Parameter estimation and forecasting

Cristiano Porciani
AIfA, Uni-Bonn

Questions?

Cosmological parameters

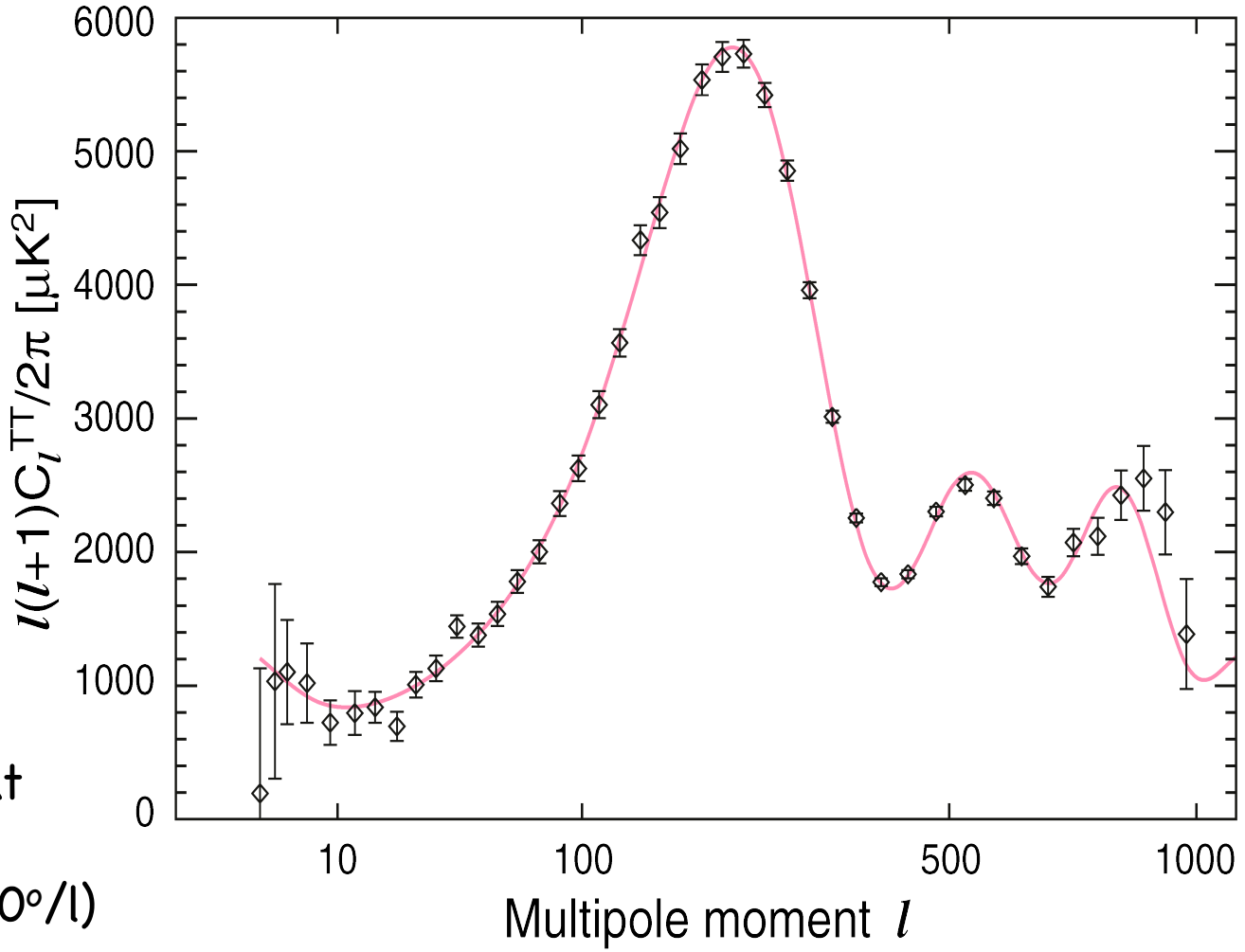
- A branch of modern cosmological research focuses on measuring cosmological parameters from observed data (e.g. the Hubble constant, the cosmic density of matter, etc.).

- 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

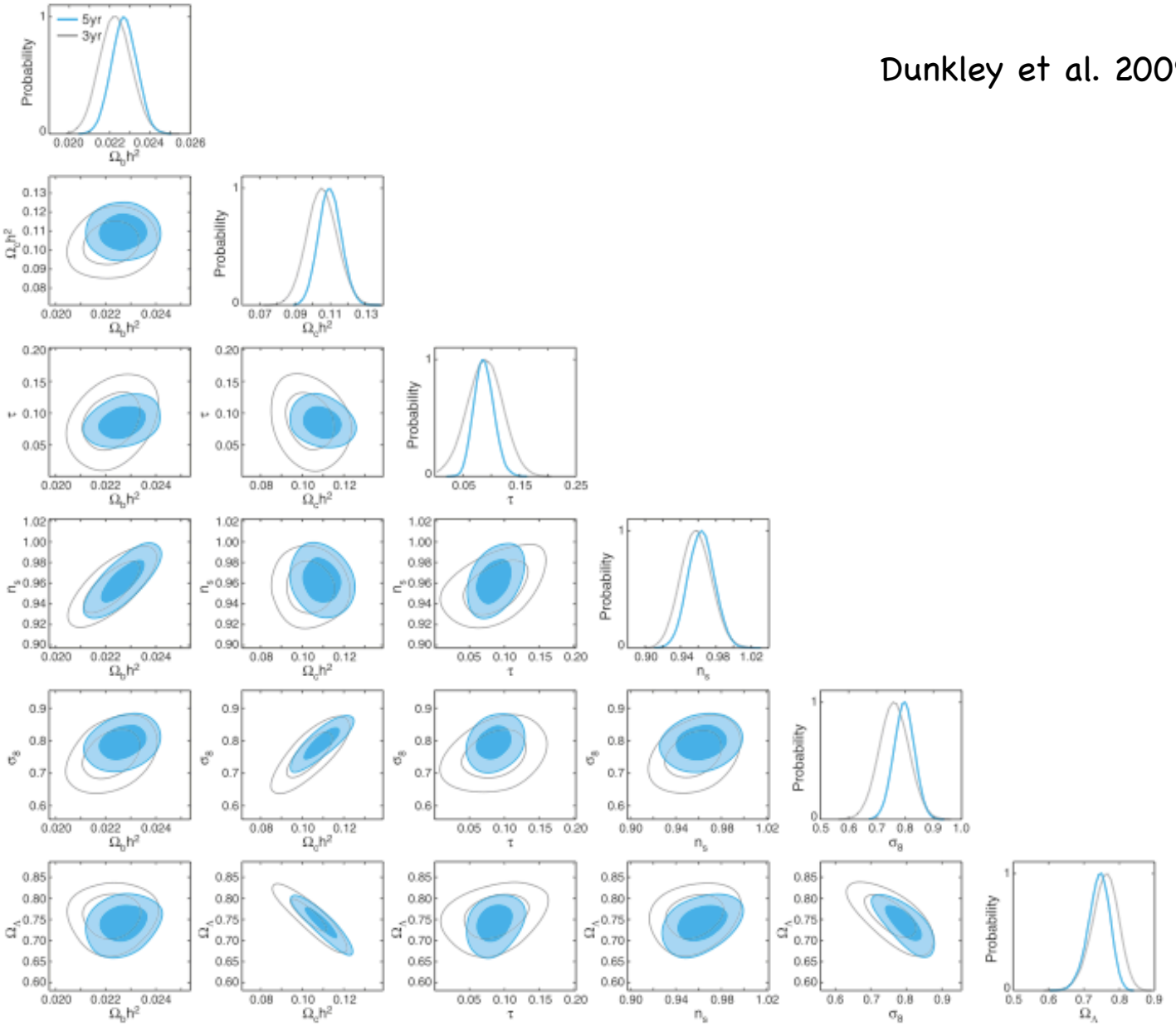
- How do you fit a model to data?
- How do you incorporate prior knowledge?
- How do you merge multiple sources of information?
- How do you treat uncertainties in model parameters?

Example: power spectrum of CMB temperature fluctuations

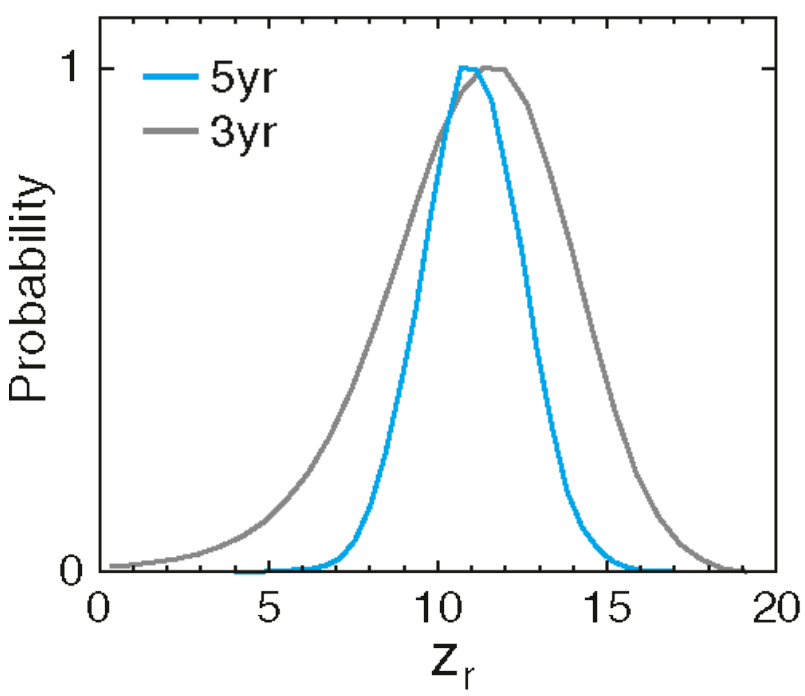


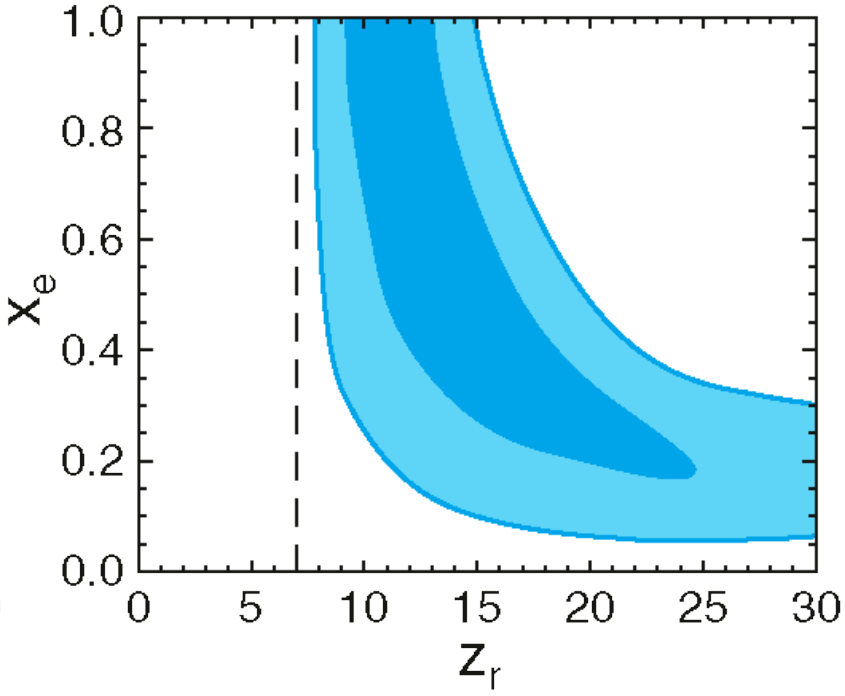
Variance at multipole l (angle ~180°/l)

Dunkley et al. 2009



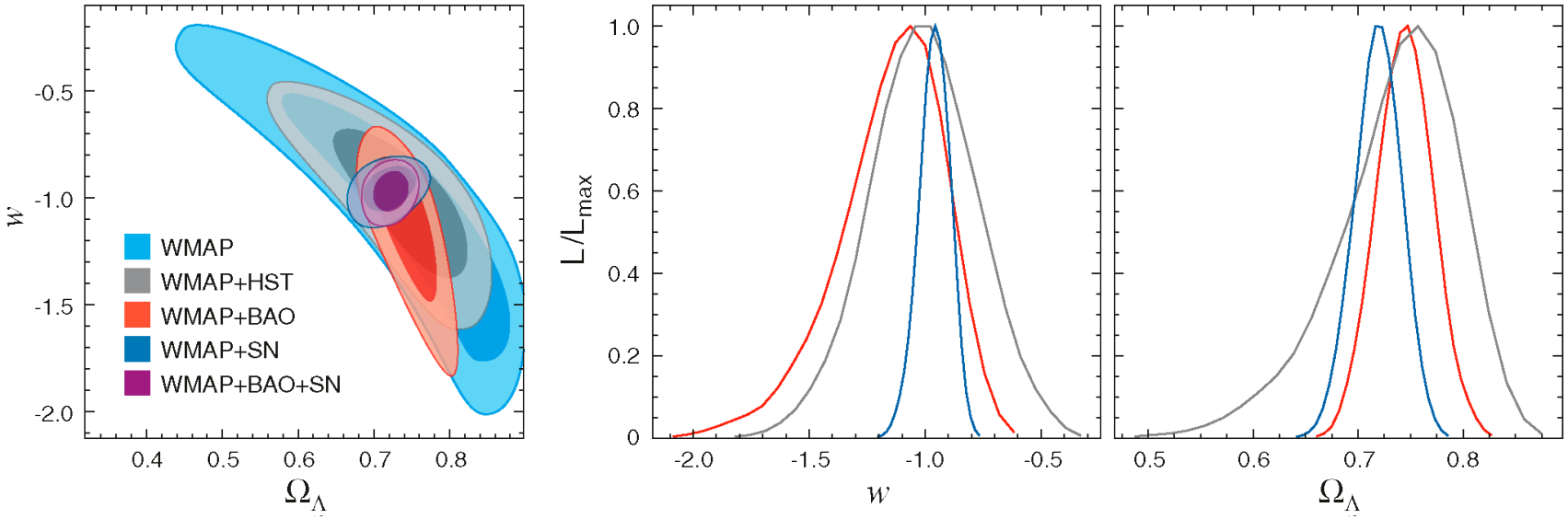
Dunkley et al. 2009





The current state of the art

Bennett 2006



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?

Basic introduction to statistics

Descriptive statistics

Input
- a set of data

Output
- The sample mean or median
- The sample 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

Statistical hypothesis testing

- One of the classical applications of statistics

- It generally consists of three steps

1. Formulate the "null hypothesis" H_0 and an "alternative hypothesis". The null hypothesis (i.e. what one is trying to rule out) is often "devil's advocate position", e.g. "treatment with a given drug does not lead to any improvement of a given physical condition". An alternative hypothesis could be "treatment with the same drug removes a specific symptom or prolongs life expectancy".

Statistical hypothesis testing II

2. Pick a "test statistic" S that will be used to make the inference (i.e. a method to decide if the experimental data favor one hypothesis with respect to the other).

3. Pick a "confidence level" that will fix a threshold T on the value of the statistic for the inference process, i.e. if S>T the null hypothesis is ruled out at the chosen confidence level (this part will become clearer later on, for the moment just worry about understanding the philosophy behind the method).

Statistical hypothesis testing III

- Statistical hypothesis testing is a decisional method based on limited data (e.g. drug efficiency, metal detector at airport, pregnancy test) and as such can (and will) make mistakes.

- Statisticians generally distinguish two types of errors:

	H_0 is true	H_0 is false
Test fails to reject H_0	Correct inference (True positive)	Type II error (False negative)
Test rejects H_0	Type I error (False positive)	Correct inference (True negative)

- The details of the test method (i.e. which test statistic and what confidence level) can be tuned to optimize the blend of Type I and Type II errors depending on the application.

Sampling and estimation

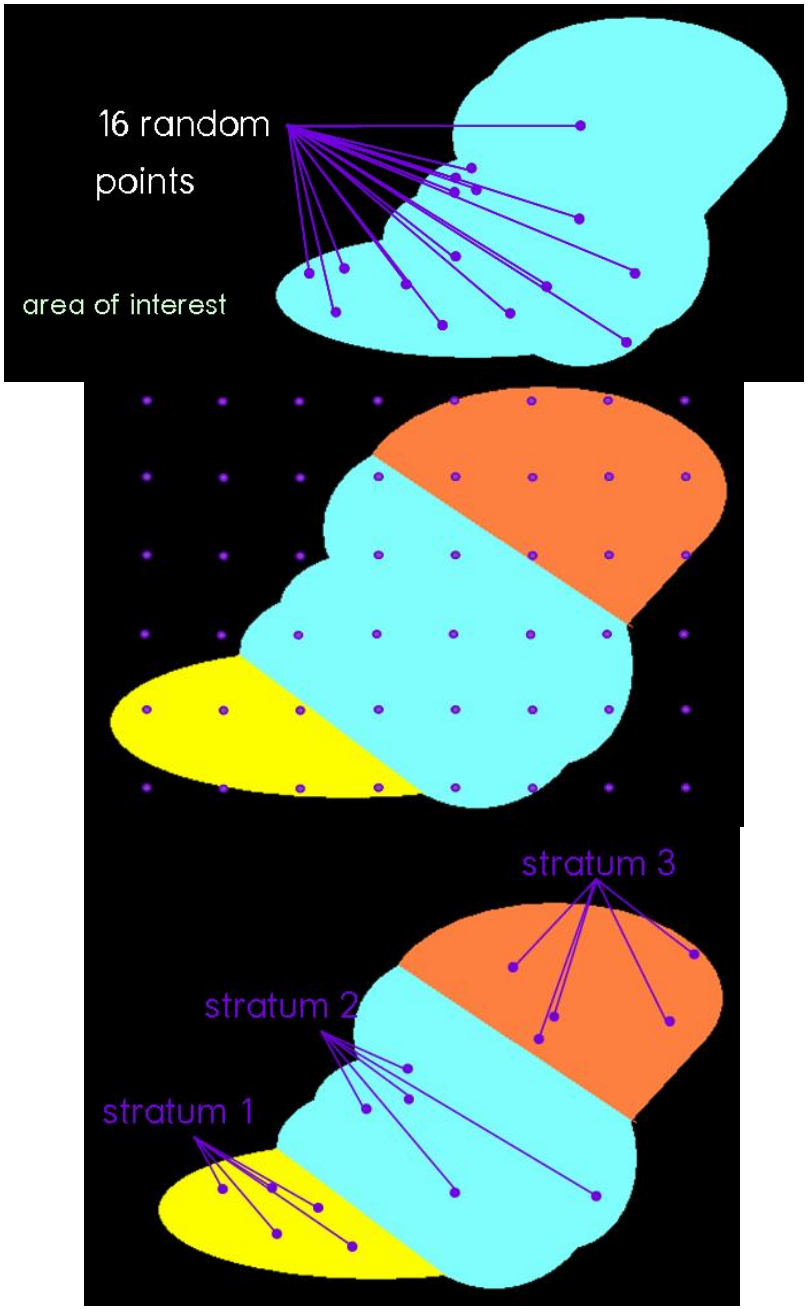
Population and sample

- 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.

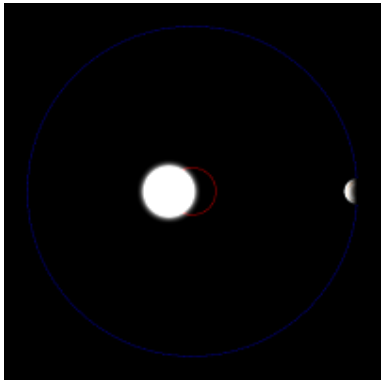
- A sample is a group of units selected from the population for study because the population is too large to study in its entirety. For each population there are many possible samples.

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

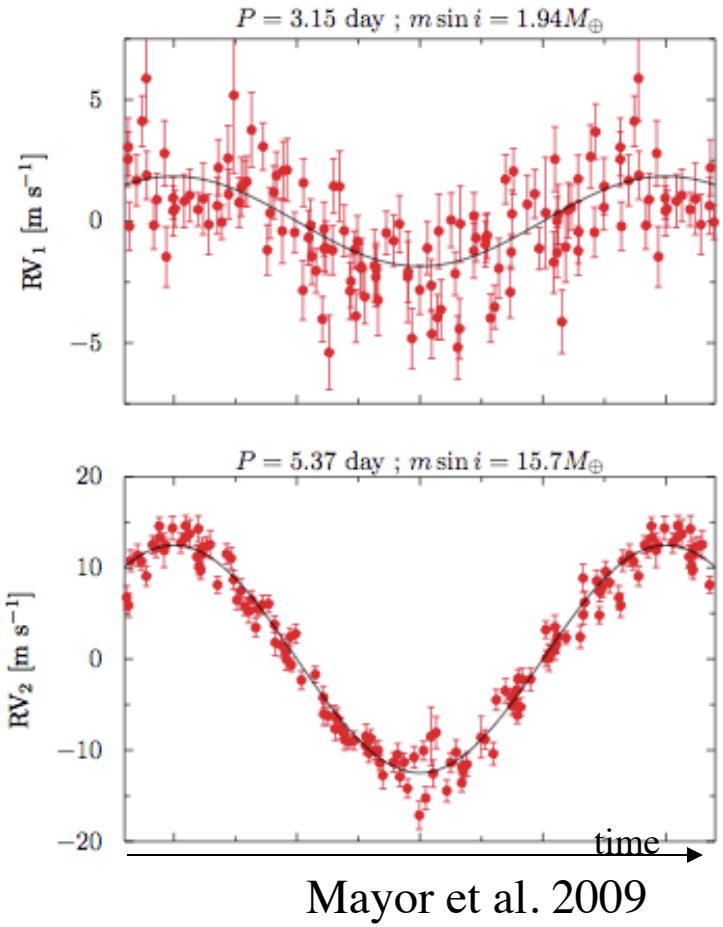


Example: extra-solar planets from Doppler surveys



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

The method is best at detecting "hot Jupiters", very massive planets close to the parent star. Current ground-based spectrographs (e.g. HARPS, HIRES) 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.

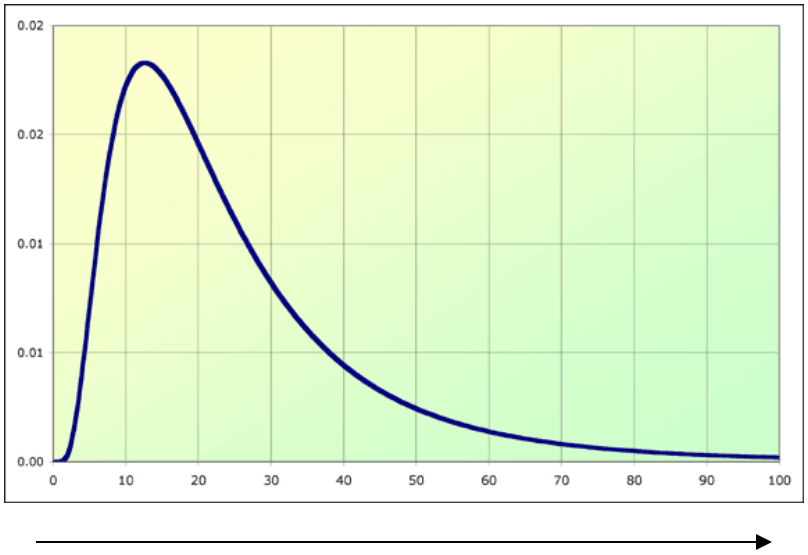
A simple example

- 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

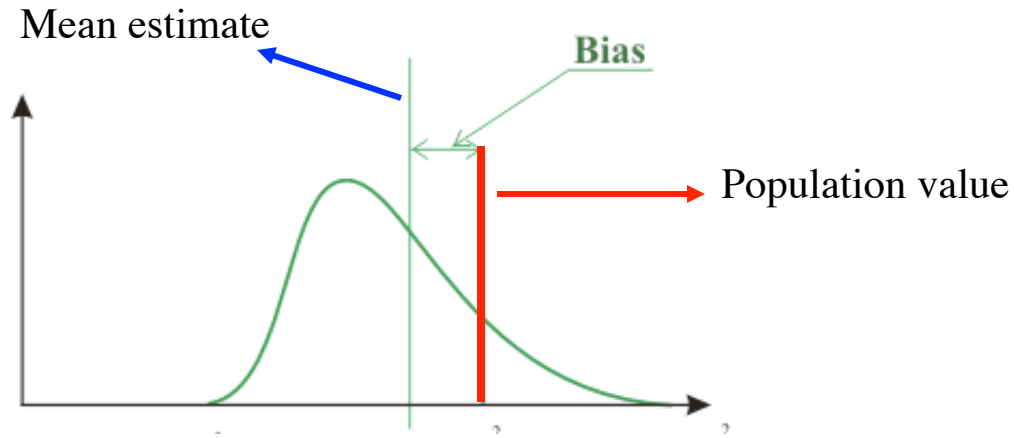
PDF of an estimator

- 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



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

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

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

Examples

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

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

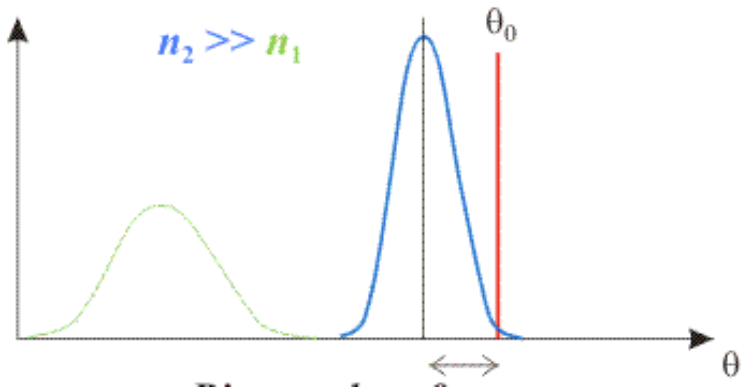
- Exercise: Is the sample variance an unbiased estimator of the population variance?

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

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

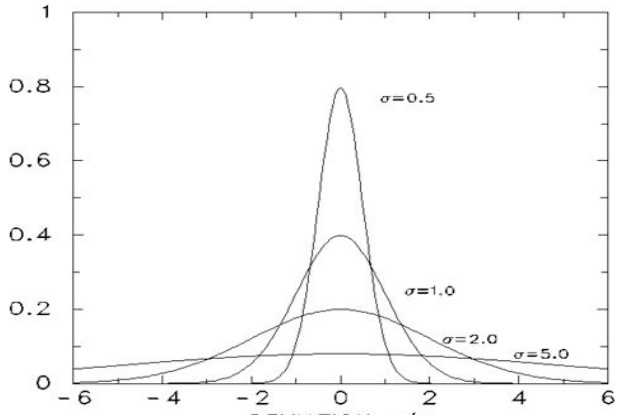
- The sample mean is a consistent estimator of the population mean

$$Var[\bar{x}] = E[(\bar{x} - \mu)^2] = E[(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu)^2] = \frac{1}{N^2} E[(\sum_{i=1}^{N} x_i)^2] - 2\frac{\mu}{N} N\mu + \mu^2 =$$

$$= \frac{1}{N^2} N(\mu^2 + \sigma^2) + \frac{N(N-1)}{N^2}\mu^2 - \mu^2 = \frac{\sigma^2}{N}$$

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

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

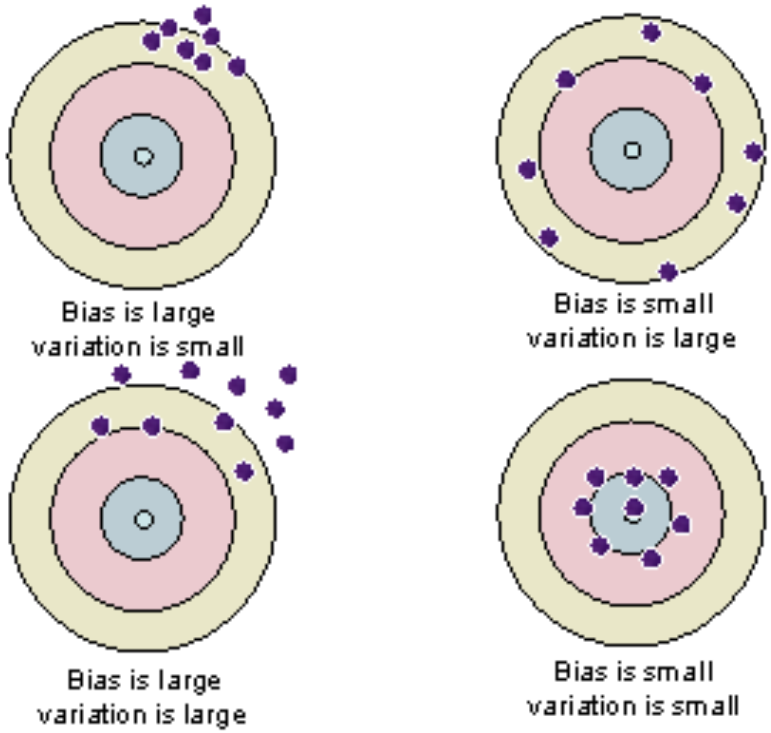
- A theorem known as the Cramer-Rao bound (see slide 45) proves that the variance of an unbiased estimator must be larger of or equal to a specific value which only depends on the sampling strategy (it corresponds to the reciprocal of the Fisher information of the sample)

- 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



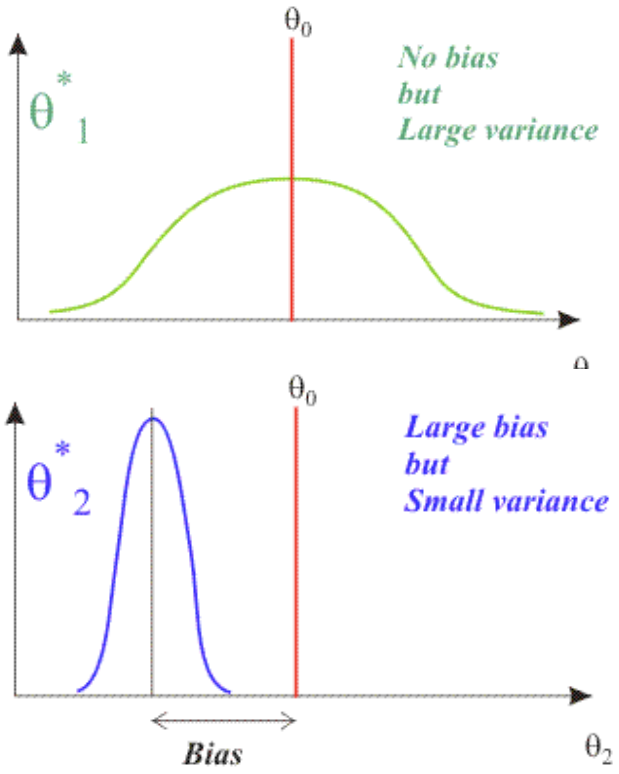
Desirable properties of an estimator

- ✓ 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,

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

- 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 $X_1, \ldots, X_n$ 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 $\theta_1, \ldots, \theta_k$ with $k<n$.

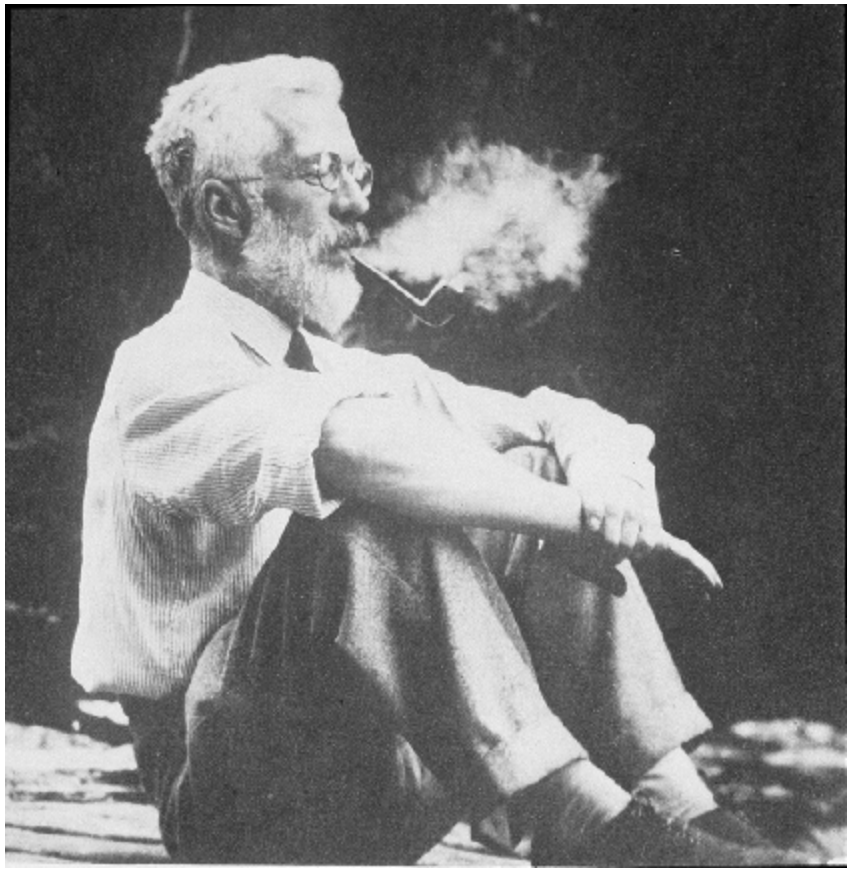
The old way: method of moments

- 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)

Ronald Aylmer 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 (IR. 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:

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

- 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:

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

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.