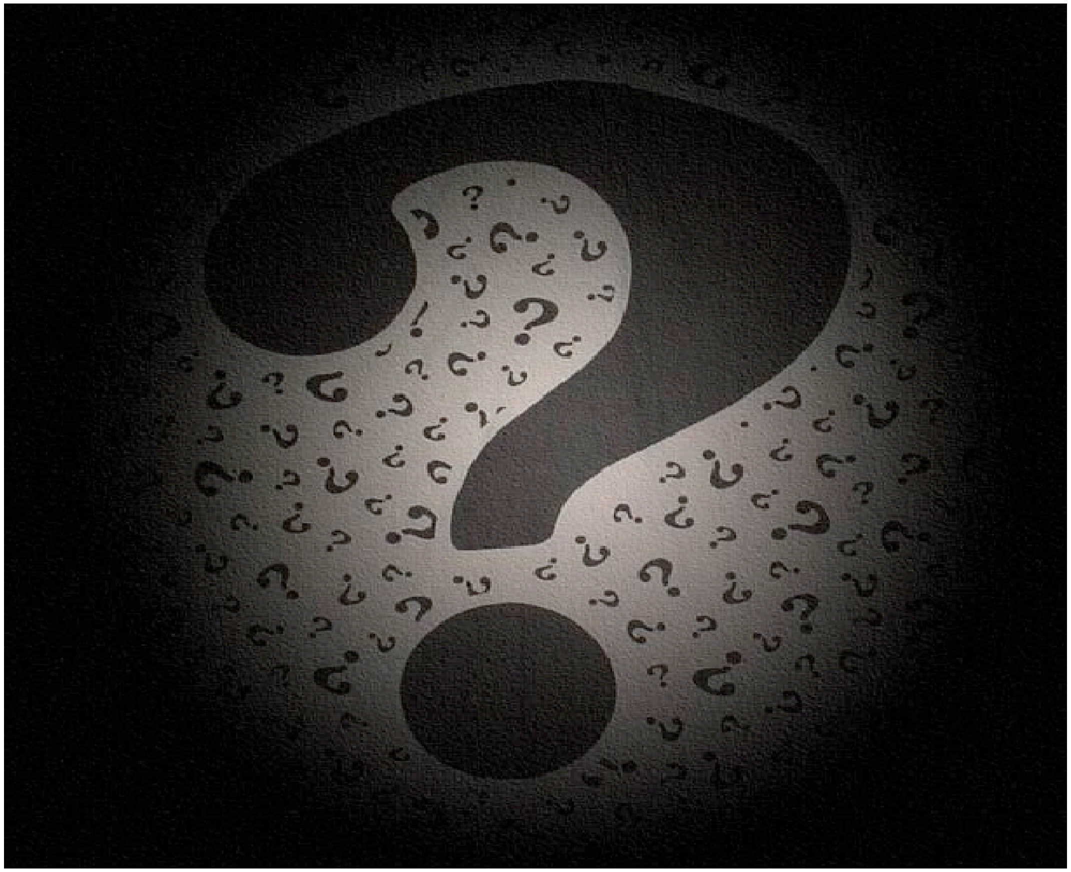
Statistical inference and resampling statistics

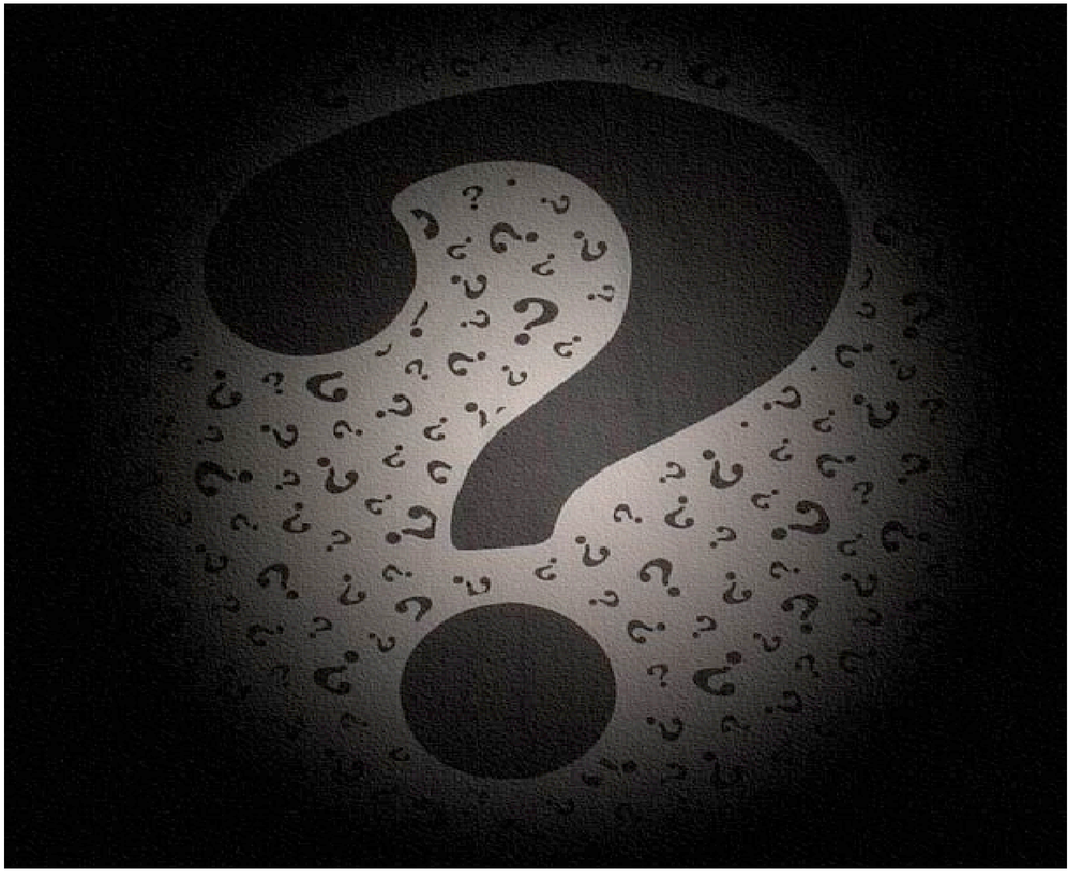
Cristiano Porciani

AIfA, Bonn

Questions



…and answers!



The grand challenge

- d=s+n
- s is a Gaussian variable with zero mean and known variance $\sigma^2{}_s$
- n is a Gaussian variable with zero mean and known variance $\sigma^2{}_n$
- You measure d, what is your best Bayesian point estimate for s?
- What is the posterior PDF for s?

The taste test

- Suppose you work for a wine-producing company in Bertinoro

- You want to know if people can tell the difference between your latest (fancy) production and the past one

- Twenty-four tasters are given three identical glasses, one contains the new wine, the other two the older one.

- The tasters are told that one glass contains a different product and are trying to correctly identify it

- Eleven tasters give the correct answer. What should you conclude?

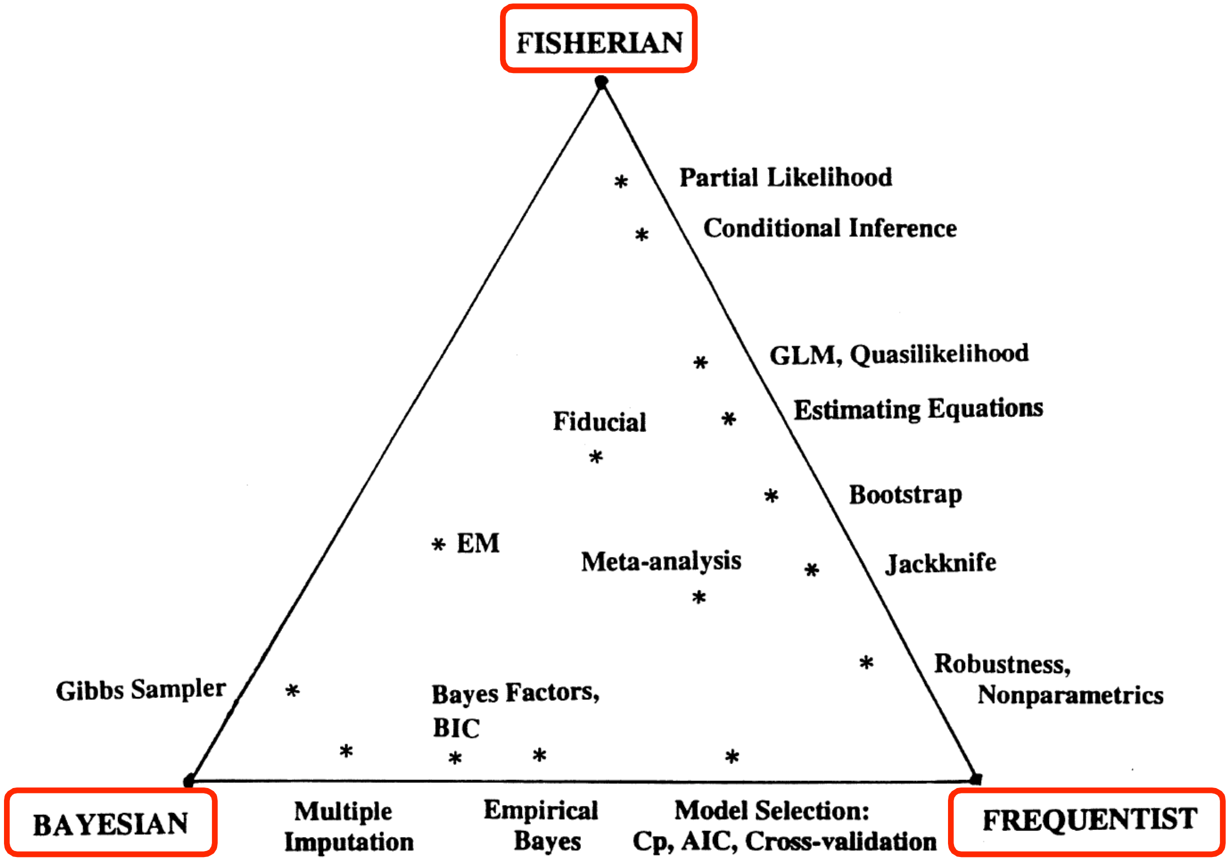
Statistical inference

- Estimation of population parameters
- Confidence intervals
- Hypothesis testing

The coordinates of statistics

Bradley Efron's triangle (1998)

Fisher



de Finetti, Savage
Jeffrey

Neyman, the Pearson's

Jerzy Neyman (1894-1981)

"Each morning before breakfast every single one of us approaches an urn filled with white and black balls. We draw a ball. If it is white, we survive the day. If it is black, we die. The proportion of black balls in the urn is not the same for each day, but grows as we become older….Still there are always some white balls present, and some of us continue to draw them day after day for years."

Hypothesis testing
(Neyman & Pearson 1933)

1. State the null hypothesis H_0 (usually, that the observations are due to pure chance). This hypothesis is tested against possible rejection under the assumption that it is true.

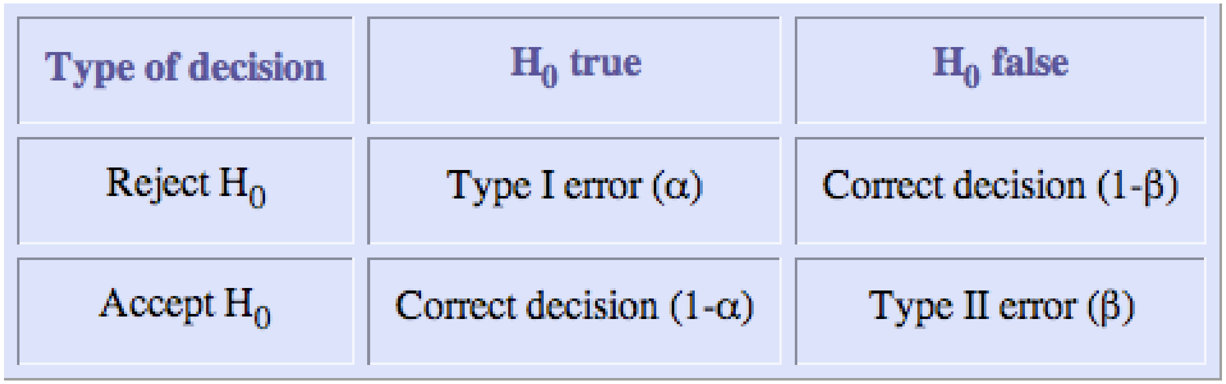
2. Formulate an alternate hypothesis H_A which is mutually exclusive with H_0 (usually, that the observation are due to a combination of a real effect and chance)

3. Identify a test statistic to assess the truth of the null hypothesis and evaluate the statistic using the sample data.

4. Assuming that the null hypothesis were true, compute the probability p that the test statistic assumes a value at least as significant as the one observed. This requires knowledge of the PDF of the statistic (the sampling distribution).

5. Draw a conclusion by comparing p with a significance value (or confidence level) $1-\alpha$ ($0\leq \alpha \leq 1$). If $p<1-\alpha$ the observed effect is statistically significant and the null hypothesis is rejected in favour of H_A. If $p> \alpha$ there is not enough evidence to reject H_0 .

Type I and II errors



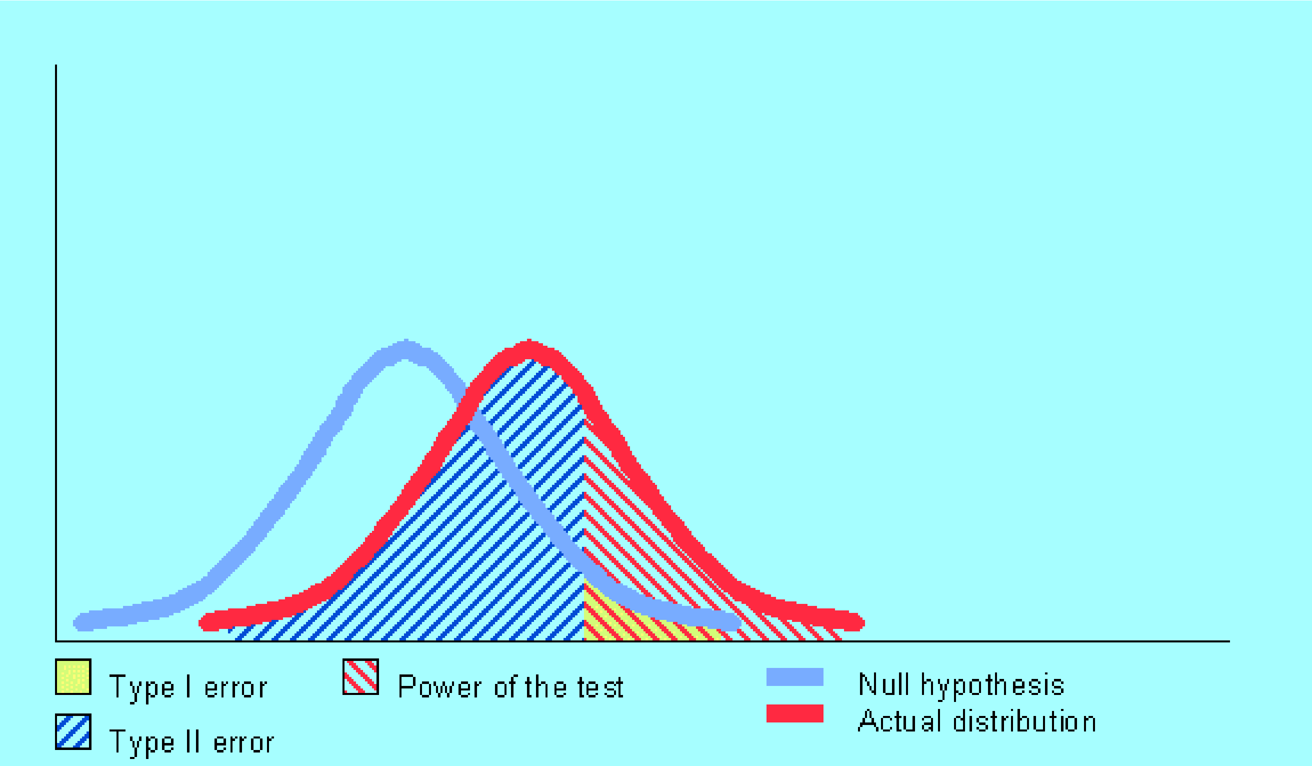
Type of decision	H_0 true	H_0 false
Reject H_0	Type I error (α)	Correct decision ($1-\beta$)
Accept H_0	Correct decision ($1-\alpha$)	Type II error (β)

- When using statistical tests there is always a chance of drawing wrong conclusions!

- Even for a confidence level of 95% there is a 5% chance of rejecting H_0 when it was actually correct. This is called type I error and its rate α is called the size of the test.

- It is also possible not to reject H_0 when it is actually incorrect. This is called type II error and its rate is indicated with the letter β. The quantity $1-\beta$ is commonly called the "power" of a test.

Type II errors



- There is little control on β because it also depends on the actual difference being evaluated which is usually unknown

- A type II error is a missed opportunity to reject H_0 correctly but it is not an error in the sense that an incorrect conclusion was drawn since NO CONCLUSION is drawn when H_0 is not rejected.

Example: the taste test

- 24 tasters are given 3 glasses, one of the three with a different wine. The tasters were attempting to correctly identify the one that was different.

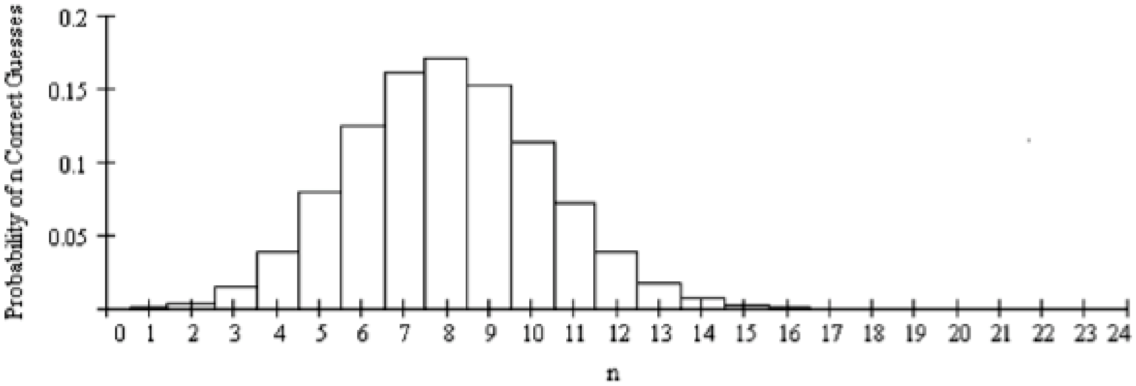
- H_0: f=1/3 against H_A: f>1/3 with f the fraction of successes

- Statistic: Y=number of successes by the tasters

- If H_0 is true (i.e. the different wine is guessed at random), the probability than Y tasters make the correct choice follows a binomial distribution with expectation value E(Y)=24/3=8.

- However, in a single experiment, values higher than 8 could happen by chance. What critical value Y_C should we choose to reject H_0 if Y> Y_C?

The taste test



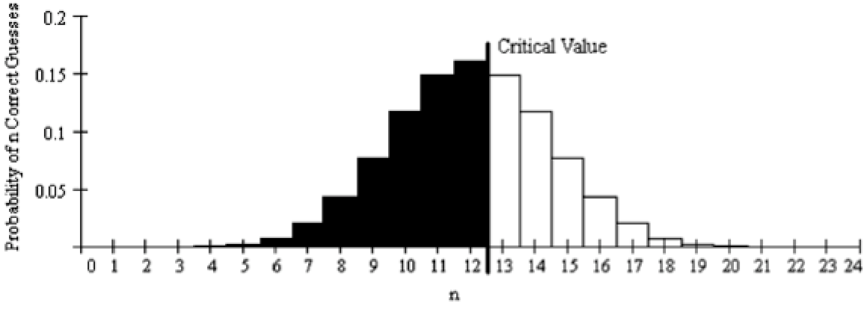
- We want to minimize Type I errors and choose α=0.05.

- Under the null hypothesis, P(Y>11)=0.068 and P(Y>12)=0.028, therefore we reject the null hypothesis if Y>12.

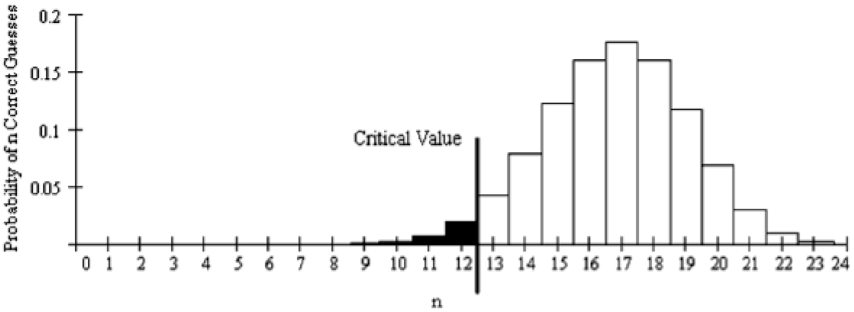
- The tasting survey found 11 correct choices and the company concluded that the results were not statistically significant as P(Y>10)=0.14.

The taste test

f=0.5 β=0.581



f=0.7 β=0.031



- What is the probability they made a type II error?

- Answer: P(Y<13 | f) which depends on the sample size, n, the type I error probability, α, and the population value of f (which is unknown)

The Neyman-Pearson criterion

- Type I errors are “false-alarm” errors
- Type II errors are missed detections
- Example, HIV testing. Type I errors would lead to treat someone who did not contract HIV infection. Type II errors would leave someone who is infected untreated (and unaware). What is worse? Would you choose a test with (α=0.05, β=0.02) or with (α=0.02, β=0.05)?
- The Neyman-Pearson criterion says that we should construct our decision rule to minimize β while not allowing α to exceed a certain value, i.e. the most powerful test of fixed size.

The Neyman-Pearson lemma

- Suppose we wish to test the simple null hypothesis H_0: $\theta=\theta_0$ (i.e. a point hypothesis where the parameters of a model assume a specific set of values) versus the simple alternative hypothesis H_A: $\theta=\theta_A$, based on a random sample $x_1,\ldots, x_n$ from a distribution with parameters θ.

- Let $L(x_1,\ldots, x_n|\theta)$ denote the likelihood of the sample when the value of the parameters is θ.

- Then the likelihood-ratio test which rejects H_0 when

$$\Lambda(x_1,\ldots,x_n) = \frac{L(x_1,\ldots,x_n \mid \theta_0)}{L(x_1,\ldots,x_n \mid \theta_A)} \leq \eta \quad where \quad P(\Lambda \leq \eta \mid H_0) = \alpha$$

is the most powerful test of size α.

The minimum-χ^2 fitting method

- If you have good reasons to think that your measurements contain random errors following a Gaussian distribution, then the likelihood function will be given by the product of Gaussian functions

- The log-likelihood is thus proportional to the sum of the squared residuals:

$$\chi^2 = \sum_i \frac{[d_i - f(x_i,\vartheta)]^2}{\sigma_i^2}$$

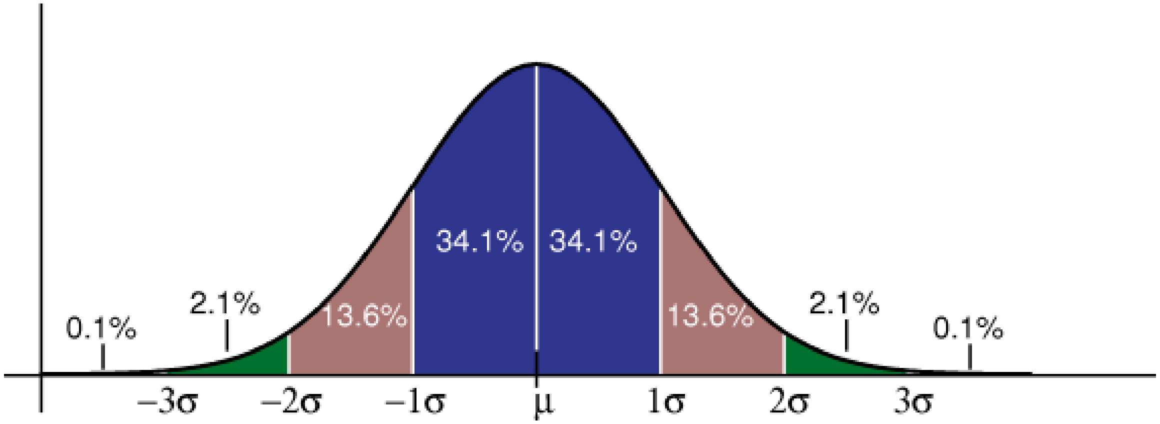
- The χ^2-fitting method consists of minimizing the χ^2 statistic by varying the model parameters.

- Following the Neyman-Pearson lemma, the confidence intervals for the model parameters will thus correspond to fixed values of $\Delta\chi^2 = \chi^2 - \chi^2_{min}$ (Avni 1976).

Values of $\Delta\chi^2$ corresponding to commonly used significance levels

Significance	1 parameter	2 parameters	3 parameters
68.3%	1.00	2.30	3.50
90%	2.71	4.61	6.25
99%	6.63	9.21	11.30

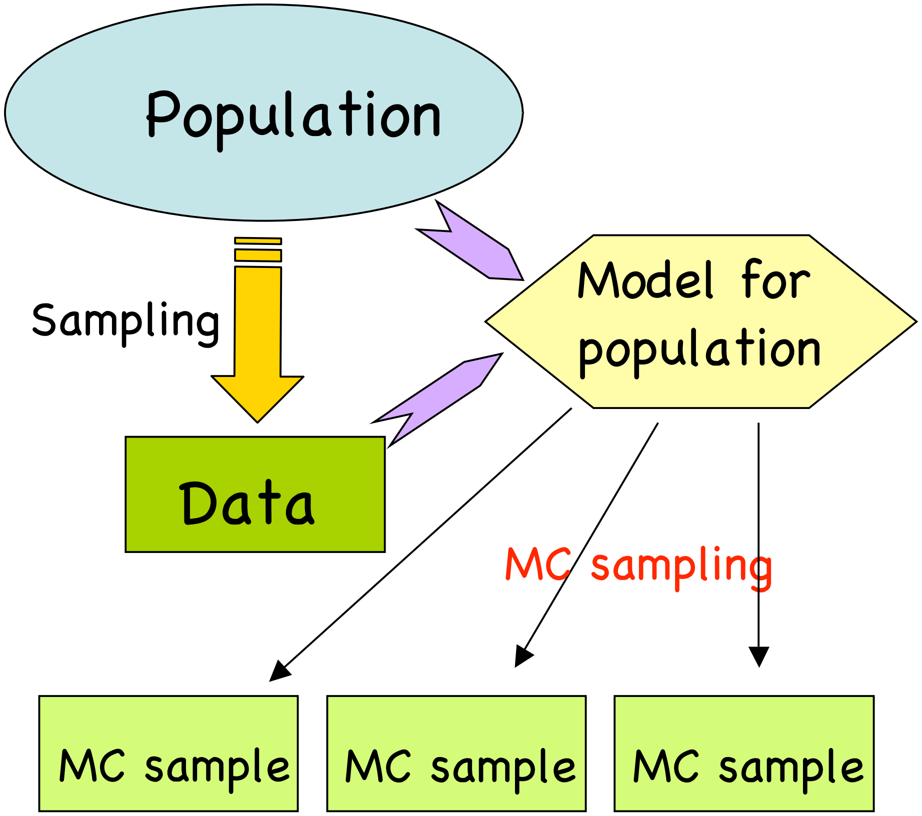
Conventional approach



- Assume (or derive) a theoretical distribution (e.g. Gaussian or Poisson) for the PDF of a statistic

- Compute confidence levels analytically or numerically

Montecarlo approach



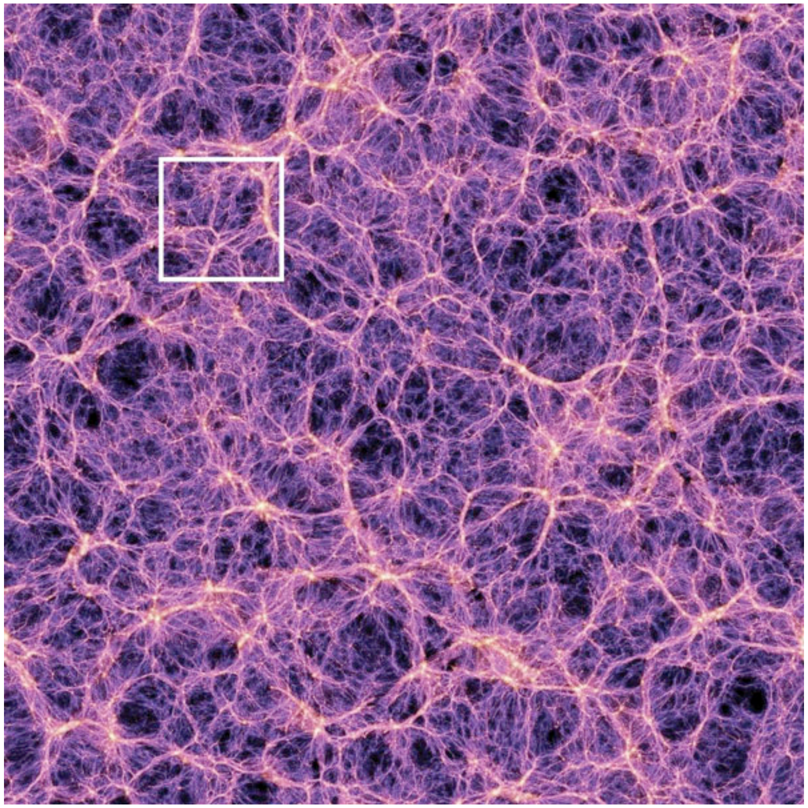
- Assume a theoretical distribution with some parameters motivated by the data

- Generate many samples extracted from the theoretical distribution

- Compute a statistic from the simulated data

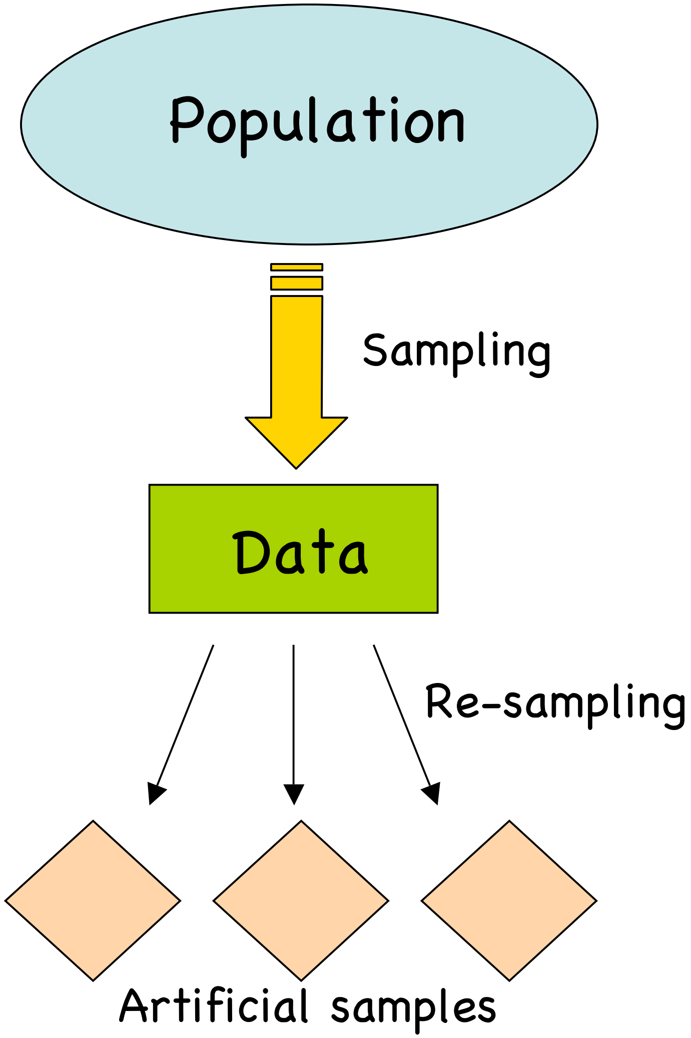
- Compare the PDF of the statistic with the actual observed data

Mock sample approach



- Simulate the process under study
- Extract statistical properties by considering many different realizations
- Popular in astronomy for complex quantities (e.g. the galaxy distribution) for which it is not easy to guess a theoretical model for the PDF
- Not reliable if mocks do not resemble reality

Resampling approach



- "Sampling within a sample" philosophy. It assumes that the data are representative of the population but no assumptions are made on the population distribution and parameters.

- The original data are used to build a large number of hypothetical samples from which one computes bias, standard errors and confidence intervals of the statistic of interest

- Resampling methods became very popular after the 1980s when fast and cheap computing resources started to be available

Jackknife resampling

(Quenouille 1949, Tukey 1958)

- The (delete-1) jackknife focuses on the samples that leave out one observation at a time

- The i^{th} jackknife sample consists of the original data with the i^{th} observation removed

- The i^{th} jackknife replica is the value of the statistic of interest evaluated using the corresponding jackknife sample

Jackknife resampling

- The jackknife estimate of the expectation value is obtained by averaging over all the replicas.

- The difference between this average and the observed value is the jackknife estimate of the bias.

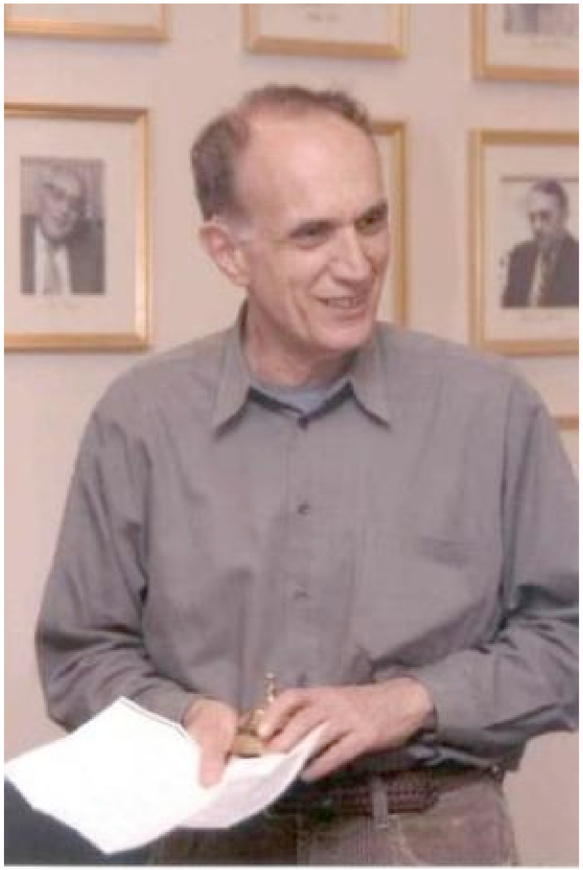
- The jackknife estimate of the variance is obtained by computing the sample variance of the N replicas and rescaling the result to account for the fact we removed only 1 object:

$$\hat{\sigma}^2_{jackknife} = \frac{N-1}{N} \sum_{i=1}^{N} (\hat{\vartheta}_i - \frac{1}{N} \sum_{j=1}^{N} \hat{\vartheta}_j)^2$$

Jackknife in practice

- Consider the sample (1, 2, 3, 4) and the sample-mean statistic (2.5)
- All jackknife samples are: (2, 3, 4), (1, 3, 4), (1, 2, 4) and (1, 2, 3)
- The corresponding jackknife replicas are: 3, 2.67, 2.33 and 2
- The mean of the jackknife replicas is 2.5 and coincides with the sample mean. This says that the sample mean is an unbiased estimator of the population mean.
- The standard deviation of jackknife replicas is 0.37 and, since we have 4 replicas, the jackknife estimate of the standard error is 0.65

Bradley Efron (b. 1938)



Bootstrap resampling

(Efron 1979)

“pull up by your own bootstrap”
i.e. “rely on your own resources”

- Suppose our dataset contains N iid measurements. A bootstrap sample is a collection of N measurements drawn with replacement from the original data

- The statistic of interest for a bootstrap sample is called a bootstrap replica

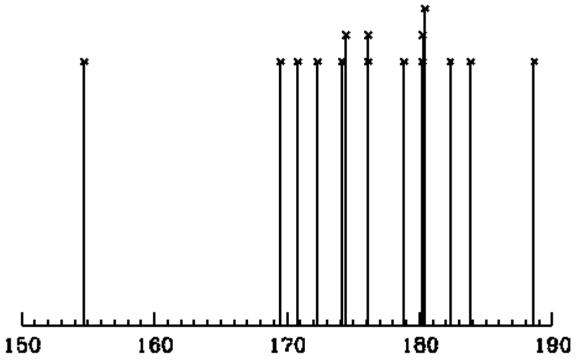
- The mean and the dispersion of the replicas thereby provides an estimate of bias and variability of the statistic

Bootstrapping in practice

- Consider the sample (1, 2, 3, 4) and the sample-mean statistic (2.5)
- Possible bootstrap samples are: (2, 1, 2, 3), (4, 3, 3, 3) and (1, 2, 1, 4)
- The corresponding bootstrap replicas are: 2, 3.25 and 2
- After generating very many bootstrap replicas, the difference between the value of the original statistic and the mean of the replicas will be the bootstrap estimate of the bias
- Similarly, the variance around the mean will be a measure of the dispersion

Comparative example: confidence interval for the median

- We want to study the height of adult males in Europe. We have a sample with 15 entries: 180.2, 174.1, 154.7, 170.8, 172.3, 169.5, 174.4, 180.3, 180.4, 182.3, 183.8, 176.1, 178.8, 176.1, 188.6 (all in cm)



- We wish to construct a 95% confidence level for the population median
- How would you proceed?

Classical approach (naïve)

- If we have good reasons to think that the data are iid and obtained by sampling a normal distribution, we could use a classical approach

- For a normal distribution, it can be demonstrated that the standard error of the median for large samples of size N asymptotically tends to (for $N \rightarrow \infty$)

$$\sigma_{med} = 1.253 \frac{\sigma}{\sqrt{N}}$$

- Using the sample standard deviation 7.60 as an estimate of σ and the sample median 176.10 as the central value, we obtain σ_{med}=2.46. Assuming that also the PDF of the median estimates is normal, we get that the 95% CL corresponds to 1.96 σ_{med} or:

  171.3 <median< 180.9

  (note that only 3 datapoints are smaller than 171.3 and only 3 datapoints are larger than 180.9)

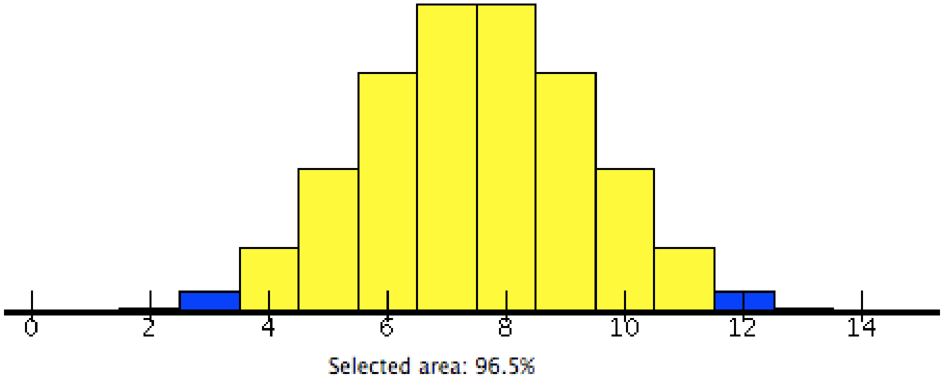
Classical approach (smart)

- This is an example of non-parametric inference, where we do not assume a model for the data (Gaussian or whatever) and we do not estimate any parameter (e.g. mean or variance) to characterize a model.

- The probability that a number drawn from the population is above the median is p=0.5, the probability that it is below is 1-p=0.5.

- The data are drawn from the population independently, so the number of data that are below the median has a binomial distribution with N=15 trials and p=0.5

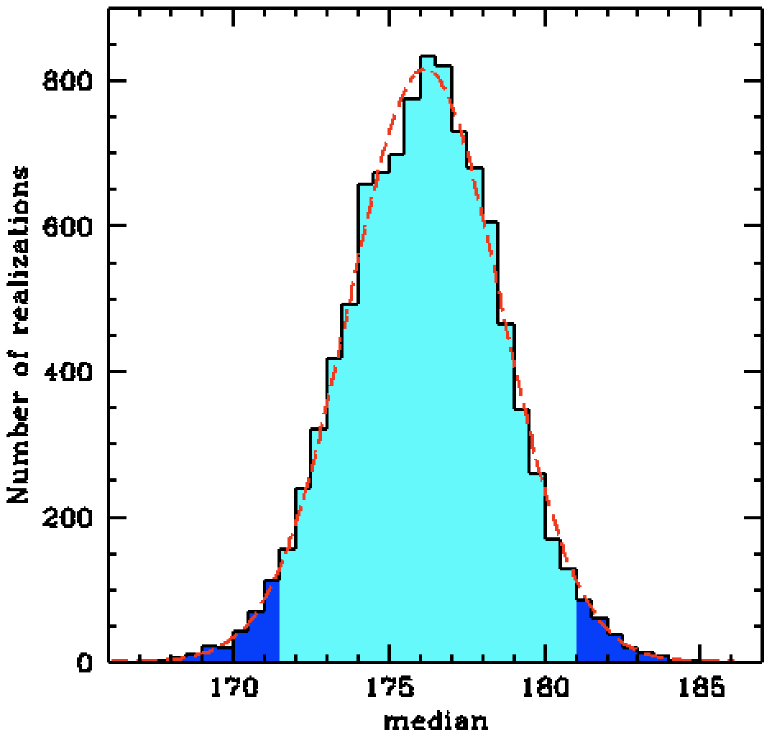
...segue...



- This binomial distribution has 88.2% probability that there are 5 to 10 datapoints below the true median and 96.5% that there are 4 to 11 datapoints below the true median.

- Therefore, the 95% CL interval is: 172.3< median<180.4 (i.e. from the 4th smallest measurement to the 4th largest)

Montecarlo



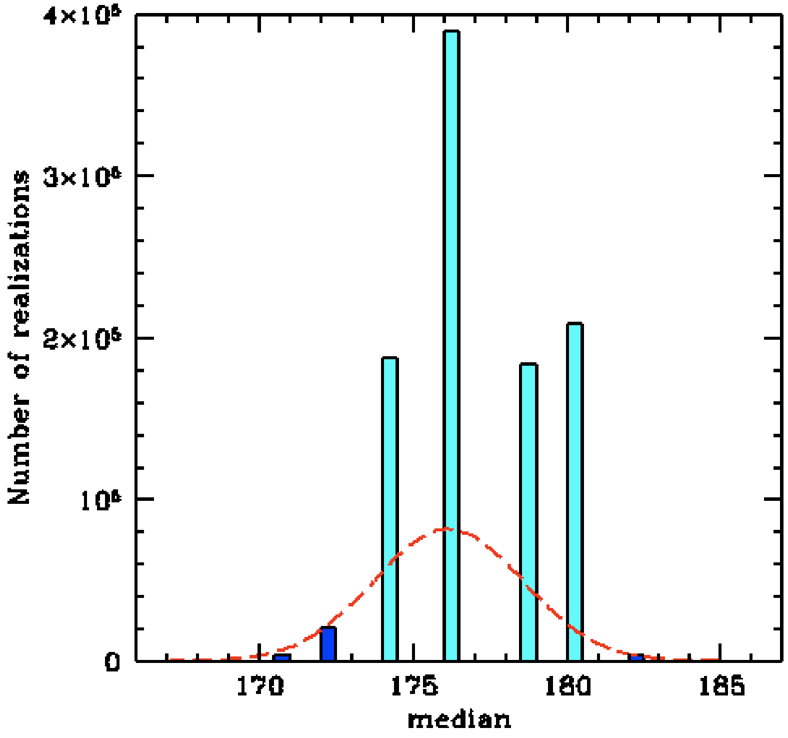
- 10,000 Montecarlo samples with N=15 drawn from a Gaussian distribution with μ= sample mean and σ^2 = the sample variance (plug-in principle)

- The empirical PDF is shown on the right together with a Gaussian with the same mean and variance

- The interval corresponding to the 95% confidence level is

  171.3< median<180.9

Bootstrap



- Bootstrap resampling of the 15 measurements

- Does not assume anything about the population

- PDF is spiky because only one of the measurements can be the median

- 95% CL interval:

10^4 res. 174.1<median<180.4

10^5 res. 172.3<median<180.4

10^6 res. 172.3<median<180.4

Jackknife

- Delete-1 jackknife fails!

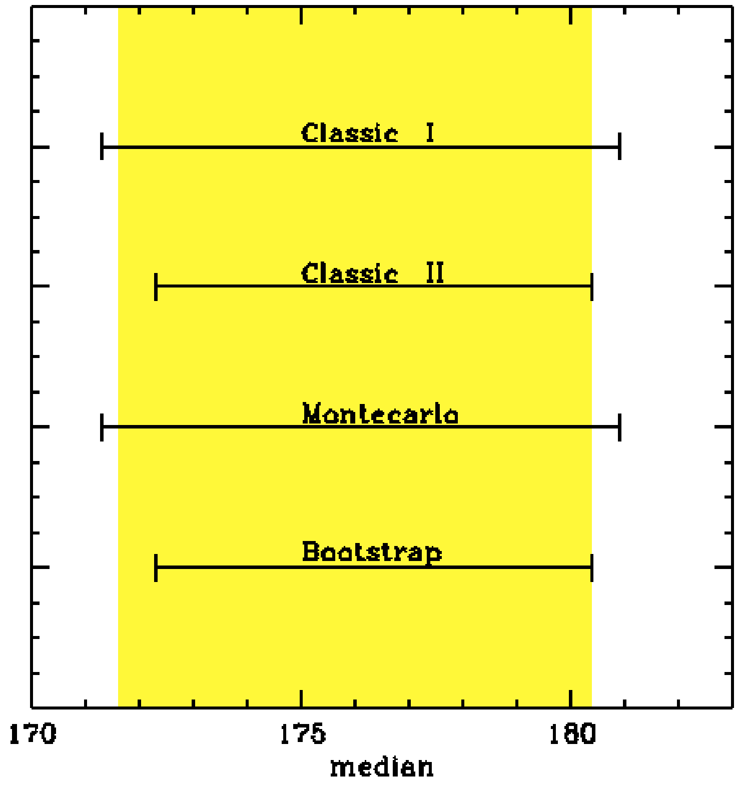
- Only 15 resamples are possible and 13 of them give exactly the same value: 175.2. The remaining two give 179.2 and 180.2.

- It can be shown that jackknife estimates of the median are inconsistent. This happens because the median is not a smooth statistic (i.e. it can jump when small changes are made to the data)

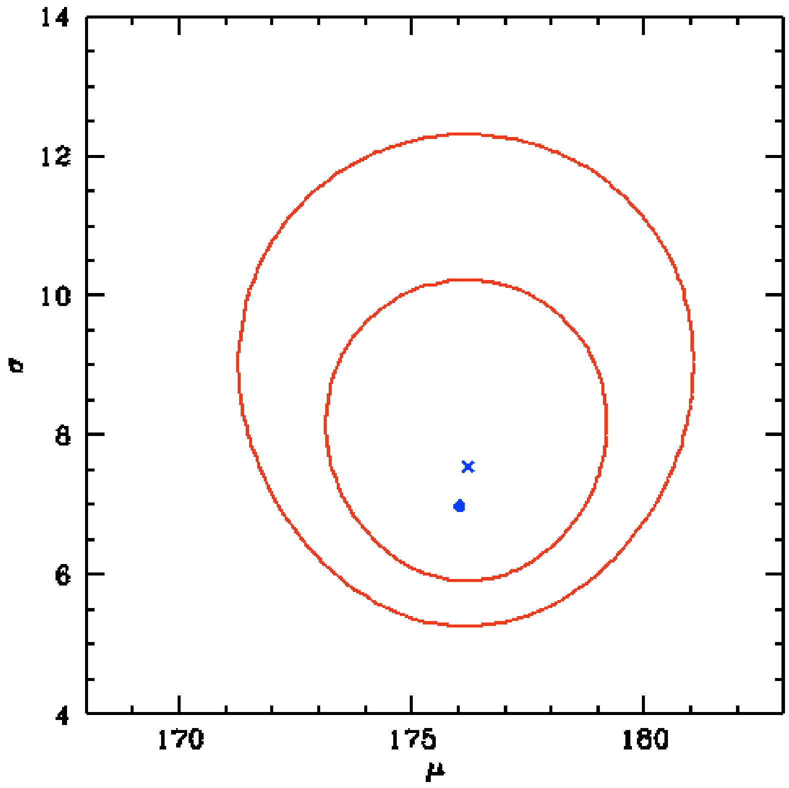
- More sophisticated versions of the jackknife where groups of d observations are removed (delete-d jackknife) are better suited for this application

Summary

- Classic I: 171.3<median<180.9
- Classic II: 172.3<median<180.4
- Montecarlo: 171.3<median<180.9
- Bootstrap: 172.3<median<180.4
- Correct: 171.6<median<180.4



Using the likelihood function



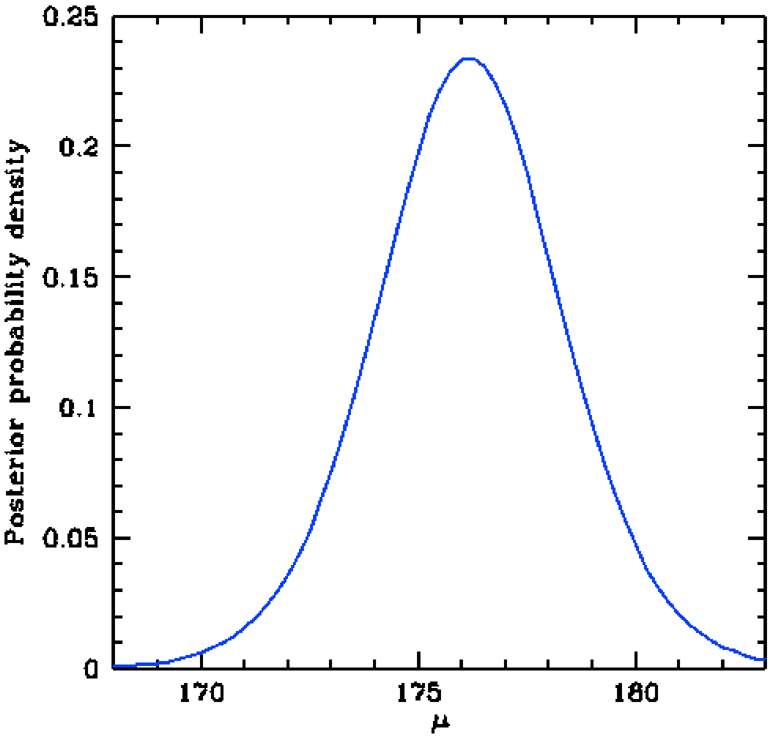
- We can easily build the likelihood function under the assumption that our data come from a Gaussian distribution (see the figure on the left)

- The blue dot indicates the model parameters that maximize the likelihood function

- The cross indicates the "population" parameters from which I generated the data with a Monte Carlo technique

- Note that the contours progressively depart from ellipses as you go away from the peak

The Bayesian way



- Assuming a flat prior in the μ-σ plane, we can build the posterior probability (the contours in the previous figure actually contain 68.3% and 95.4% of the posterior probability)

- Marginalizing over σ (i.e. integrating the posterior along σ at fixed μ), we obtain the posterior for μ

- The central credibility interval for μ (at 95% significance) is then: $171.6<\mu<180.4$

Confidence vs. credibility intervals

- Confidence intervals (Frequentist): measure the variability due to sampling from a fixed distribution with the TRUE parameter values. If I repeat the experiment many times, what is the range within which 95% of the results will lie?

- Credibility interval (Bayesian): For a given significance level, what is the range I believe the parameters of a model can assume given the data we have measured?

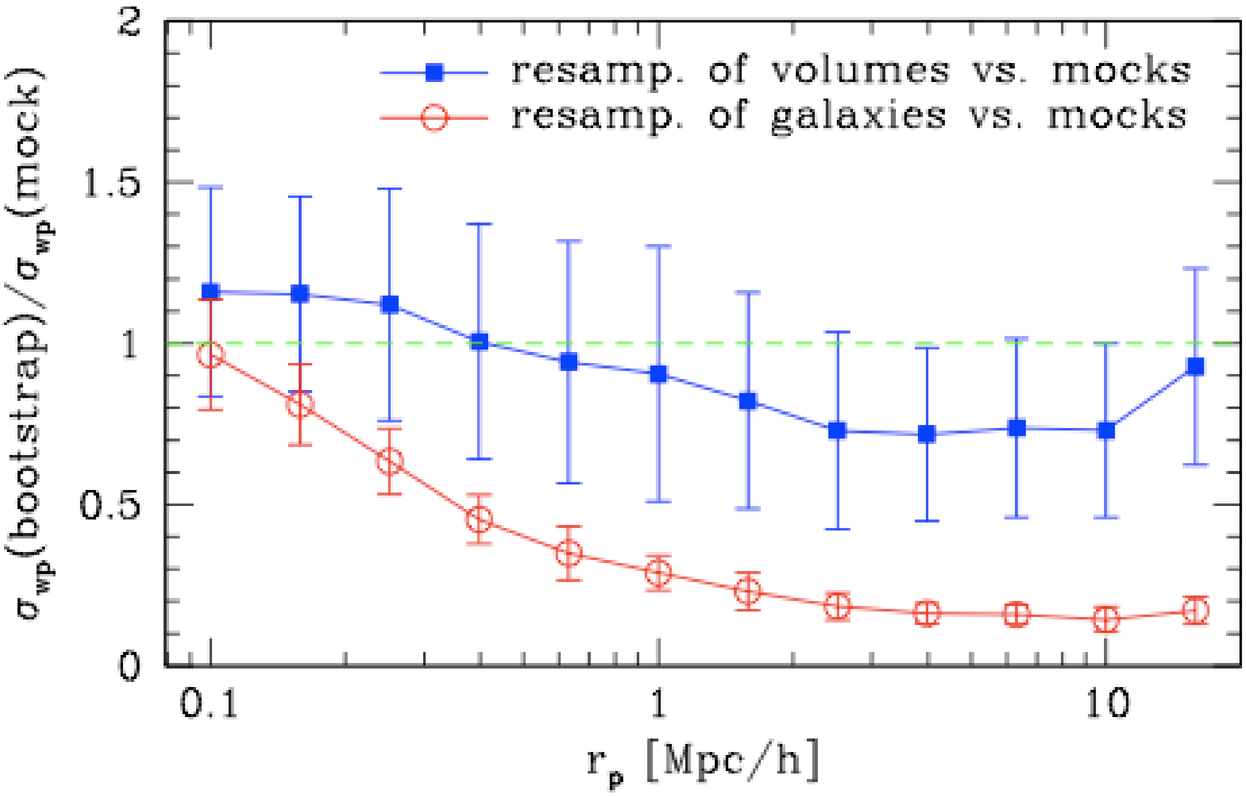
- They are profoundly DIFFERENT things even though they are often confused. Sometimes practitioners tend use the term “confidence intervals” in all cases and this is ok because they understand what they mean but this might be confusing for the less experienced readers of their papers. PAY ATTENTION!

Resampling methods in astronomy

- Image fidelity assessment in radioastronomy
- Significance of point-source (or spectral-line) detection and secular time variability in gamma-ray astronomy
- Errorbars for the Hubble diagram from supernovae Ia
- Errorbars for estimates of the galaxy luminosity-function and 2-point correlation function
- Loads of minor applications to compute errors in intermediate steps

The pitfalls of bootstrap

Meneux et al. 2009



Make sure your data are nearly iid before you start resampling them!