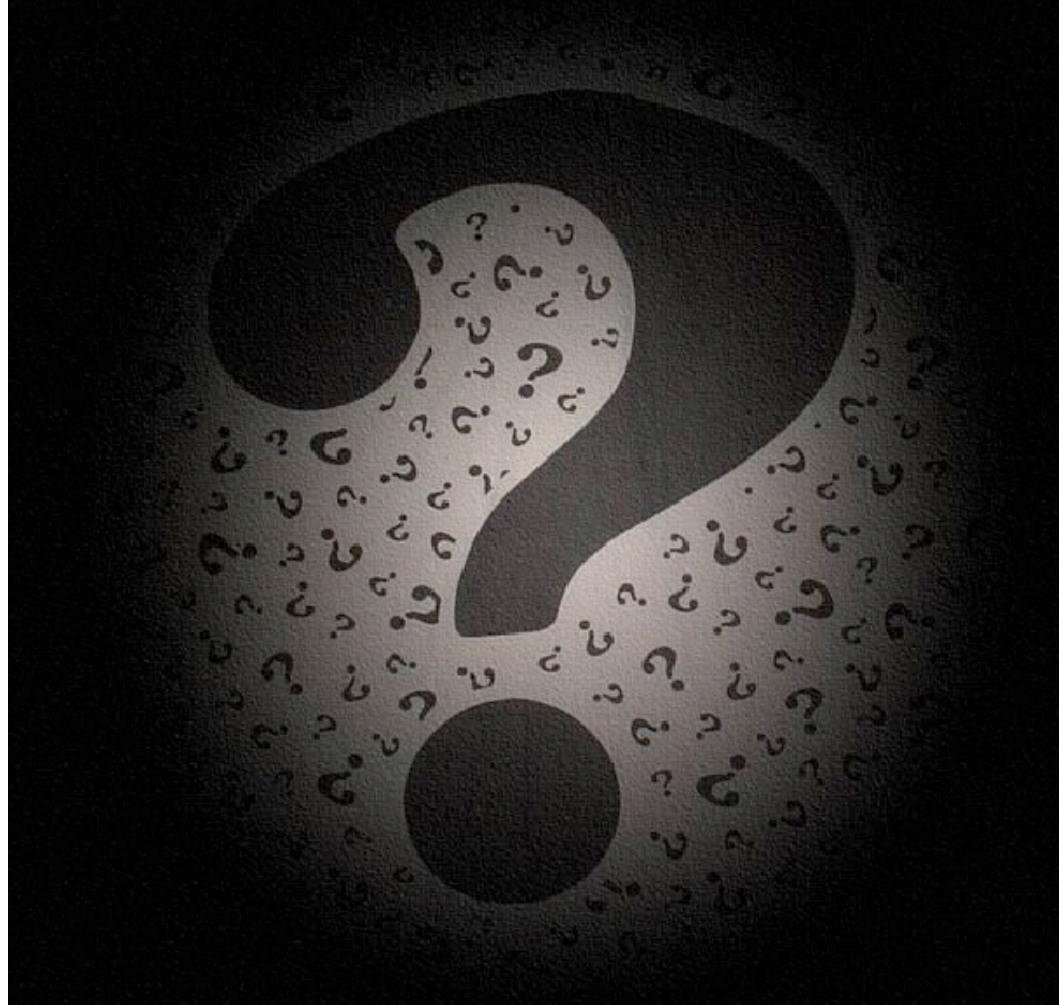


Parameter estimation and forecasting

Cristiano Porciani
AIfA, Uni-Bonn

Questions?



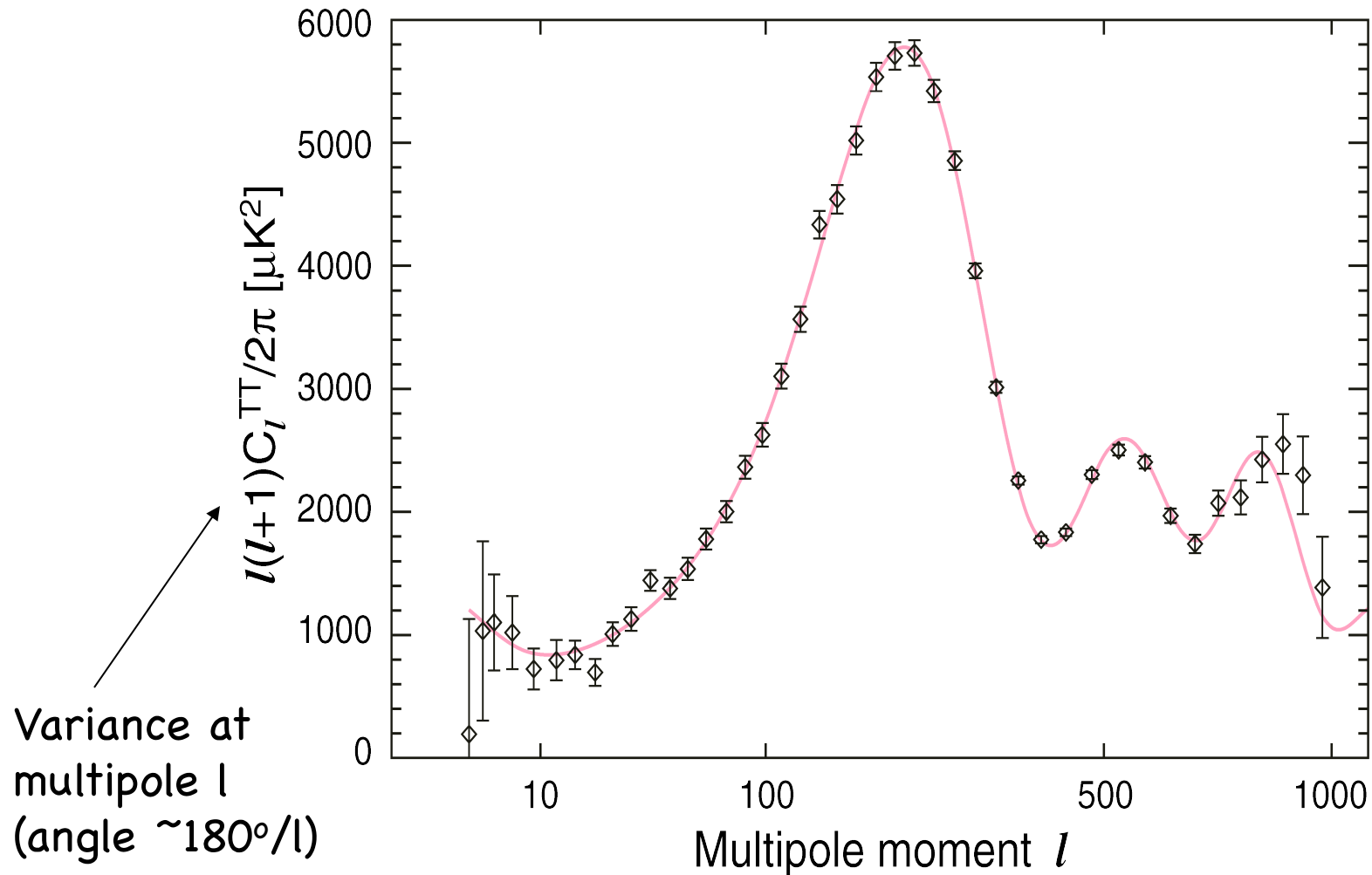
Cosmological parameters

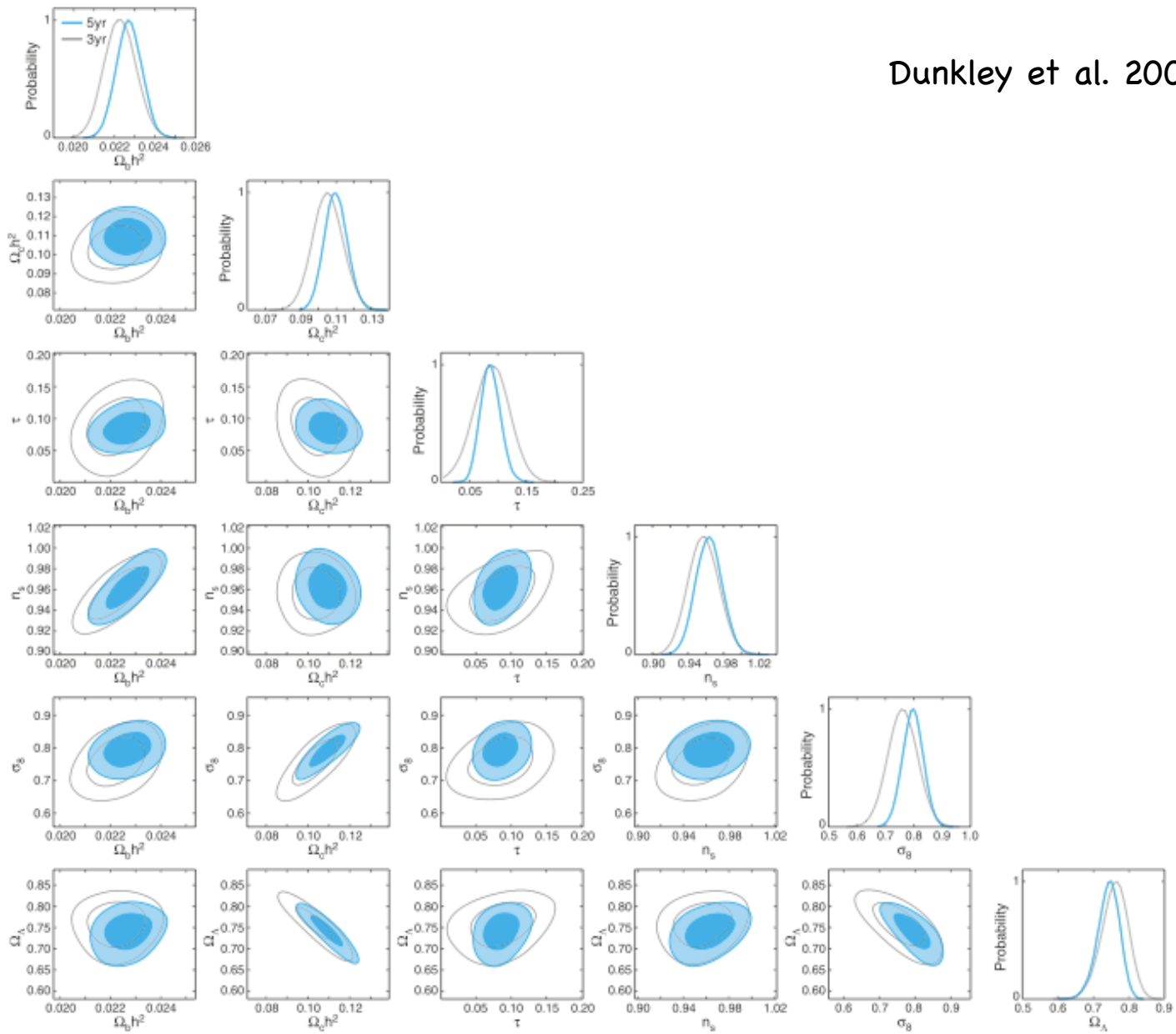
- A branch of modern cosmological research focuses on measuring cosmological parameters from observed data (e.g. the Hubble constant, the cosmic density of matter, etc.).
- In this class we will review the main techniques used for **model fitting** (i.e. extracting information on cosmological parameters from existing observational data) and **forecasting** (i.e. predicting the uncertainty on the parameters when future experiments will become available). The latter is a crucial ingredient for optimizing experimental design.

Key problems

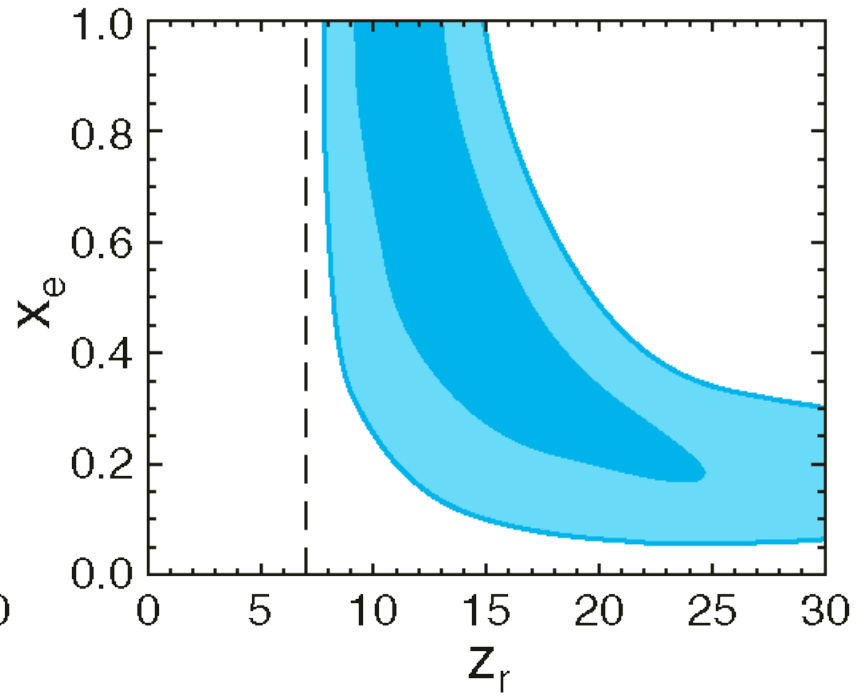
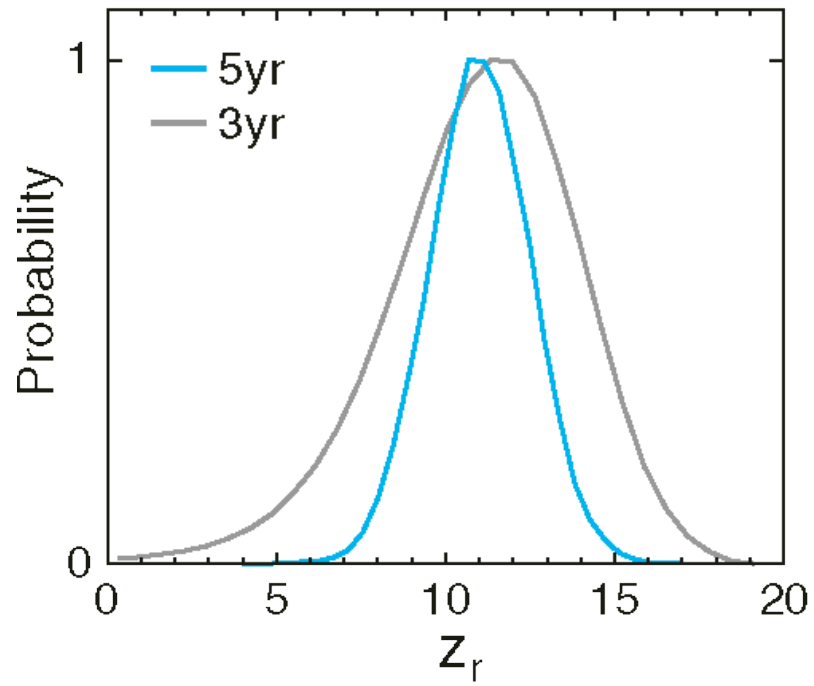
- How do you fit a model to data?
- How do you incorporate prior knowledge?
- How do you merge multiple sources of information?
- How do you treat uncertainties in model parameters?

Example: power spectrum of CMB temperature fluctuations

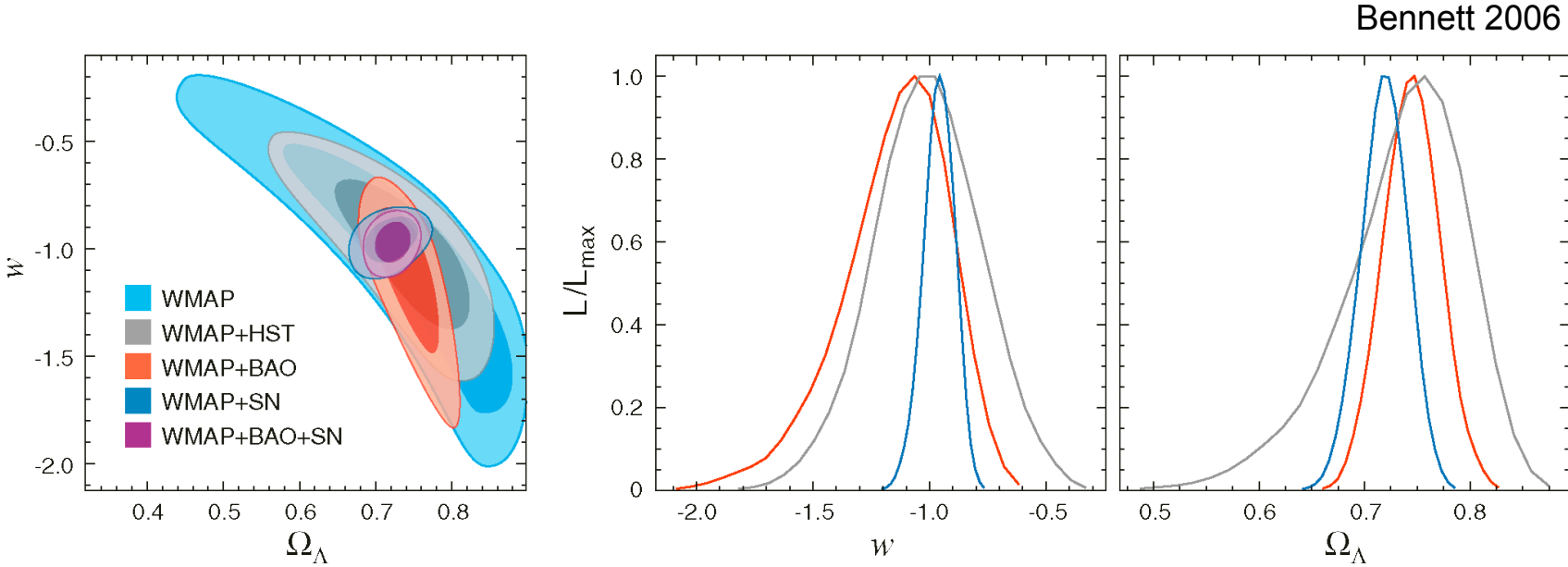




Dunkley et al. 2009



The current state of the art



What is the meaning of these plots?

- What's the difference between the 1D and the 2D plots?
- What is a confidence interval?
- What is a credibility interval?
- What does marginalisation mean?
- What's the difference between the frequentist and the Bayesian interpretation of statistics?

Descriptive statistics

Input

- a set of data

Output

- The sample mean or median
- The variance
- A histogram of one or more variables
- A scatterplot between different variables
- Correlation coefficient between different variables

Statistical inference

Input:

- a set of data
- a statistical model of the random process that generates the data
(a set of assumptions, e.g. Gaussian or Poisson distributed with some free parameters)

Output:

- A point estimate (e.g. the value that best approximates a model parameter)
- A set estimate (e.g. a confidence or credibility interval for a model parameter)
- Rejection of a hypothesis to some confidence level
- Classification of data points into groups

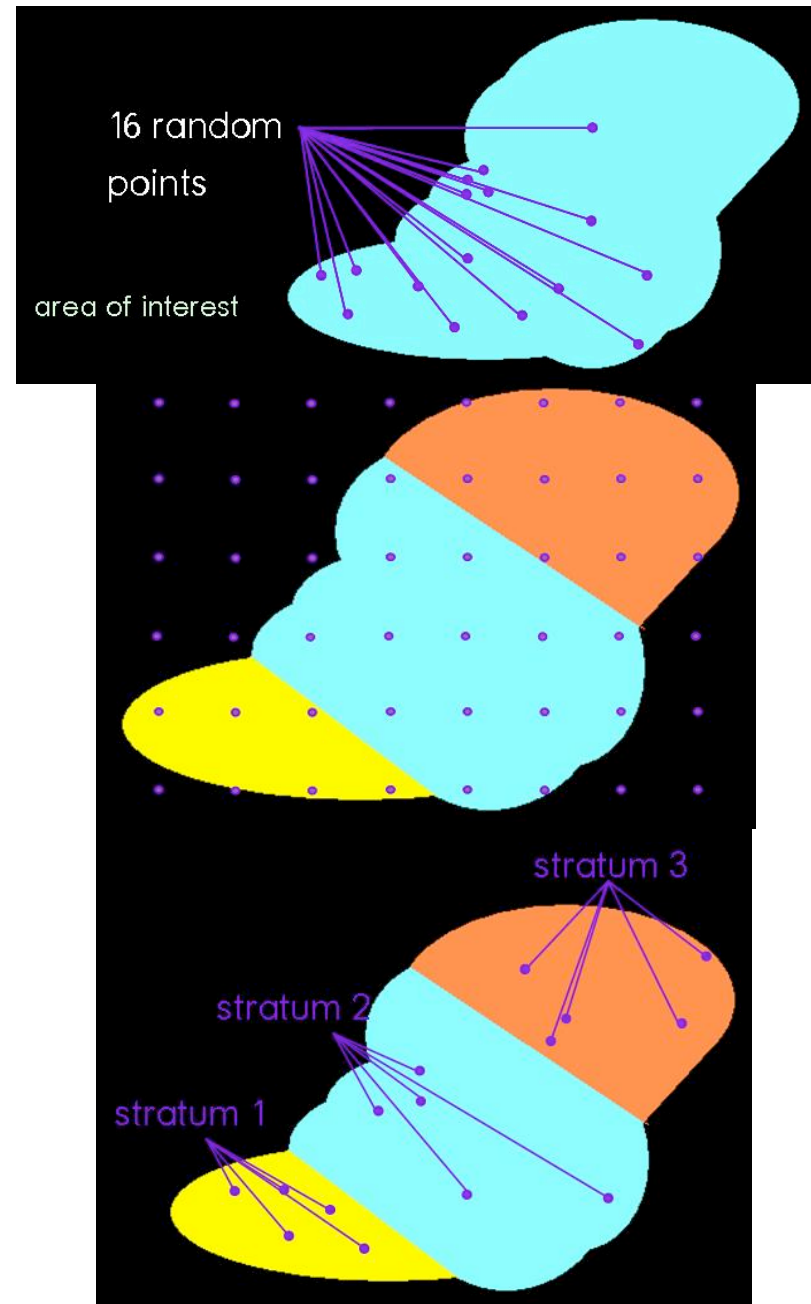
Population and sample



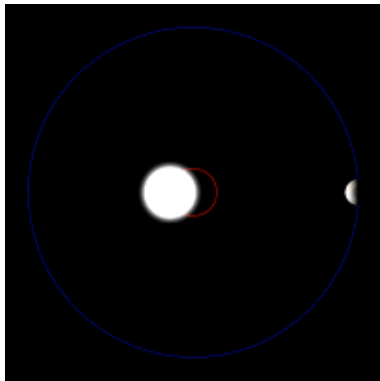
- A population is any entire collection of objects (people, animals, galaxies) from which we may collect data. It is this entire group we are interested in, which we wish to describe or draw conclusions about.
- A sample is a group of units selected from the population for study because the population is too large to study in its entirety. For each population there are many possible samples.

Sampling

- Selection of observations intended to yield knowledge about a population of concern
- Social sciences: census, simple random sampling, systematic sampling, stratified sampling, etc.
- Sample-selection biases (also called selection effects) arise if the sample is not representative of the population
- In astronomy often observational selection effects must be modeled a posteriori because sample-selection is determined by instrumental limits

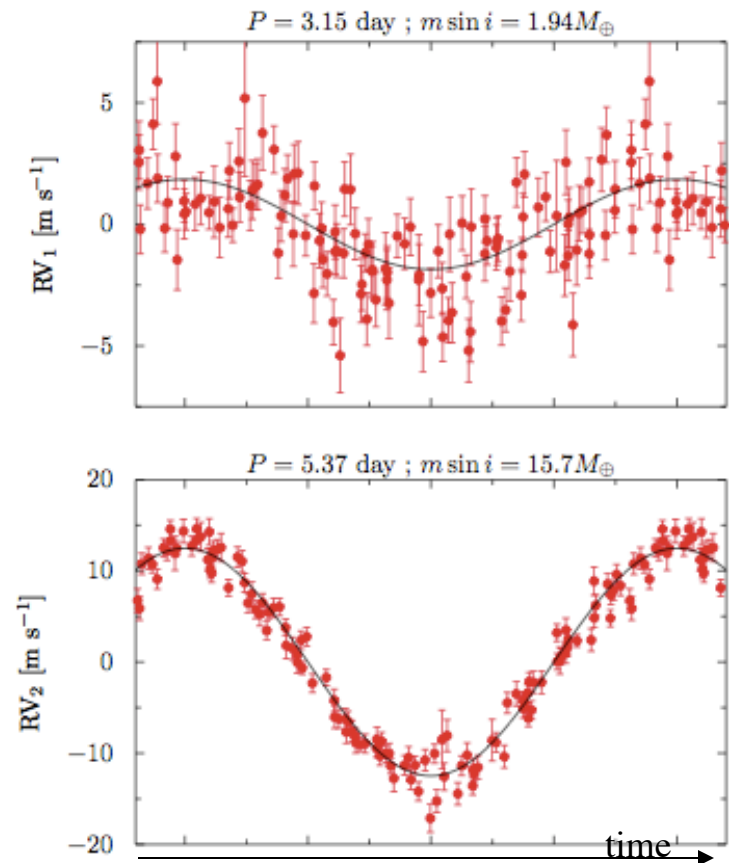


Example: extra-solar planets from Doppler surveys



$$v_{obs} = \frac{m_p}{M_s} \sqrt{\frac{GM_s}{r}} \sin(i)$$

The method is best at detecting “hot Jupiters”, very massive planets close to the parent star. Current experiments (HARPS) can measure radial velocities of approximately 1 m/s corresponding to 4 Earth masses at 0.1 AU and 11 Earth masses at 1 AU.



Mayor et al. 2009

Understanding the selection effects
is often the crucial element of a
paper in astronomy!

What is estimation?

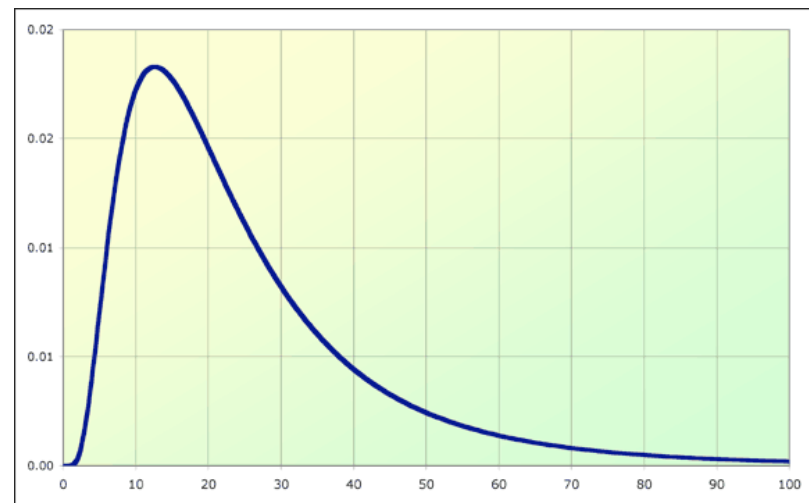
- In statistics, **estimation** (or inference) refers to the process by which one makes inferences (e.g. draws conclusions) about a population, based on information obtained from a sample.
- A **statistic** is any measurable quantity calculated from a sample of data (e.g. the average). This is a stochastic variable as, for a given population, it will in general vary from sample to sample.
- An **estimator** is any quantity calculated from the sample data which is used to give information about an unknown quantity in the population (the estimand).
- An **estimate** is the particular value of an estimator that is obtained by a particular sample of data and used to indicate the value of a parameter.

A simple example

- Population: people in this room
- Sample I: people sitting in the middle row
Sample II: people whose names start with the letter M
- Statistic: average height
- I can use this statistic as an estimator for the average height of the population obtaining different results from the two samples

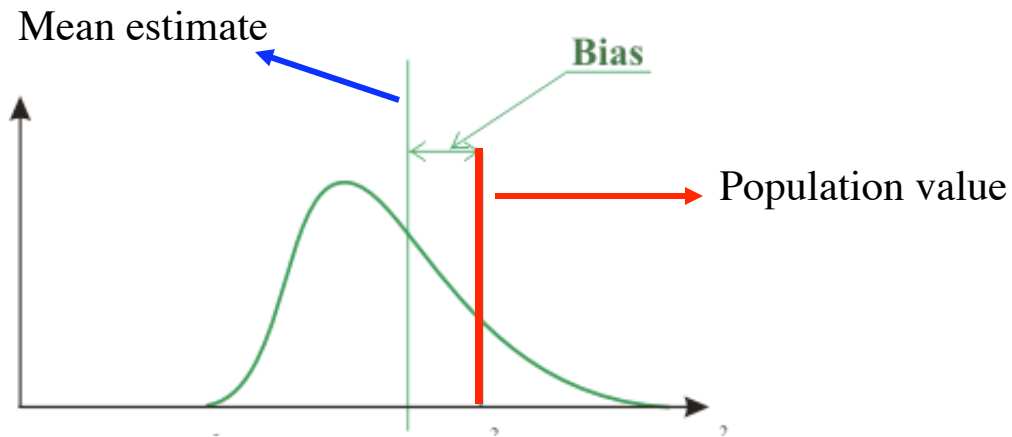
PDF of an estimator

- Ideally one can consider all possible samples corresponding to a given sampling strategy and build a probability density function (PDF) for the different estimates
- We will use the characteristics of this PDF to evaluate the quality of an estimator



Value of estimated statistic

Bias of an estimator



- The bias of an estimator is the difference between the expectation value over its PDF (i.e. its mean value) and the population value

$$b(\hat{\theta}) = E(\hat{\theta}) - \theta_0 = \langle \hat{\theta} \rangle - \theta_0 = \langle \hat{\theta} - \theta_0 \rangle$$

- An estimator is called unbiased if $b=0$ while it is called biased otherwise

Examples

- The sample mean is an unbiased estimator of the population mean

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i, \quad E[\bar{x}] = \frac{1}{N} \sum_{i=1}^N E[x_i] = \frac{N}{N} \mu = \mu$$

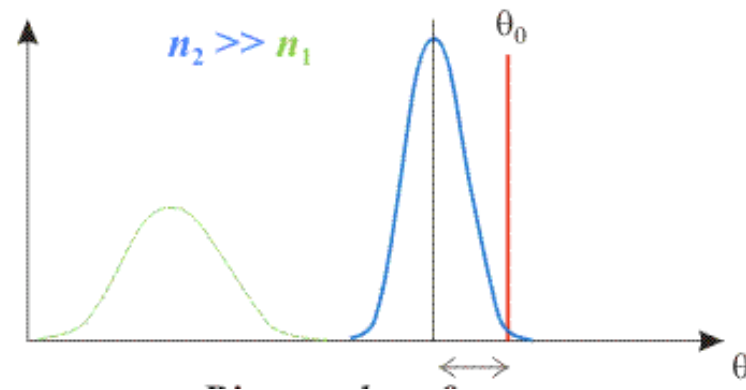
- Exercise: Is the sample variance an unbiased estimator of the population variance?

$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2, \quad E[s^2] = ???$$

Examples

- Note that functional invariance does not hold.
- If you have an unbiased estimator S^2 for the population variance σ^2 and you take its square root, this will NOT be an unbiased estimator for the population rms value σ !
- This applies to any non-linear transformation including division.
- Therefore avoid to compute ratios of estimates as much as you can.

Consistent estimators



- We can build a sequence of estimators by progressively increasing the sample size
- If the probability that the estimates deviate from the population value by more than $\varepsilon \ll 1$ tends to zero as the sample size tends to infinity, we say that the estimator is consistent

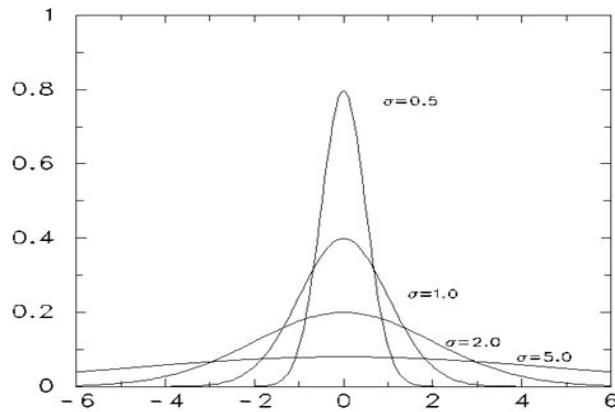
Example

- The sample mean is a consistent estimator of the population mean

$$\begin{aligned} \text{Var}[\bar{x}] &= E[(\bar{x} - \mu)^2] = E\left[\left(\frac{1}{N} \sum_{i=1}^N x_i - \mu\right)^2\right] = \frac{1}{N^2} E\left[\left(\sum_{i=1}^N x_i\right)^2\right] - 2\frac{\mu}{N} N\mu + \mu^2 = \\ &= \frac{1}{N^2} N(\mu^2 + \sigma^2) + \frac{N(N-1)}{N^2} \mu^2 - \mu^2 = \frac{\sigma^2}{N} \end{aligned}$$

$$\text{Var}[\bar{x}] \rightarrow 0 \quad \text{when} \quad N \rightarrow \infty$$

Relative efficiency



Suppose there are 2 or more unbiased estimators of the same quantity, which one should we use? (e.g. should we use the sample mean or sample median to estimate the centre of a Gaussian distribution?)

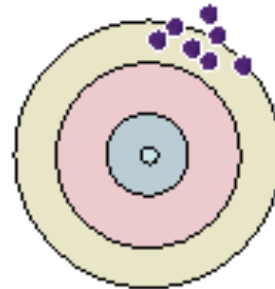
- Intuition suggests that we should use the estimator that is closer (in a probabilistic sense) to the population value. One way to do this is to choose the estimator with the lowest variance.
- We can thus define a relative efficiency as: $E[(\hat{\vartheta}_1 - \theta_0)^2] / E[(\hat{\vartheta}_2 - \theta_0)^2]$
- If there is an unbiased estimator that has lower variance than any other for all possible population values, this is called the minimum-variance unbiased estimator (MVUE)

Efficient estimators

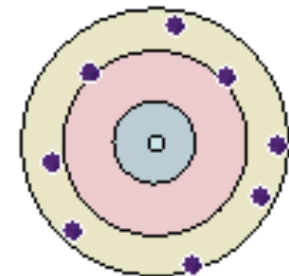
- A theorem known as the Cramer-Rao bound (see Alan Heaven's lectures) proves that the variance of an unbiased estimator must be larger or equal to a specific value which only depends on the sampling strategy (it corresponds to the reciprocal of the Fisher information of the sample)
- We can thus define an absolute efficiency of an estimator as the ratio between the minimum variance and the actual variance
- An unbiased estimator is called efficient if its variance coincides with the minimum variance for all values of the population parameter θ_0

Accuracy vs precision

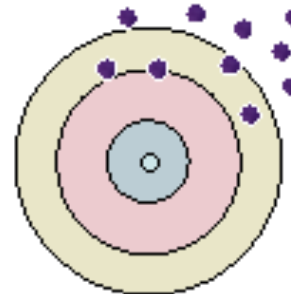
- The bias and the variance of an estimator are very different concepts (see the bullseye analogy on the right)
- Bias quantifies accuracy
- Variance quantifies precision



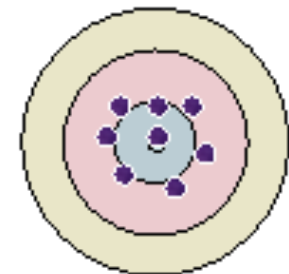
Bias is large
variation is small



Bias is small
variation is large



Bias is large
variation is large

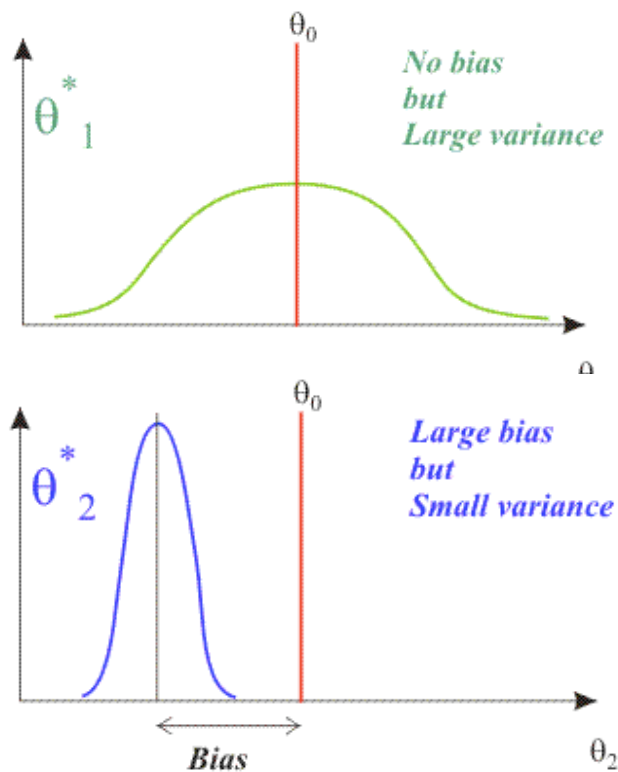


Bias is small
variation is small

Desirable properties of an estimator

- ✓ Consistency
 - ✓ Unbiasedness
 - ✓ Efficiency
- However, unbiased and/or efficient estimators do not always exist
 - Practitioners are not particularly keen on unbiasedness. So they often tend to favor estimators such that the mean square error, $MSE = E[(\hat{\theta} - \theta_0)^2]$, is as low as possible independently of the bias.

Minimum mean-square error



- Note that,

$$\begin{aligned}MSE &= E[(\hat{\theta} - \theta_0)^2] = E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta_0)^2] = \\ &= E[(\hat{\theta} - E[\hat{\theta}])^2] + (E[\hat{\theta}] - \theta_0)^2 = \sigma^2(\hat{\theta}) + b^2(\hat{\theta})\end{aligned}$$

- A biased estimator with small variance can then be preferred to an unbiased one with large variance
- However, identifying the minimum mean-square error estimator from first principles is often not an easy task. Also the solution might not be unique (the bias-variance tradeoff)

Point vs interval estimates

- A **point estimate** of a population parameter is a single value of a statistic (e.g. the average height). This in general changes with the selected sample.
- In order to quantify the uncertainty of the sampling method it is convenient to use an **interval estimate** defined by two numbers between which a population parameter is said to lie
- An interval estimate is generally associated with a confidence level. Suppose we collected many different samples (with the same sampling strategy) and computed confidence intervals for each of them. Some of the confidence intervals would include the population parameter, others would not. A 95% confidence level means that 95% of the intervals contain the population parameter.

This is all theory but how do we build an estimator in practice?

Let's consider a simple (but common) case.

Suppose we perform an experiment where we measure a real-valued variable X .

The experiment is repeated n times to generate a random sample X_1, \dots, X_n of independent, identically distributed variables (iid).

We also assume that the shape of the population PDF of X is known (Gaussian, Poisson, binomial, etc.) but has k unknown parameters $\theta_1, \dots, \theta_k$ with $k < n$.

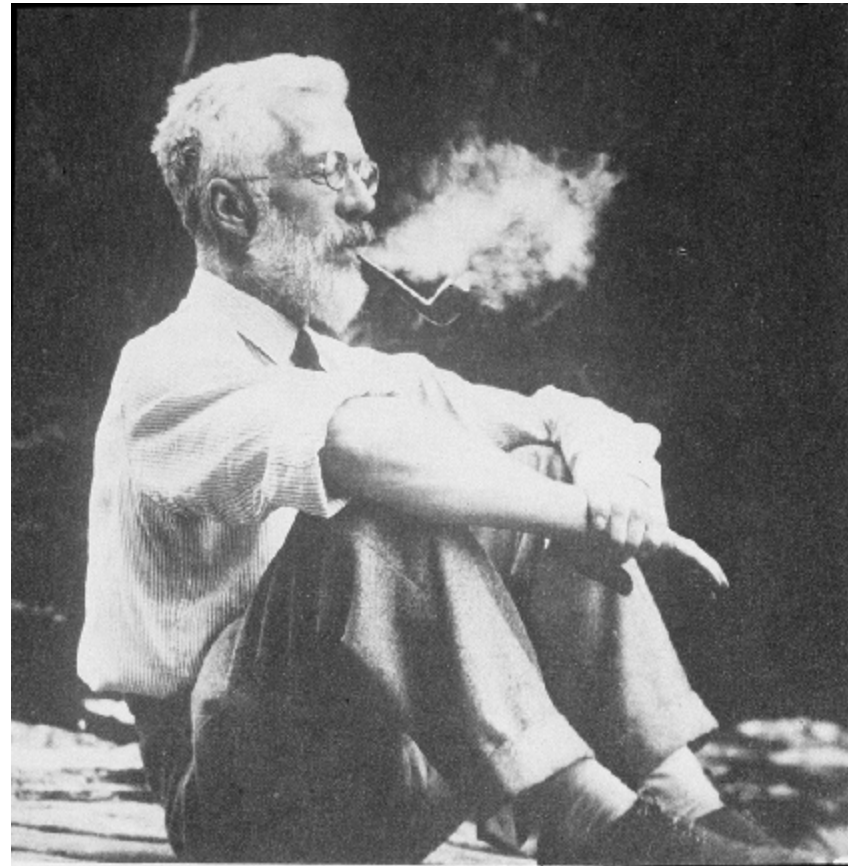
The old way: method of moments

- The method of moments is a technique for constructing estimators of the parameters of the population PDF
- It consists of equating sample moments (mean, variance, skewness, etc.) with population moments
- This gives a number of equations that might (or might not) admit an acceptable solution
- There is a much better way that we are going to describe now

I occasionally meet geneticists who ask me whether it is true that the great geneticist R.A. Fisher was also an important statistician (L. J. Savage)

R.A. Fisher (1890-1962)

"Fisher was to statistics what Newton was to Physics" (R. Kass)



"Even scientists need their heroes, and R.A. Fisher was the hero of 20th century statistics" (B. Efron)

The greatest biologist since Darwin (J.R. Dawkins)

Fisher's concept of likelihood

- “Two radically distinct concepts have been confused under the name of ‘probability’ and only by sharply distinguishing between these can we state accurately what information a sample does give us respecting the population from which it was drawn.” (Fisher 1921)
- “We may discuss the probability of occurrence of quantities which can be observed...in relation to any hypotheses which may be suggested to explain these observations. We can know nothing of the probability of the hypotheses...We may ascertain the likelihood of the hypotheses...by calculation from observations:...to speak of the likelihood...of an observable quantity has no meaning.” (Fisher 1921)
- “The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed.” (Fisher 1922)

Probability of the data versus likelihood of the parameters

- Suppose you are counting how many cars pass in front of your window on Sundays between 9:00 and 9:02 am. Counting experiments are generally well described by the Poisson distribution. Therefore, if the mean counts are λ , the probability of counting n cars follows the distribution:

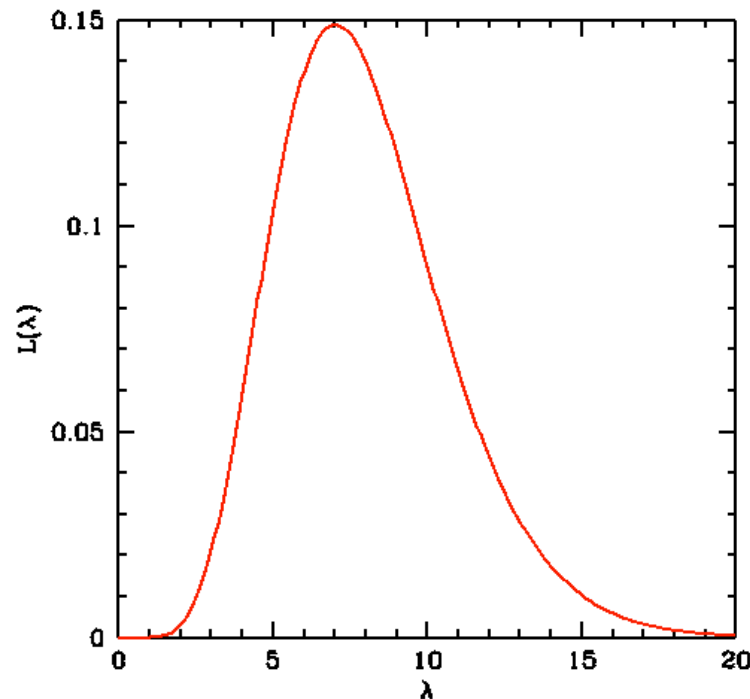
$$P(n | \lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

- This means that if you repeat the experiment many times, you will measure different values of n following the frequency $P(n)$. Note that the sum over all possible n is unity.
- Now suppose that you actually perform the experiment once and you count 7. Then, the likelihood for the model parameter λ GIVEN the data is:

$$L(\lambda) = P(7 | \lambda) = \frac{\lambda^7 e^{-\lambda}}{5040}$$

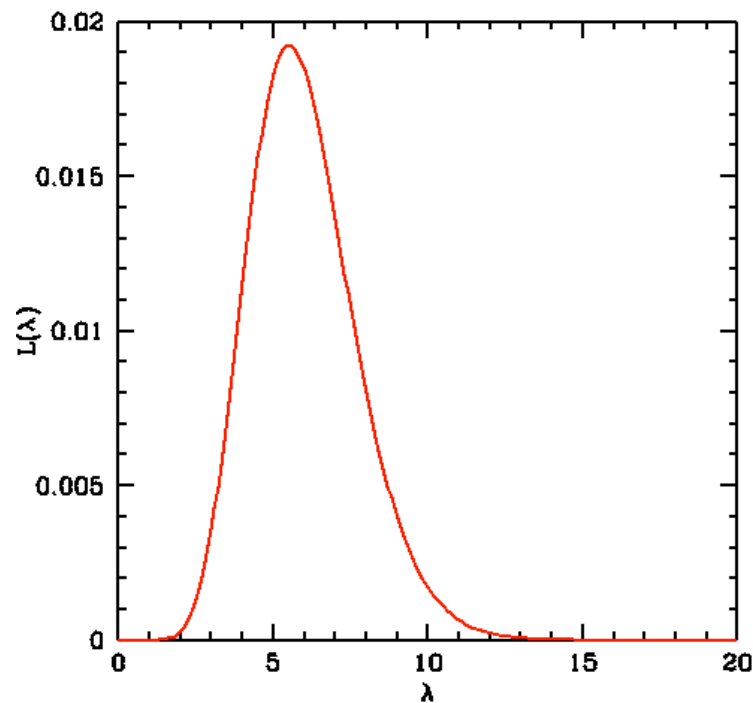
The likelihood function

- This is a function of λ only but it is NOT a probability distribution for λ ! It simply says how likely it is that our measured value of $n=7$ is obtained by sampling a Poisson distribution of mean λ . It says something about the model parameter GIVEN the observed data.



The likelihood function

- Let us suppose that after some time you repeat the experiment and count 4 cars. Since the two experiments are independent, you can multiply the likelihoods and obtain the curve below. Note that now the most likely value is $\lambda = 5.5$ and the likelihood function is narrower than before, meaning that we know more about λ .



Likelihood for Gaussian errors

- Often statistical measurement errors can be described by Gaussian distributions. If the errors σ_i of different measurements d_i are independent:

$$L(\theta) = P(d | \theta) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(d_i - m_i(\theta))^2}{2\sigma_i^2}\right]$$

$$-\ln L(\theta) = \sum_{i=1}^N \frac{(d_i - m_i(\theta))^2}{2\sigma_i^2} + \text{const.} = \frac{\chi^2(\theta, d)}{2} + \text{const.}$$

- Maximizing the likelihood corresponds to finding the values of the parameters $\theta = \{\theta_1, \dots, \theta_n\}$ which minimize the χ^2 function (weighted least squares method).

The general Gaussian case

- In general, errors are correlated and

$$-\ln L(\theta) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N [d_i - m_i(\theta)] C_{ij}^{-1} [d_j - m_j(\theta)] + \text{const.} = \frac{\chi^2(\theta, d)}{2} + \text{const.}$$

where $C_{ij} = \langle \varepsilon_i \varepsilon_j \rangle$ is the covariance matrix of the errors.

- For uncorrelated errors the covariance matrix is diagonal and one reduces to the previous case.
- Note that the covariance matrix could also derive from a model and then depend on the model parameters. We will encounter some of these cases in the rest of the course.

The Likelihood function: a summary

- In simple words, the likelihood of a model given a dataset is proportional to the probability of the data given the model
- The likelihood function supplies an order of preference or plausibility of the values of the free parameters θ_i by how probable they make the observed dataset
- The likelihood ratio between two models can then be used to prefer one to the other
- Another convenient feature of the likelihood function is that it is functionally invariant. This means that any quantitative statement about the θ_i implies a corresponding statements about any one to one function of the θ_i by direct algebraic substitution

Maximum Likelihood

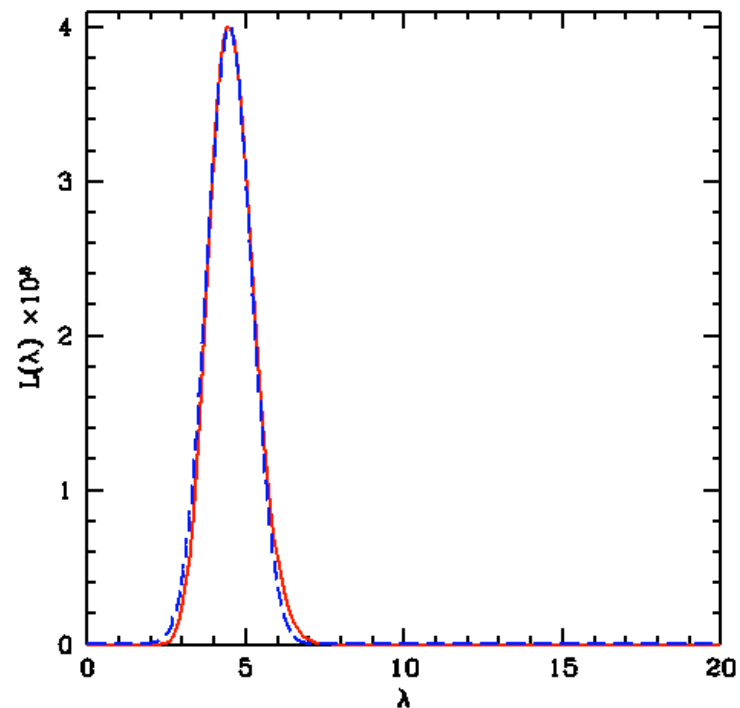
- The likelihood function is a statistic (i.e. a function of the data) which gives the probability of obtaining that particular set of data, given the chosen parameters $\theta_1, \dots, \theta_k$ of the model. It should be understood as a function of the unknown model parameters (but it is NOT a probability distribution for them)
- The values of these parameters that maximize the sample likelihood are known as the Maximum Likelihood Estimates or MLE's.
- Assuming that the likelihood function is differentiable, estimation is done by solving

$$\frac{\partial L(\theta_1, \dots, \theta_k)}{\partial \theta_i} = 0 \quad \text{or} \quad \frac{\partial \ln L(\theta_1, \dots, \theta_k)}{\partial \theta_i} = 0$$

- On the other hand, the maximum value may not exist at all.

Back to counting cars

- After 9 experiments we collected the following data: 7, 4, 2, 6, 4, 5, 3, 4, 5. The new likelihood function is plotted below, together with a Gaussian function (dashed line) which matches the position and the curvature of the likelihood peak ($\lambda = 4.44$). Note that the 2 curves are very similar (especially close to the peak), and this is not by chance.



Score and information matrix

- The first derivative of the log-likelihood function with respect to the different parameters is called the Fisher score function:

$$S_i = \frac{\partial \ln L(\theta)}{\partial \theta_i}$$

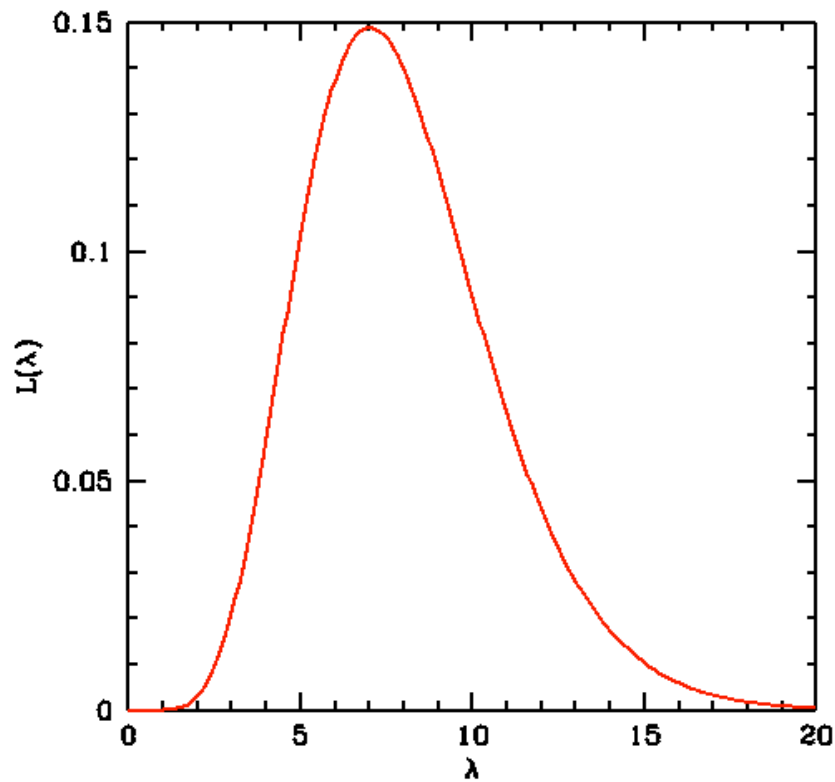
- The Fisher score vanishes at the MLE.
- The negative of the Hessian matrix of the log-likelihood function with respect to the different parameters is called the observed information matrix:

$$O_{ij} = -\frac{\partial^2 \ln L(\theta)}{\partial \theta_i \partial \theta_j}$$

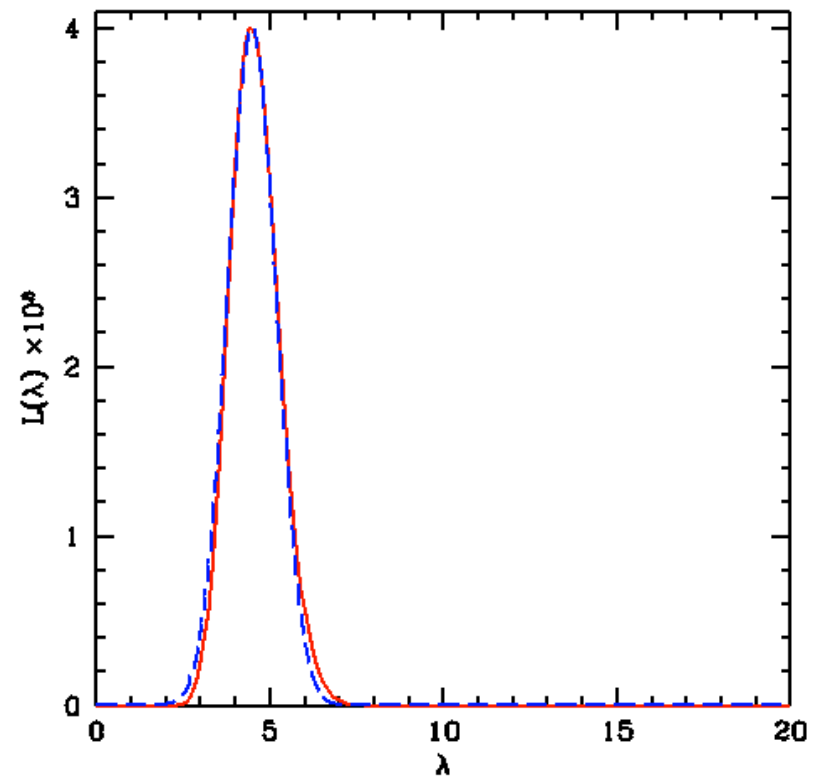
- The observed information matrix is definite positive at the MLE. Its elements tell us how broad is the likelihood function close to its peak and thus with what accuracy we determined the model parameters.

Example

1 datapoint
Low information
Large uncertainty in λ



9 datapoints
High information
Small uncertainty in λ



Fisher information matrix

- If we took different data, then the likelihood function for the parameters would have been a bit different and so its score function and the observed information matrix.
- Fisher introduced the concept of information matrix by taking the ideal ensemble average (over all possible datasets of a given size) of the observed information matrix (evaluated at the true value of the parameters).

$$F_{ij} = - \left\langle \frac{\partial^2 \ln L(\theta)}{\partial \theta_i \partial \theta_j} \right\rangle$$

- Under mild regularity conditions, it can be shown that the Fisher information matrix also corresponds to

$$F_{ij} = \left\langle \frac{\partial \ln L(\theta)}{\partial \theta_i} \frac{\partial \ln L(\theta)}{\partial \theta_j} \right\rangle$$

i.e. to the covariance matrix of the scores at the MLE's.

Cramér-Rao bound

- The Cramér-Rao bound states that, for ANY unbiased estimator of a model parameter θ_i , the measurement error (keeping the other parameters constant) satisfies


$$\Delta\theta_i \geq \frac{1}{\sqrt{F_{ii}}}$$

- For marginal errors that also account for the variability of the other parameters (see slide 35 for a precise definition), instead, it is the inverse of the Fisher information matrix that matters and

$$\Delta\theta_i \geq \sqrt{F_{ii}^{-1}}$$

Fisher matrix with Gaussian errors

- For data with Gaussian errors, the Fisher matrix assumes the form (the notation is the same as in slide 20)

$$F_{ij} = \frac{1}{2} \text{Tr} \left[C^{-1} C_{,i} C^{-1} C_{,j} + C^{-1} M_{ij} \right]$$


where

Information from the noise

Information from the signal

$$M_{ij} = m_{,i} m_{,j}^T + m_{,j} m_{,i}^T$$

(note that commas indicate derivatives with respect to the parameters while data indices are understood)

Properties of MLE's

As the sample size increases to infinity (under weak regularity conditions):

- MLE's become asymptotically efficient and asymptotically unbiased
- MLE's asymptotically follow a normal distribution with covariance matrix (of the parameters) equal to the inverse of the Fisher's information matrix (that is determined by the covariance matrix of the data).

However, for small samples,

- MLE's can be heavily biased and the large-sample optimality does not apply

Maximizing likelihood functions

- For models with a few parameters, it is possible to evaluate the likelihood function on a finely spaced grid and search for its minimum (or use a numerical minimisation algorithm).
- For a number of parameters $\gg 2$ it is NOT feasible to have a grid (e.g. 10 point in each parameter direction, 12 parameters = 10^{12} likelihood evaluations!!!)
- Special statistical and numerical methods needs to be used to perform model fitting.
- Note that typical cosmological problems consider models with a number of parameters ranging between 6 and 20.

Forecasting

- Forecasting is the process of estimating the performance of future experiments for which data are not yet available
- It is a key step for the optimization of experimental design (e.g. how large must be my survey if I want to determine a particular parameter to 1% accuracy?)
- The basic formalism has been developed by Fisher in 1935

Figure of merit

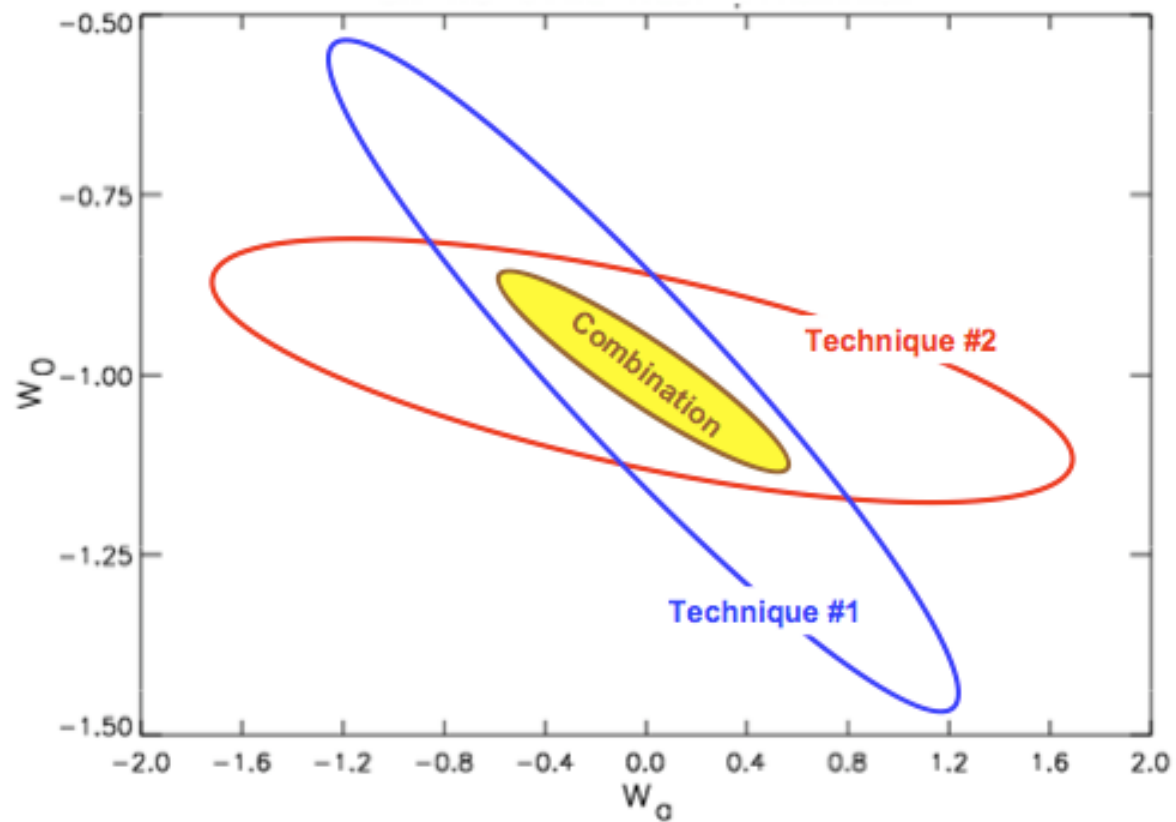



Figure of merit = $1 / (\text{area of the ellipse})$

iCOSMO.org

[Initiative](#) [Tools](#) [Resources](#) [Help](#) [Contact Us](#) [FAQs](#)

INITIATIVE FOR COSMOLOGY 

Welcome!

This site is designed to make cosmology calculations easy and pain-free. Here, you will find a host of tools and resources for performing calculations, ranging from distance calculations to cosmological error predictions for future surveys.

The site also contains a set of tutorials and links that are useful whether you are a newbie to cosmology or a seasoned professional. These resources have been made available in an easy-to-access format and will be continually updated and expanded.

[COSMOLOGY TOOLS:](#)
You can perform a calculation either by using your web browser or by [downloading the source code](#). To get started you can either go to [tools](#), and you will be guided through each step. Alternatively, you can use the QuickStart Calculator to the right.

[COSMOLOGY RESOURCES:](#)
Here you will find general cosmology support materials, such as tutorials and links to external sites. To find the material you need go to [resources](#) or use the QuickStart Tutorial to the right. If you wish to create your own interactive web pages you can use the templates available [here](#). A discussion forum for the tools and resources is provided at [Cosmocoffee](#).

NEWS:
21/05/2009 - **w(z) eigenfunctions**. Module for [astro-ph/0905.3383](#) to be included in [iCosmo v1.2](#).
20/05/2009 - **Hardware-Software balance**. Code for [astro-ph/0905.3176](#) can be downloaded here [iCosmo PublicAstroCodes](#).
11/02/2009 - **Redshift Distortion & ISW**. Module for [astro-ph/0902.1759](#) to be included in [iCosmo v1.2](#).
21/01/2009 - **Cloud Cosmology**. Article available [here](#). Template web pages available [here](#).

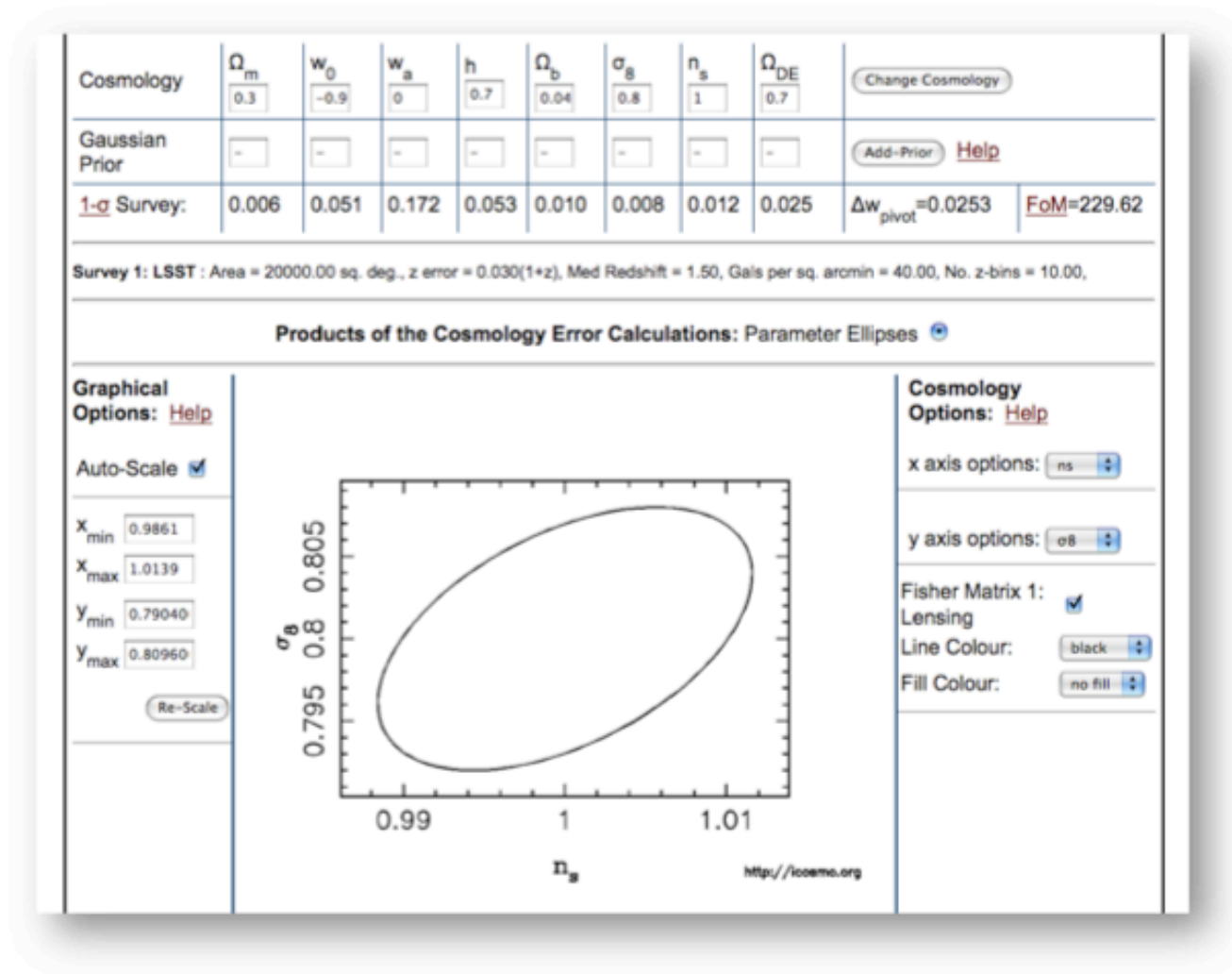
QuickStart Calculator

Ω_m	<input type="text" value="0.3"/>	Ω_{DE}	<input type="text" value="0.7"/>
Ω_b	<input type="text" value="0.045"/>	w_0	<input type="text" value="-0.95"/>
h	<input type="text" value="0.7"/>	w_s	<input type="text" value="0.0"/>
σ_8	<input type="text" value="0.8"/>	n_s	<input type="text" value="1.0"/>

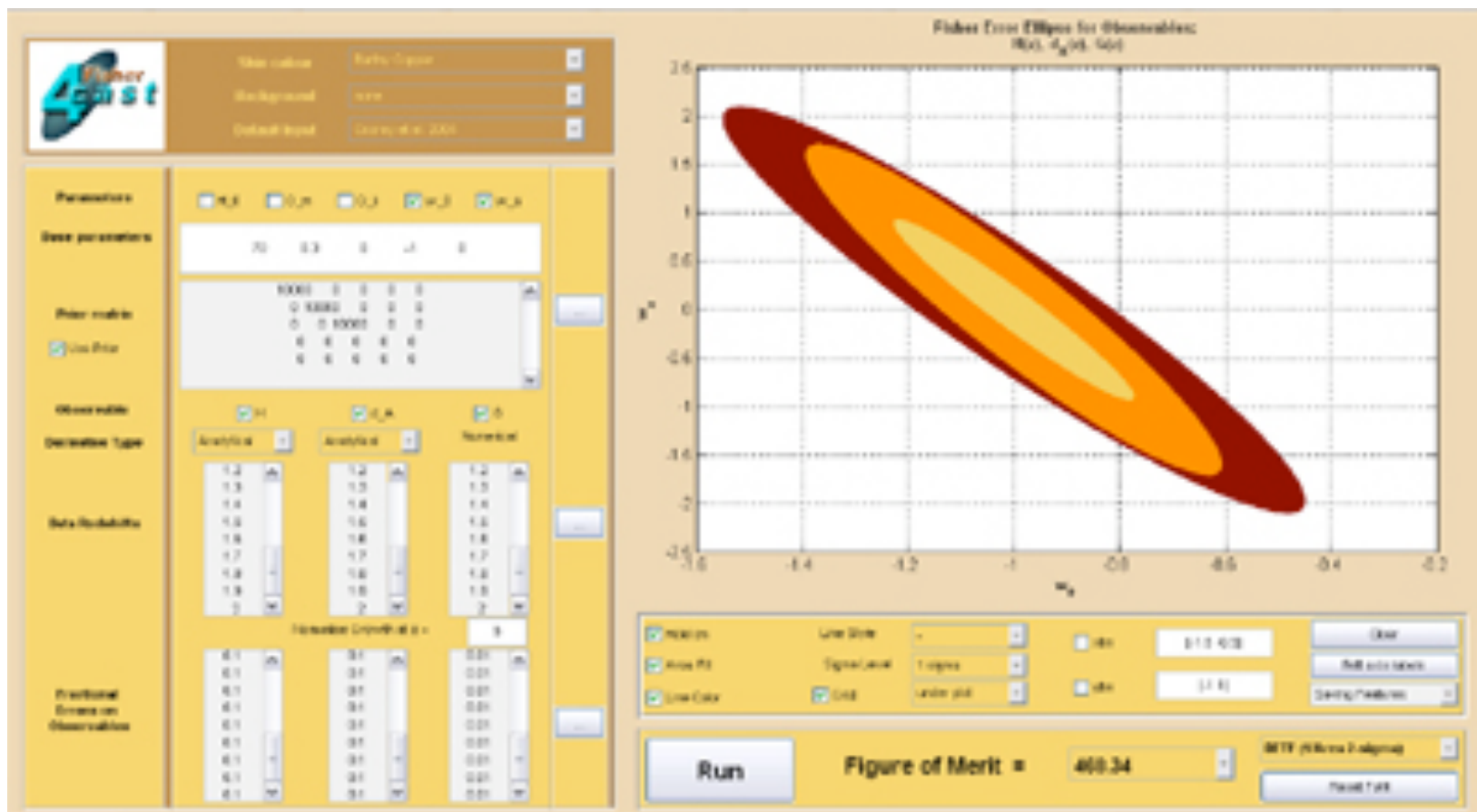
QuickStart Tutorial

- Gravitational Lensing
- Galaxy Correlations
- CMB

Open source Fisher matrices



Fisher 4cast (Matlab toolbox)



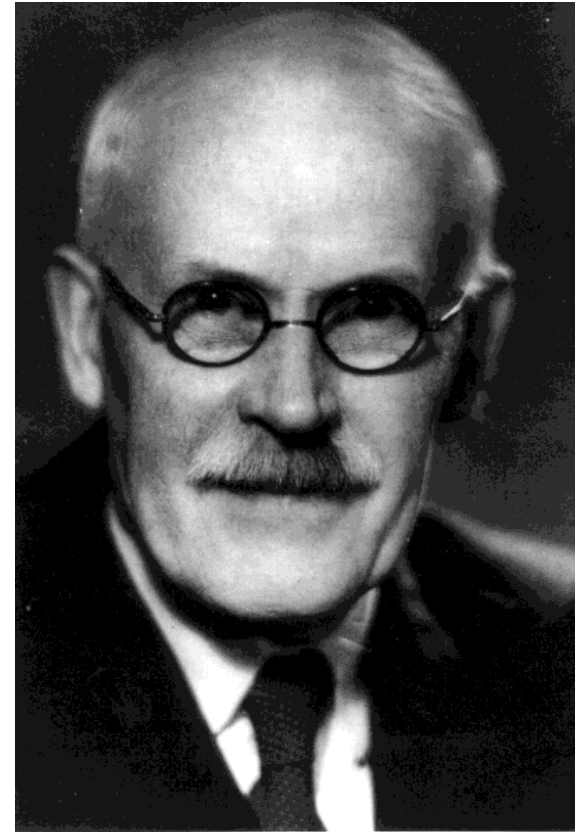
Counting cars, again

- In our study of the car counts we implicitly assumed that all the values of λ are equally likely a priori (i.e. before we started taking the data). However, we didn't consider that an automatic gate regulates the traffic in our street and does not allow more than 8 cars to enter every 10 minutes. Therefore λ cannot be larger than 8 and the likelihood derived from our counts should have been truncated at $\lambda = 8$.
- Also, we live close to a church and whenever there is a wedding the traffic is more intense than usual. This means that on wedding days a higher value of λ is more likely than on non-wedding days.
- Moreover, a fellow that had been living in our flat before us did the same exercise and told us that he obtained $\lambda = 4.2 \pm 0.5$.
- Is there a way to account for all this information in our study?

The Bayesian way



Bruno de Finetti (1906-1985)



Harold Jeffreys (1891-1989)

What is probability?

- **Probability is a modern concept** first discussed in a correspondence between Blaise Pascal and Pierre de Fermat in 1654
- There is no unique definition of probability, statisticians are divided into different schools with contrasting views
- **Classic definition:** The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible. (Pierre-Simon de Laplace, [*Théorie analytique des probabilités*](#), 1812)
- This is based on the “principle of insufficient reason” (or principle of indifference) which states that when cases are only distinguishable by their name they should be assigned the same probability

What is probability?

- **Frequentist:** the long-run expected frequency of occurrence of a random event
- **Axiomatic:** given a sample space Ω , a σ -algebra F of events E (a set of subsets of Ω), we call probability measure a real function on F such that $P(E) \geq 0$, $P(\Omega) = 1$, and for any countable series of pairwise disjoint events $P(E_1 \cup E_2 \cup \dots \cup E_N) = P(E_1) + P(E_2) + \dots + P(E_N)$. These are known as Kolmogorov axioms.
- **Bayesian:** a measure of the degree of belief (the plausibility of an event given incomplete knowledge)

Reasoning with beliefs

- There is 90% chance that today it will rain
- There is a 30% chance that my favourite football team will win the league this year
- There is a 10% chance that I will fail the observational cosmology examination
- There is a 0.1% chance that I will die before being 30
- There is 68.3% chance that H_0 lies between 67 and 73 km/s/Mpc

De Finetti's game

Can you measure degree of belief?

Suppose we are on a trip and you say that you are "pretty sure" you locked the door of your flat. I want to determine how sure you are.

- I offer you to play a game: I propose you to draw a marble from a bag containing 95 red and 5 blue marbles. If you pick at random a red marble, I give you one million euros. Alternatively, I offer you to go back home and check the door. If you choose this option and the door is locked indeed, I give you one million euros.
- If you choose to pick a marble, it means that your degree of belief is lower than 95%
- I can then propose many other rounds of the game by progressively reducing the fraction of red marbles until you choose to go back. This would measure your degree of belief.