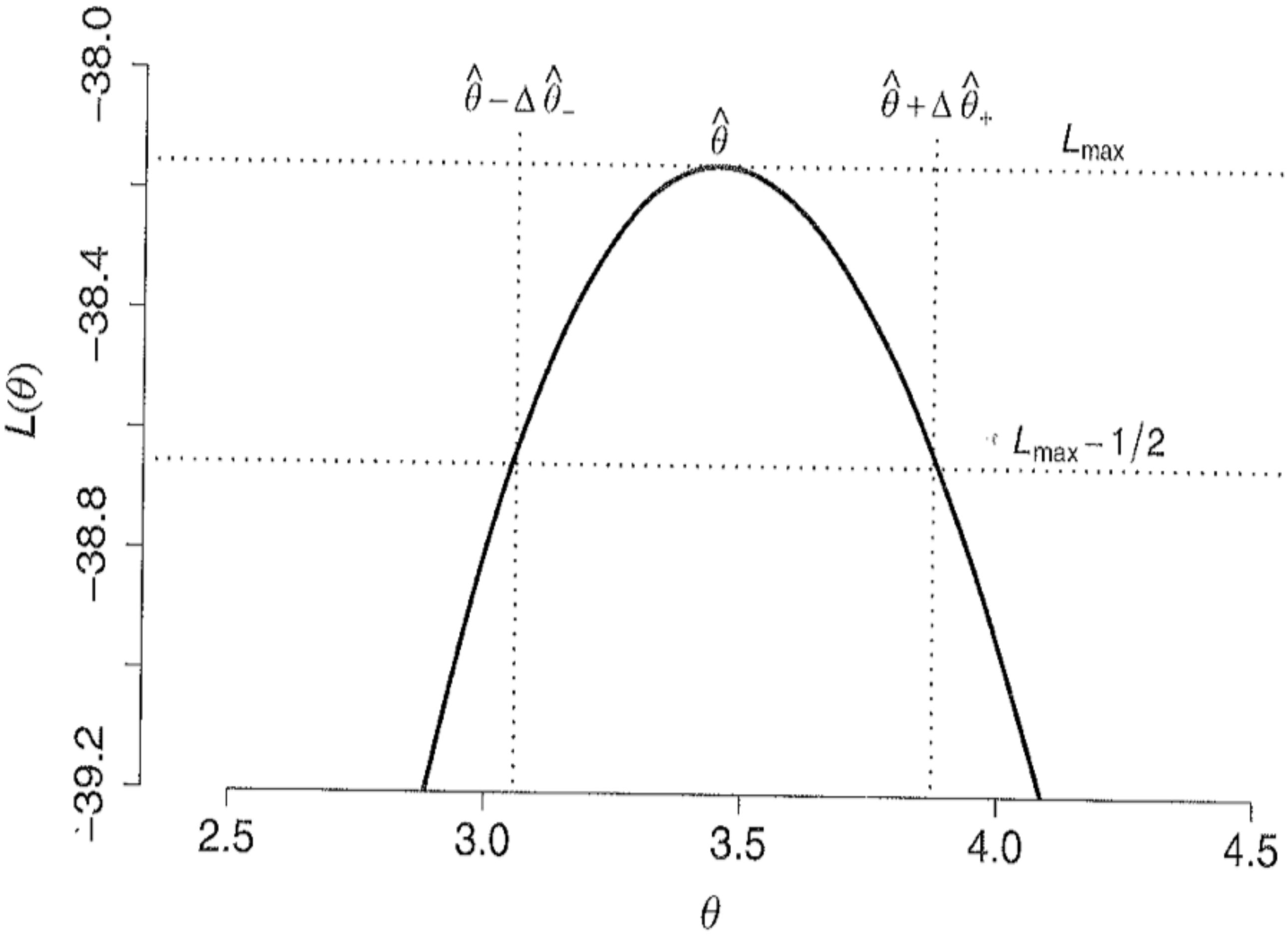
$$L(\theta) \approx L_{\max} + \frac{(\theta - \hat{\theta})^2}{2}\left[\frac{\partial^2 L}{\partial \theta^2}\right]_{\theta=\hat{\theta}}$$

- Since the variance of the MLE is $\mathrm{V}[\hat{\theta}] = \sigma_{\hat{\theta}}^2$ we can write the true value in terms of the expected deviation: $\theta = \hat{\theta} \pm \sigma_{\hat{\theta}}$
- Thus (and using our previous result for the variance of the MLE):

$$L(\hat{\theta} \pm \sigma_{\hat{\theta}}) \approx L_{\max} + \frac{([\hat{\theta} \pm \sigma_{\hat{\theta}}] - \hat{\theta})^2}{2}\left[\frac{\partial^2 L}{\partial \theta^2}\right]_{\theta=\hat{\theta}} = L_{\max} - \frac{\sigma_{\hat{\theta}}^2}{2}\mathrm{V}[\theta]^{-1}$$

$$\Rightarrow L(\hat{\theta} \pm \sigma_{\hat{\theta}}) \approx L_{\max} - \frac{1}{2}$$

‘Graphical method’: relation to $L_{\max}$



$$L(\hat{\theta} \pm \sigma_{\hat{\theta}}) \approx L_{\max} - \frac{1}{2}$$

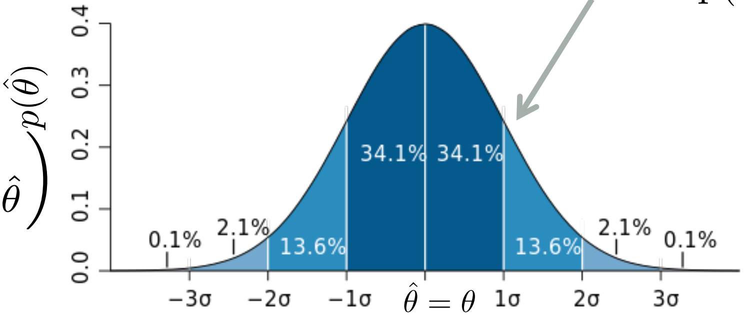
Coverage and confidence intervals

- What does the uncertainty on the MLE actually mean?
- As the sample size increases ($n \rightarrow \infty$), the MLE is distributed with a normal pdf centred on the true value θ
- Since the pdf is:

$$p(z|0,1) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}$$

$$L_{\max} - \Delta L \Leftrightarrow p\left(\hat{\theta} = \theta \pm \sqrt{-2\Delta L}\ \sigma_{\hat{\theta}}\right)$$

$$p(\hat{\theta} = \theta \pm \sigma_{\hat{\theta}}) = \quad p(\hat{\theta} = \theta) \times \exp(-1/2)$$



- The MLE lies within a 1-*σ* deviation of the true value 68.3% of the time. In this case the ***coverage*** is 68.3%.

$\hat{\theta}$

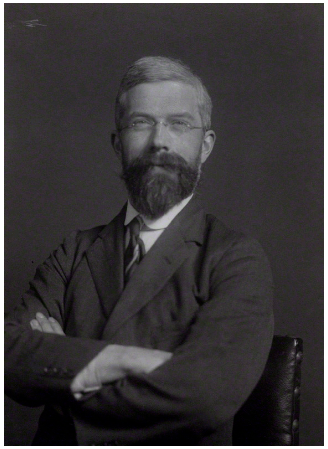
- I.e. the probability that θ lies in the range $[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is 68.3%.
- Similarly, the probability that θ lies in the range $[\hat{\theta} - 2\sigma_{\hat{\theta}}, \hat{\theta} + 2\sigma_{\hat{\theta}}]$ is 95.4%.
- We say that these ranges are respectively the 68.3% and 95.4% ***confidence intervals*** for θ.
- Note that this approximation works in the large sample limit – caution must be applied when the MLE is not likely to be normally distributed.

The many parameter case

- If our model has multiple parameters: $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$

we can generalise our approach. We first compute the $m \times m$ ***Fisher information matrix***:

$$\hat{I}_{ij} = -\frac{\partial^2 L}{\partial\theta_i \partial\theta_j}\bigg|_{\theta=\hat{\theta}}$$



Ronald Fisher (1890-1962)

- The covariance matrix is the inverse of the Fisher information matrix:

$$V_{ij} = \left(\hat{I}^{-1}\right)_{ij}$$

- Note that the element V_{ij} is not the reciprocal of $\hat{I}_{ij}$, it is the element ij of the inverse matrix $\hat{I}^{-1}$.
- If the covariance between two parameter estimates is zero, the parameters are independent of each other.
- The covariance matrix is symmetric [this follows since cov(*x*,*y*) = cov(*y*,*x*)], i.e.:

$$V_{ij} = V_{ji}$$

Covariance: reminder

- We define the covariance of *X* and *Y*:

$$\mathrm{Cov}(X, Y) = \sigma_{xy} = \mathrm{E}[(X - \mu_x)(Y - \mu_y)] = \mathrm{E}[XY] - \mu_x \mu_y$$

- Note that the *Cauchy-Schwarz inequality* states that:

$$|\sigma_{xy}|^2 \leq \sigma_x^2 \sigma_y^2 \qquad \text{or} \qquad |\mathrm{Cov}(X, Y)|^2 \leq \mathrm{V}[X]\mathrm{V}[Y]$$

- The correlation coefficient is:

$$\rho(X, Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

By the Cauchy-Schwarz inequality, this must lie in the range: $-1 \leq \rho(X, Y) \leq 1$

- For independent variables *X* and *Y* we have:

$$\mathrm{E}[XY] = \mu_x \mu_y$$

i.e. covariance is zero

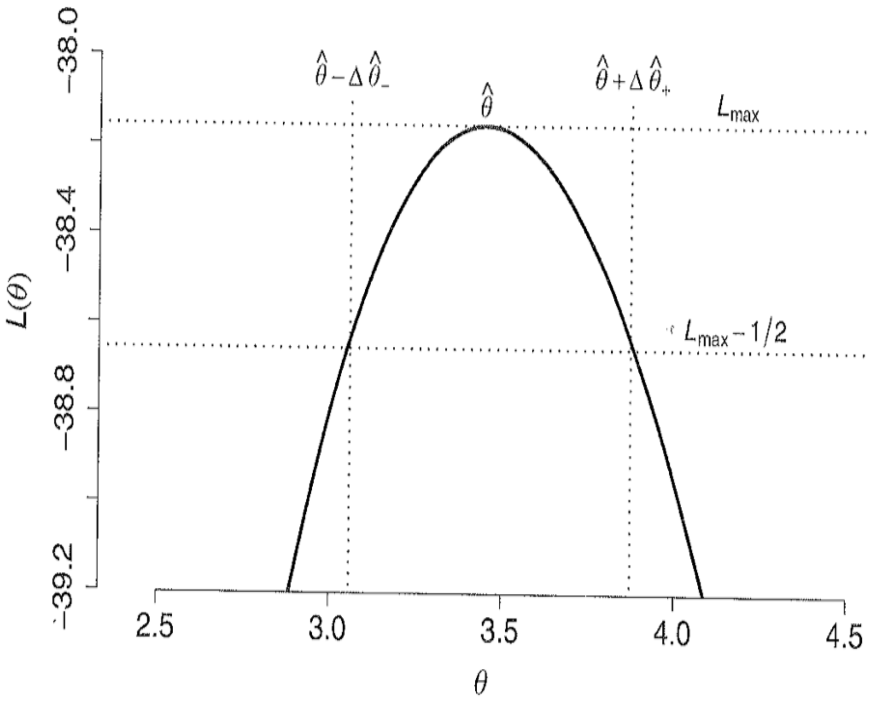
- For multivariate distributions, E.g. for *X*, *Y*, *Z* we can expand the covariance to measure a ***covariance matrix***.

$$\Sigma \quad = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix}$$

The *leading diagonal* contains the variances.
The matrix is symmetric about the leading diagonal

Practical application

- The matrix of 2nd order partial derivatives of a function is called the ***Hessian***
- Many optimisation methods estimate the Hessian and will keep it as output, so you can use it to estimate confidence intervals.
- Or you can use the graphical method: the log-likelihood can be mapped out via brute force (grid of parameter values) or Monte Carlo (e.g. Markov Chain approaches)



Weighted least squares: reminder

- 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, $X^2(\boldsymbol{\theta})$

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

Confidence intervals in weighted least-squares (chi-squared) fitting

- For least-squares fitting we have:

$$X^2(\theta) = -2L(\theta) + const$$

- So that the confidence intervals in the single parameter case are:

$$X^2(\hat{\theta} \pm \sigma_{\hat{\theta}}) = X^2_{\mathrm{min}} + 1$$

$$X^2(\hat{\theta} \pm z\sigma_{\hat{\theta}}) = X^2_{\mathrm{min}} + \sqrt{z}$$

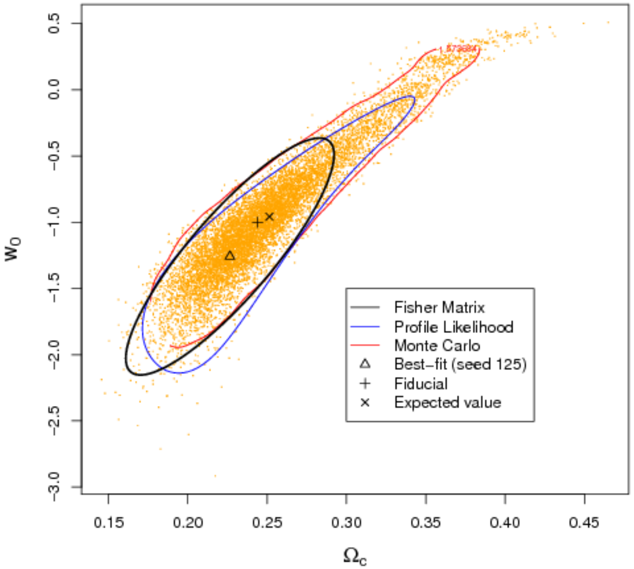
- Similarly, for the multi-parameter case we have the Fisher information matrix:

$$\hat{I}_{ij} = \frac{1}{2} \frac{\partial^2 X^2(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\bigg|_{\theta=\hat{\theta}}$$

which can be inverted to obtain the covariance matrix.

Finding confidence intervals via Monte Carlo simulation or bootstrapping

- In some cases the log-likelihood function may be difficult to obtain (e.g. if the pdf of our data is unknown or the log-likelihood is not possible to derive from the data), or the distribution of MLEs may be significantly non-normal.
- Monte Carlo simulation can be used to derive the confidence intervals, by simulating data and repeating MLE estimation many times, to map out the density of estimates and hence the confidence intervals.
- If the pdf of the data is known one can simulate individual data points assuming the best-fitting model parameters. Bootstrapping can also be used in this way to generate 'fake' data.



(Penna-Lima et al. 2014)

- If the pdf of our data is unknown one can try to simulate the population and measurement process itself given the model parameters.
- More 'Bayesian' is to randomly draw model parameters from a plausible ***prior*** distribution then simulate.