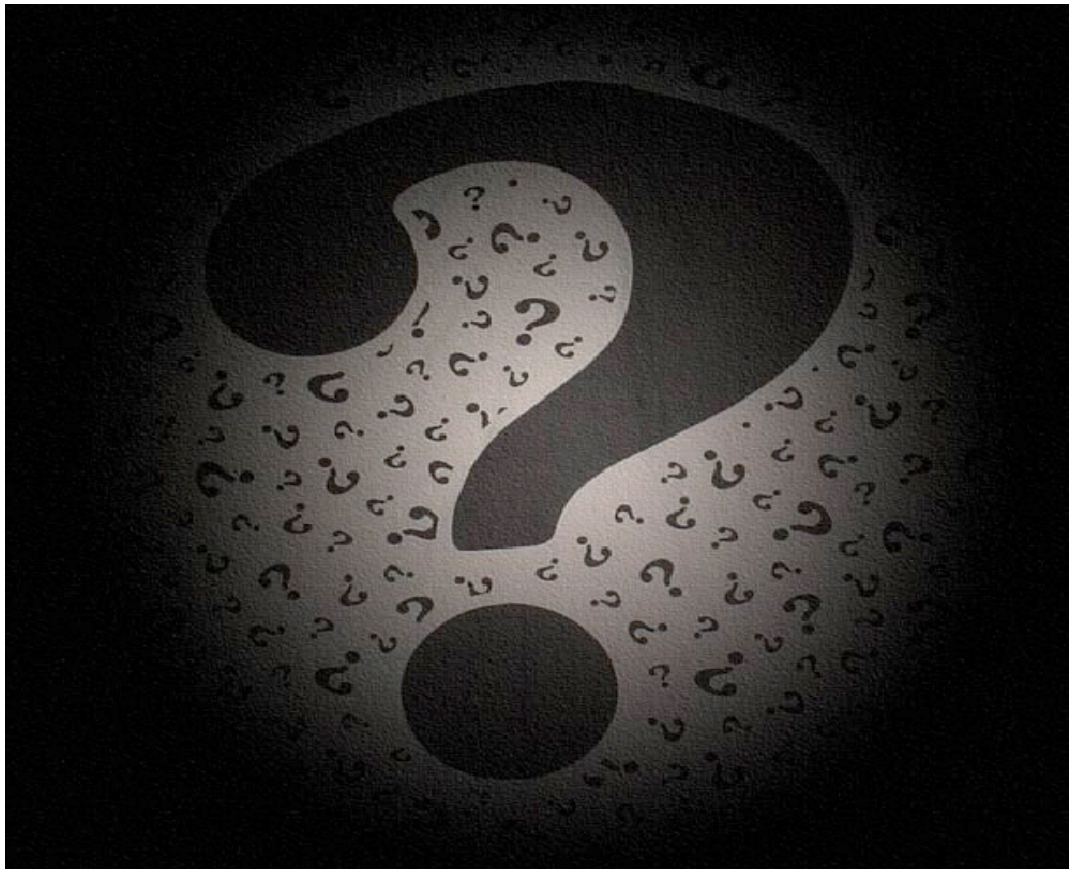
Estimation problems

Cristiano Porciani

AIfA, Bonn

Please ask questions!!!



My coordinates

- Cristiano Porciani, Argelander Institute für Astronomie, Auf dem Hügel 71, D-53121, Bonn

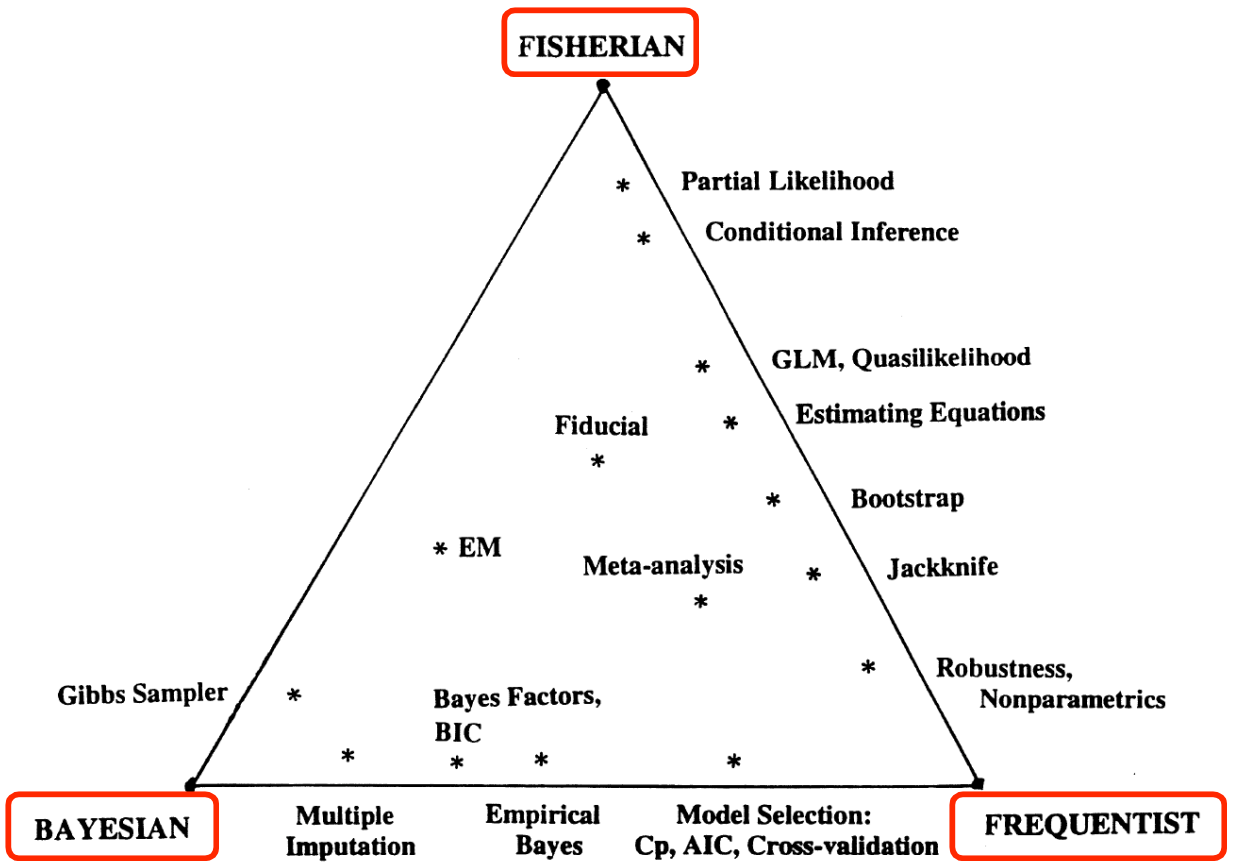
- porciani@astro.uni-bonn.de

- Cosmology, large-scale structure of the universe, intergalactic medium

The coordinates of statistics

Bradley Efron's triangle (1998)

Fisher



de Finetti, Savage
Jeffrey

Neyman, the Pearson's

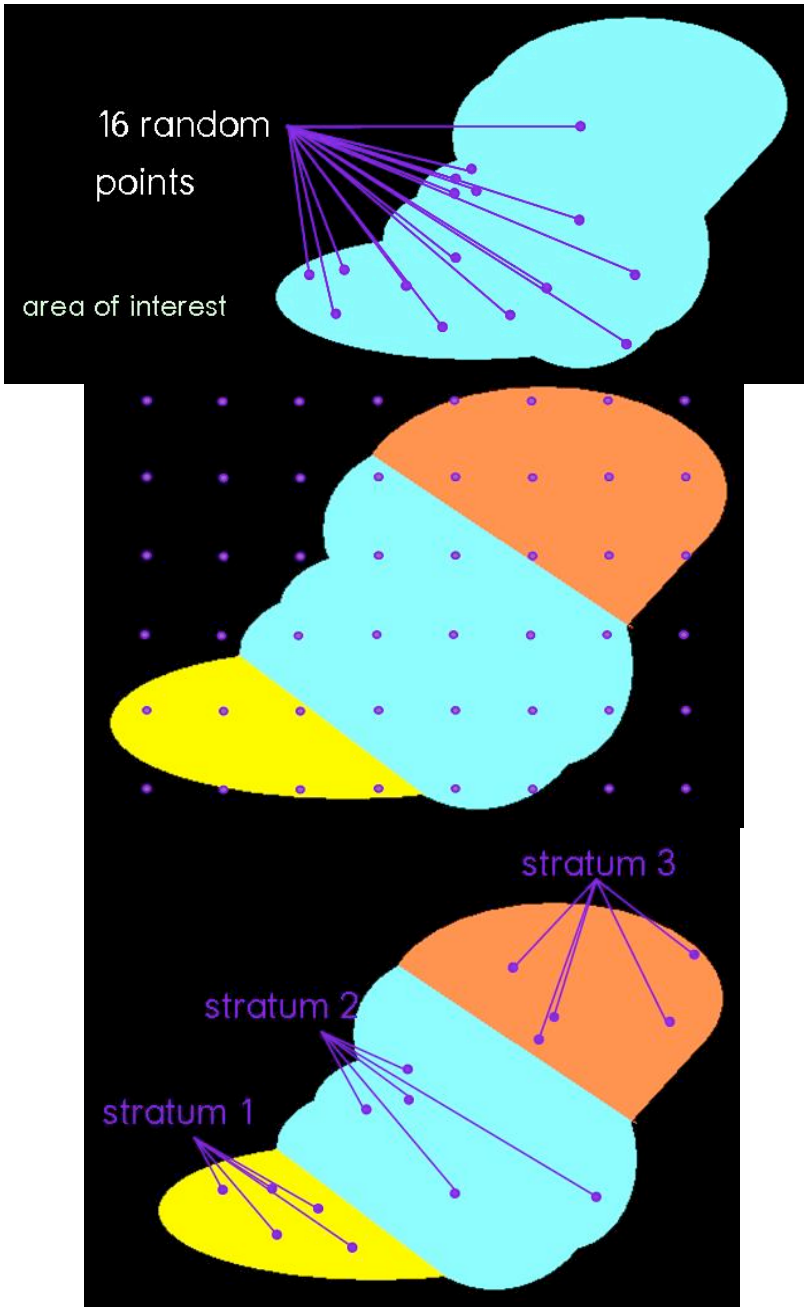
Population and sample

- 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.

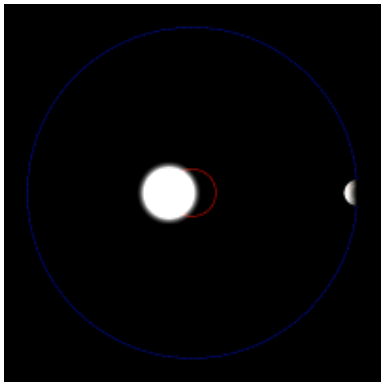
- A sample is a group of units selected from the population for study because the population is too large to study in its entirety. For each population there are many possible samples.

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

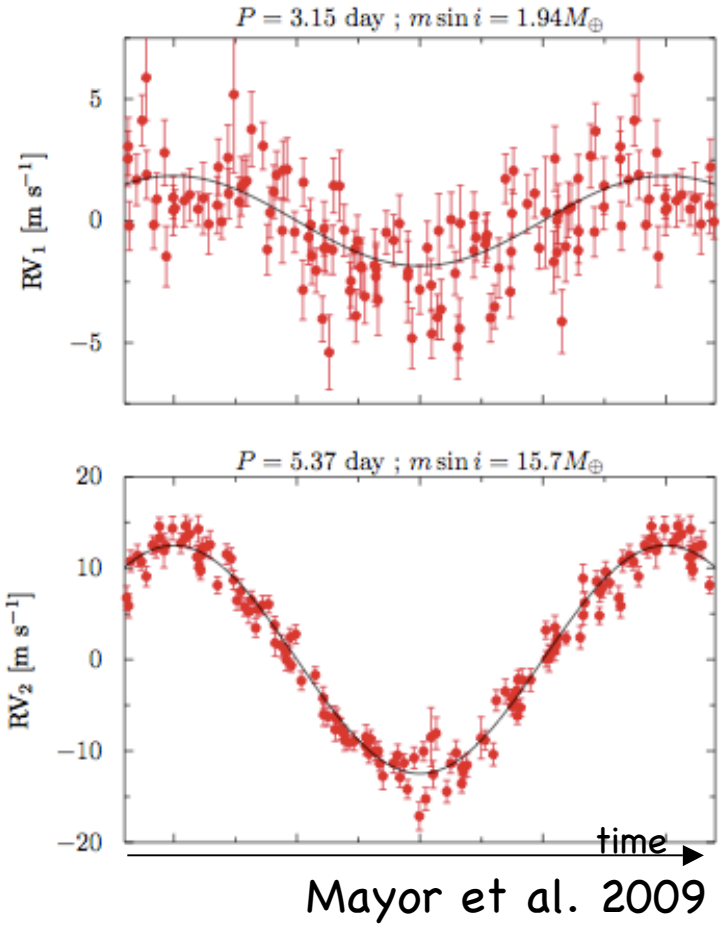


Example: extra-solar planets from Doppler surveys



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

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.

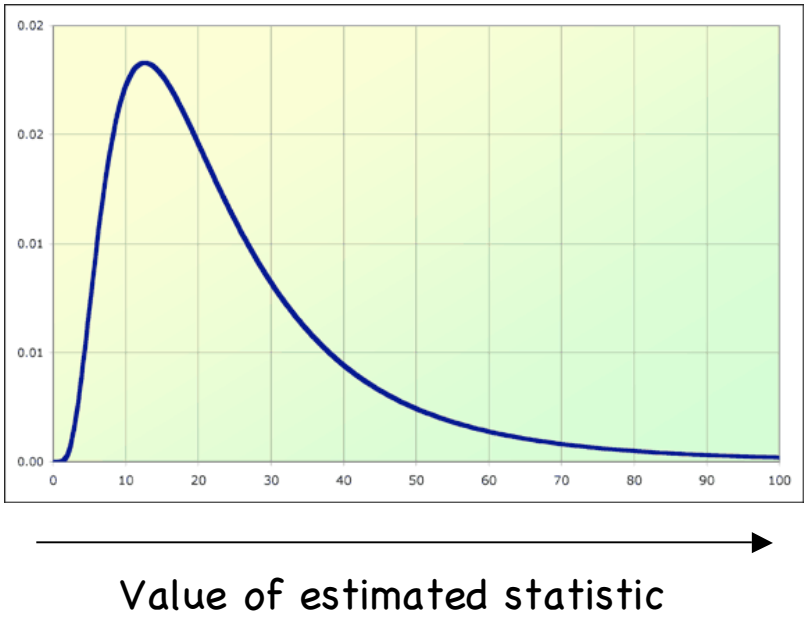
A simple example

- 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

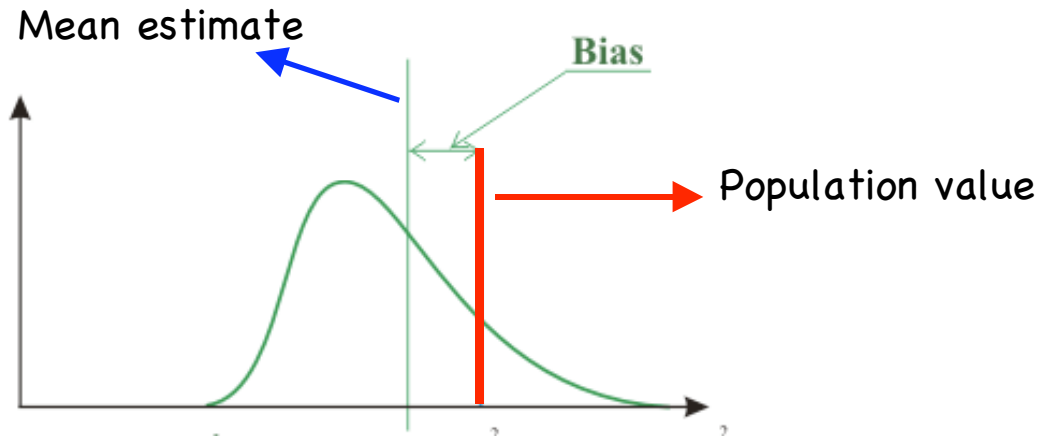
PDF of an estimator

- 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



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

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

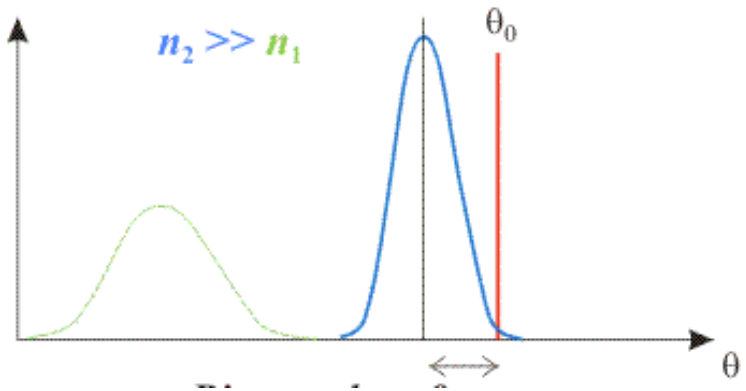
- Exercise: Is the sample variance an unbiased estimator of the population variance?

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

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

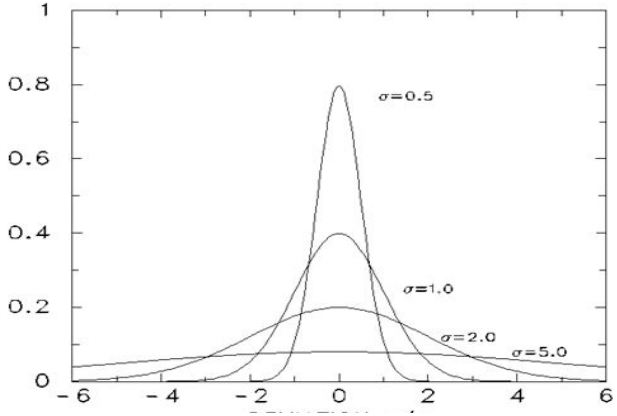
- The sample mean is a consistent estimator of the population mean

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

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

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

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

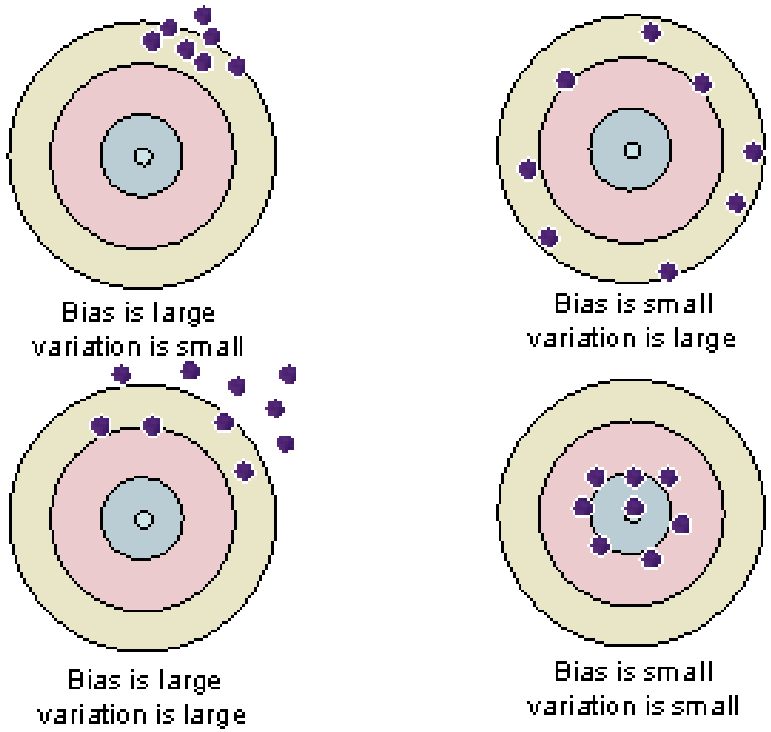
- A theorem known as the Cramer-Rao bound (see Alan Heaven's lectures) proves that the variance of an unbiased estimator must be larger of or equal to a specific value which only depends on the sampling strategy (it corresponds to the reciprocal of the Fisher information of the sample)

- 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



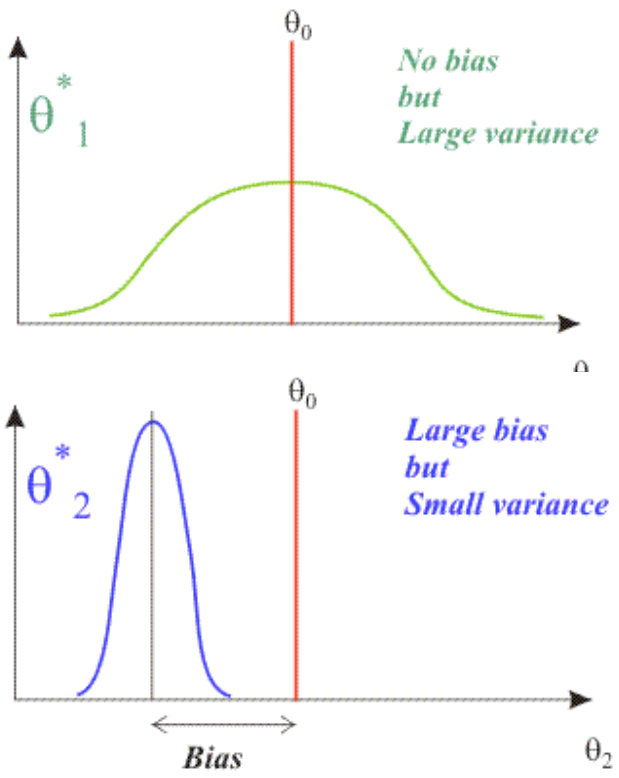
Desirable properties of an estimator

- ✓ 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, ... , X_n of independent, identically distributed variables (iid).

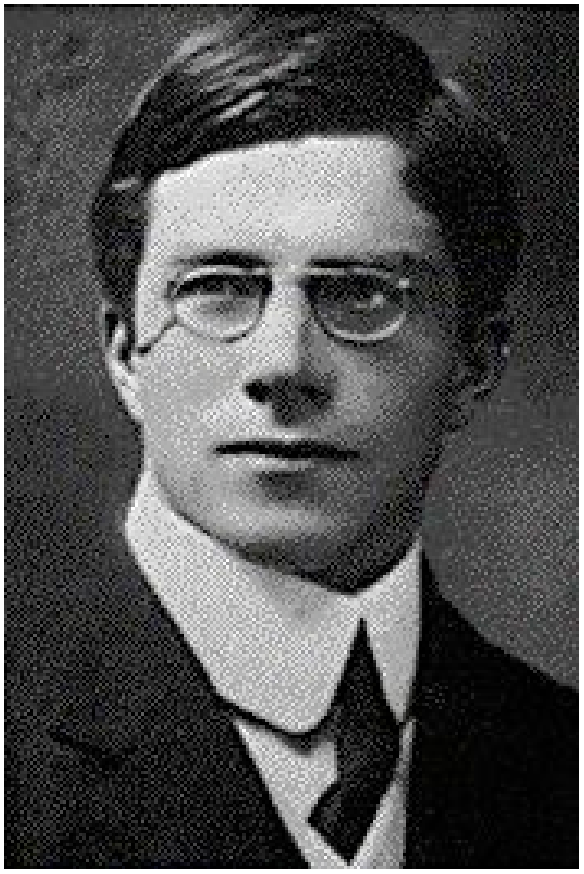
We also assume that the shape of the population PDF of X is known (Gaussian, Poisson, binomial, etc.) but has k unknown parameters θ_1, ... , θ_k with k<n.

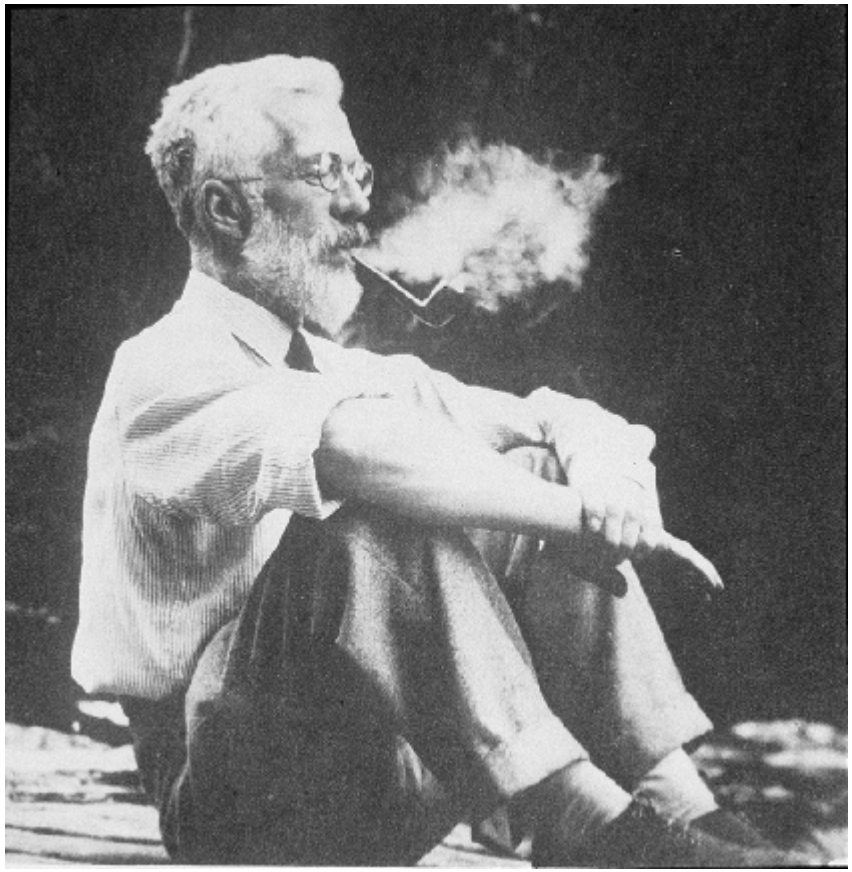
The old way: method of moments

- 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

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)

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)

The Likelihood function

- 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 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 θ_1, ... , θ_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

$$\frac{\partial L(\theta_1,...,\theta_k,x_1,...,x_n)}{\partial \theta_i} = 0 \qquad \text{or} \qquad \frac{\partial \ln L(\theta_1,...,\theta_k,x_1,...,x_n)}{\partial \theta_i} = 0$$

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

Properties of MLE's

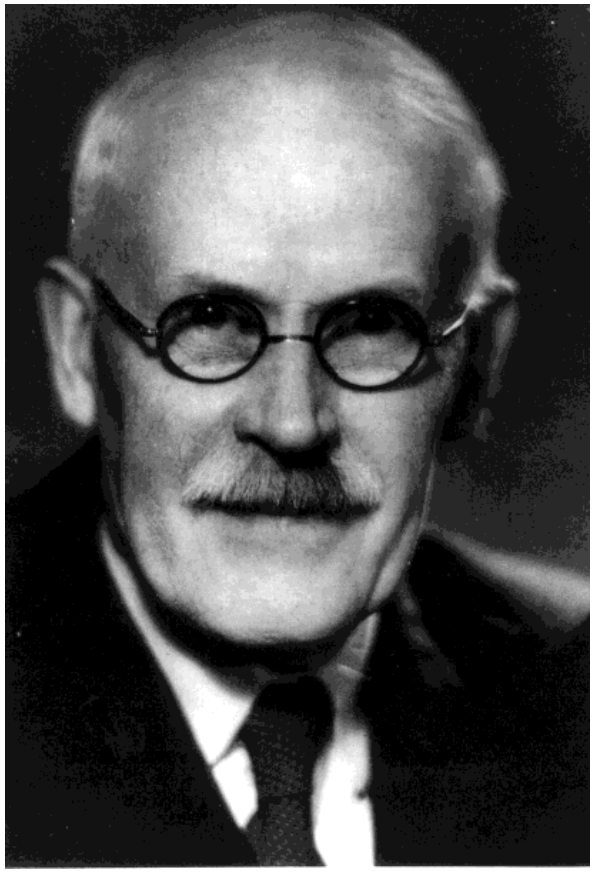
As the sample size increases to infinity (under weak regularity conditions):

- MLE's become asymptotically efficient and asymptotically unbiased
- MLE's asymptotically follow a normal distribution with covariance matrix equal to the inverse of the Fisher's information matrix (see Alan Heaven's lectures)

However, for small samples,

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

The Bayesian way



Bayesian estimation

- In the Bayesian approach to statistics (see Jasper Wall's lectures), population parameters are associated with a posterior probability which quantifies our degree of belief in the different values

- Sometimes it is convenient to introduce estimators obtained by minimizing the posterior expected value of a loss function

- For instance one might want to minimize the mean square error, which leads to using the mean value of the posterior distribution as an estimator

- If, instead one prefers to keep functional invariance, the median of the posterior distribution has to be chosen

- Remember, however, that whatever choice you make is somewhat arbitrary as the relevant information is the entire posterior probability density.

Example I:
the galaxy luminosity function

V_{max} method (non-parametric estimator)

- Find the largest distance at which a galaxy with observed abs magnitude M_i can be found in order to have apparent magnitude equal to the limit of the sample m_{lim}

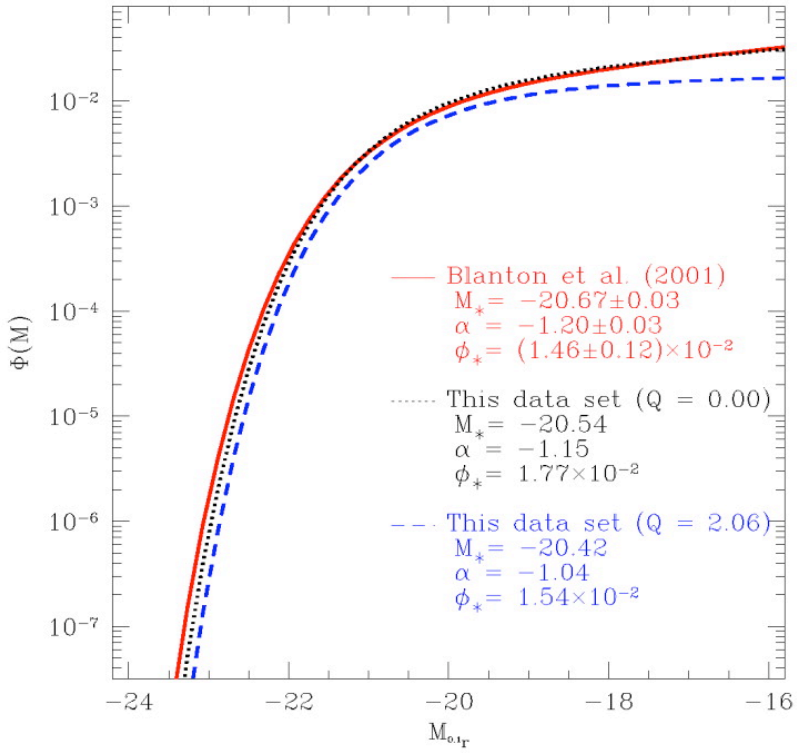
- Volume of the sample corresponding the distance is V_{max}. This is the volume available for the galaxy. The galaxy could have been anywhere inside the volume.

- Select all galaxies with abs magnitudes in the range (M,M+dM). An estimate of the luminosity function is

- $\Phi(M)dM = \sum[1/V_{max}(i)]$

Courtesy A. Klypin

An example of MLE



Blanton et al. 2003

Schechter function:

$$\phi(M) = (0.4 \ln 10)\ \phi^*\ [10^{0.4(M^*-M)}]^{1+\alpha}\ \exp[-10^{0.4(M^*-M)}]$$

$$\Phi(L) = \left(\frac{\Phi^*}{L^*}\right)\left(\frac{L}{L^*}\right)^{\alpha} \exp\left(-\frac{L}{L^*}\right)$$

Parametric maximum-likelihood method of Sandage, Tammann, Yahil(1979)

Consider a galaxy i observed at redshift z_i in a flux-limited sample. Apparent magnitude limits for the sample are m_{min} and m_{max}. The differential luminosity function of the sample is $\Phi(M)$, where M is the absolute magnitude. The probability for for a galaxy at redshift z_i to be in the sample is p_i:

$$p_i \equiv p(M_i|z_i) = \phi(M_i) \Big/ \int_{M_{\rm min}(z_i)}^{M_{\rm max}(z_i)} \phi(M) dM$$

The likelihood function $\mathfrak{L}$ for having a sample of N galaxies with abs.magnitudes Mi is the product of probabilities pi:

$$\mathfrak{L} = p(M_1, \ldots, M_N|z_1, \ldots, z_N) = \prod_{i=1}^{N} p_i \ ,$$

It is more convenient to deal with the ln of the function:

$$\ln \mathfrak{L} = \sum_{i=1}^{N} \left\{ \ln \phi(M_i) - \ln \int_{M_{\rm min}(z_i)}^{M_{\rm max}(z_i)} \phi(M) dM \right\}$$

It Assume a parametric form for $\Phi(M) = \Phi(M;p1,p2,...)$. Maximize $\mathfrak{L}$ with respect to those parameters. In practice, we use the Schechter function, which has three free parameters.

Courtesy A. Klypin

Normalization

The overall normalization $\bar{n}$ cannot be determined from this likelihood maximization procedure. We use the standard minimum variance estimator of Davis & Huchra (1982) to perform the normalization:

$$\bar{n} = \frac{\sum_{j=1}^{N_{\rm gals}} w(z_j)}{\int dV\, \phi(z) w(z)} \,, \qquad (9)$$

where the integral is over the volume covered by the survey between the minimum and maximum redshifts used for our estimate. The weight for each galaxy is

$$w(z) = \frac{f_t}{1 + \bar{n} 10^{0.4P(z-z_0)} J_3 \phi(z)} \,,$$

and the selection function is

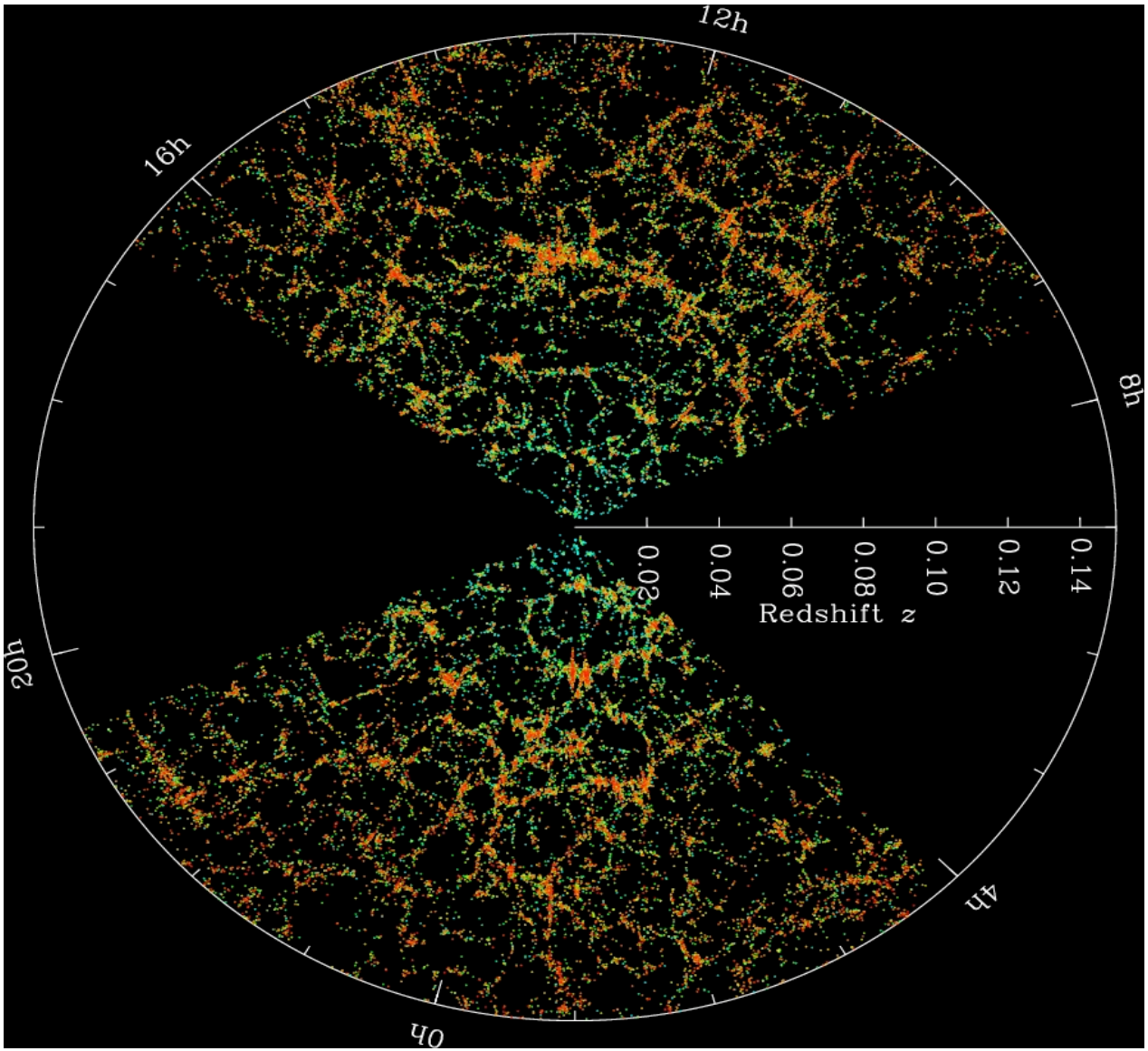
$$\phi(z) = \frac{\int_{L_{\rm min}(z)}^{L_{\rm max}(z)} dL\, \Phi(L,z)}{\int_{L_{\rm min}}^{L_{\rm max}} dL\, \Phi(L,z)} \,, \qquad (11)$$

where f_t is the galaxy sampling rate determined at each position of sky as the fraction of targets in each sector that were successfully assigned a classification. The integral of the correlation function is

$$J_3 = \int_0^\infty dr\, r^2 \xi(r) = 10,000\ h^{-3}\ {\rm Mpc}^3 \,. \qquad (12)$$

Courtesy A. Klypin

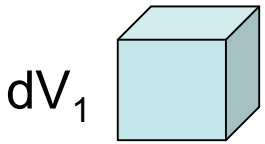
Example II:
the galaxy 2-point correlation function

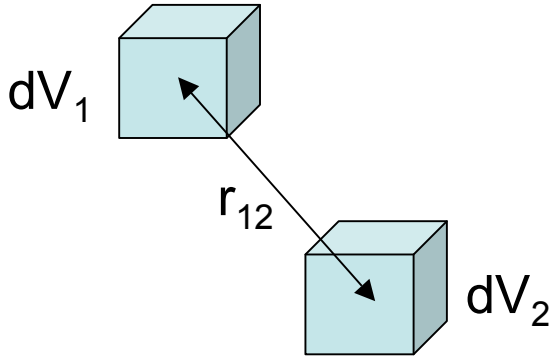


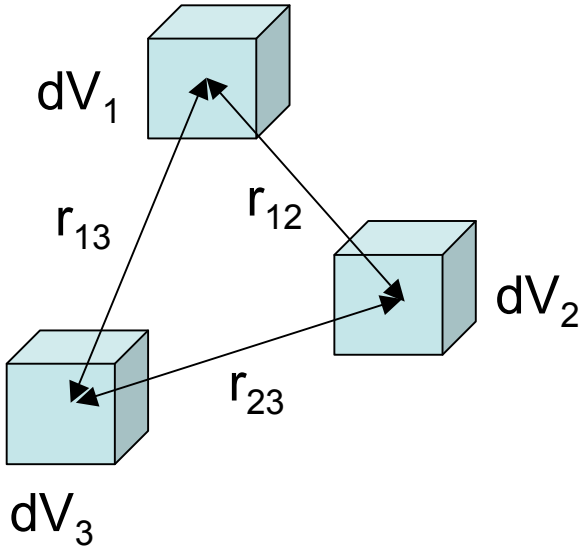
Courtesy SDSS

Correlation functions

Consider a stationary point process with mean density n and write the probability of finding N points within N infinitesimal volume elements





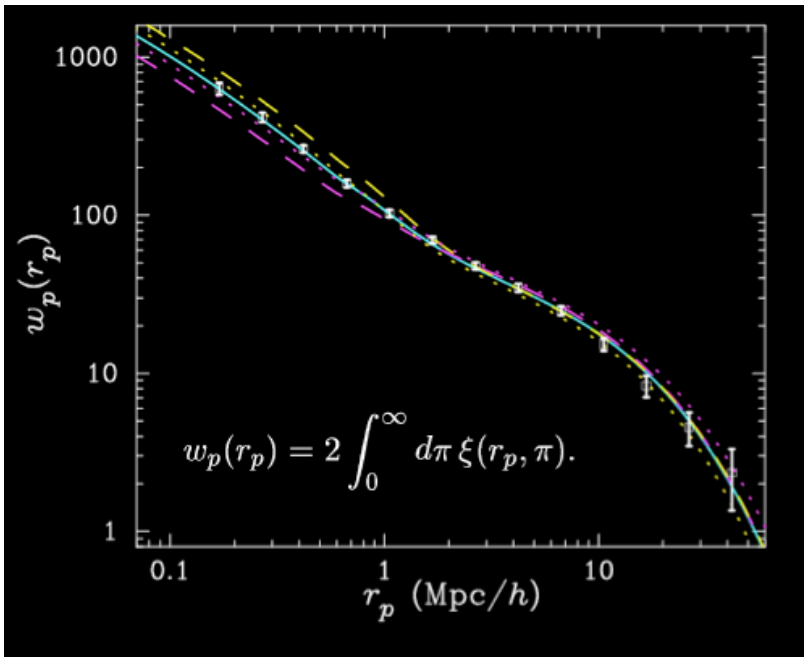


$$dP_1 = n\, dV_1$$

$$dP_{12} = n^2\,(1+\xi(r_{12}))\, dV_1\, dV_2$$

$$dP_{123} = n^3\,(1+\xi(r_{12})+\xi(r_{13})+\xi(r_{23})+\zeta(r_{12,}\, r_{13,}\, r_{23}))\, dV_1\, dV_2\, dV_3$$

Estimators for ξ



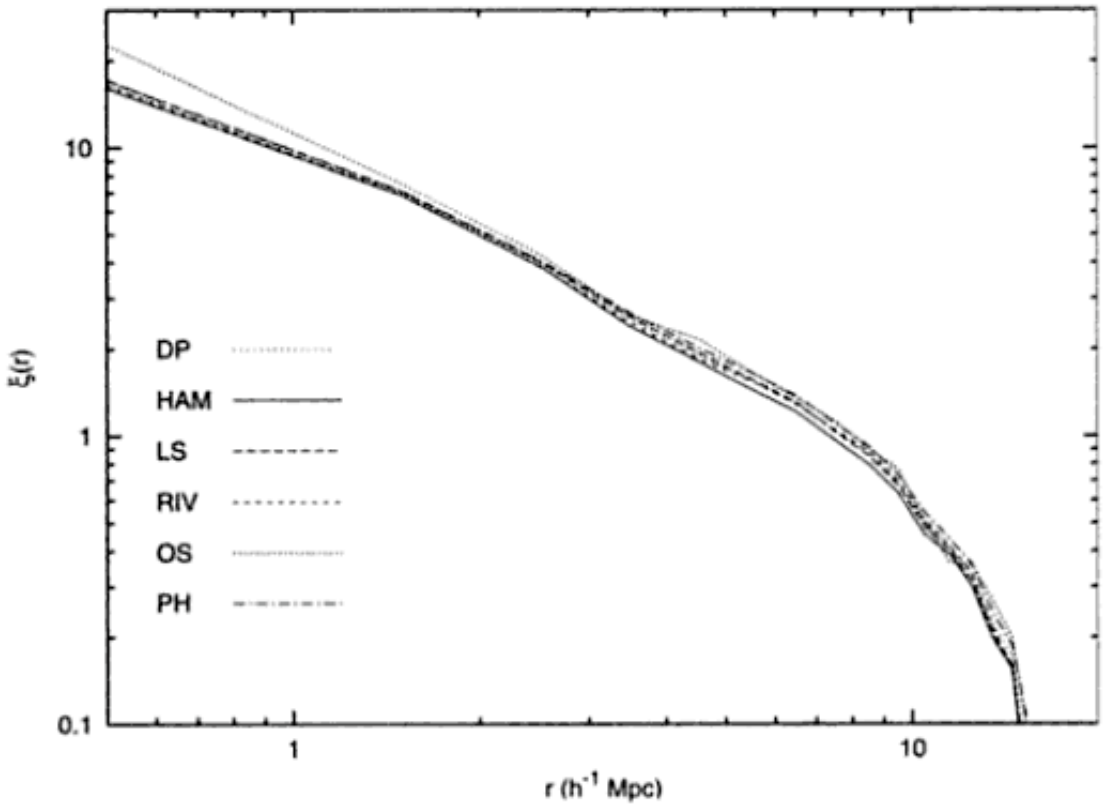
Zehavi et al. 2004

- The traditional technique is based on a Montecarlo method.
- The probability distribution of pair separations in the data is compared with that of a (computed generated) pseudo-random process having the same overall sky coverage and redshift distribution as the data.
- This requires detailed modelling of the observational sampling strategy.

Common estimators

- Peebles & Hauser (1974): $\hat{\xi} = \frac{DD}{RR} - 1$
- Davis & Peebles (1983): $\hat{\xi} = \frac{DD}{DR} - 1$
- Hamilton (1993): $\hat{\xi} = \frac{DD * RR}{DR^2} - 1$ (self-normalizing)
- Landy & Szalay (1993): $\hat{\xi} = \frac{DD - 2DR + RR}{RR}$
- All estimators are biased when applied to overdense or underdense samples (i.e. in almost all practical cases). The first two have a bias which depends linearly on the overdensity, the second two show a quadratic dependence and are therefore preferable.

A comparison of the estimators



Martinez & Saar 2002