Classical Statistical Inference

Mariano Dominguez

August 23, 2015

***Statistical inference is so pervasive throughout the astronomical and astrophysical investigations that we are hardly aware of its ubiquitous role. It arises when the astronomer: + smooths over discrete observations to understand the underlying continuous phenomenon + seeks to quantify relationships between observed properties + tests whether an observation agrees with an assumed astrophysical theory + divides a sample into subsamples with distinct properties + tries to compensate for flux limits and nondetections + investigates the temporal behavior of variable sources + infers the evolution of cosmic bodies from studies of objects at different stages + characterizes and models patterns in wavelength, images or space

- Statistical inference helps in making judgments regarding the likelihood that a hypothesized effect in data arises by chance or represents a real effect. It is particularly designed to draw conclusions about the underlying population when the observed samples are subject to uncertainties.
- Two main aspects of inference are **estimation** and the **testing of hypotheses**. Regression, goodness-of-fit, classification and many other statistical procedures fall under its framework.
- Statistical inference can be parametric, nonparametric and semip.

Parametric inference requires that the scientist makes some assumptions regarding the mathematical structure of the underlying population, and this structure has parameters to be estimated from the data at hand. Linear regression is an example.

Nonparametric procedures make no assumption about the model structure or the distribution of the population. The KS hypothesis test and the Kendall's τ correlation coefficient are examples.

Point estimation

- If the shape of the probability distribution, or relationship between variables, of the underlying population is well-understood, then it remains to find the parameters of the distribution or relationship.
- Typically a probability distribution or relationship is characterized by a p-dimensional vector of model parameters $\theta = (\theta_1, \theta_2, ..., \theta_3)$. For example:
- the model of a planet in a Keplerian orbit around a star has a vector of six parameters: semi-major axis, eccentricity, inclination, ascending node longitude, argument of periastron and true anomaly.
- for a tidally truncated isothermal sphere of stars (King model).
- a turbulent viscous accretion disk (Shakura–Sunyaev α-disk model).
- the consensus model of cosmology with dark matter and dark energy (ΛCDM model).

Principles of point estimation

- In parametric point estimation two decisions must be made:
- First, the functional model and its parameters must be specified (**model misspecification**). Statistical procedures are available to assist the scientist in **model validation** (or goodness-of-fit) and **model selection**.
- Second, the method by which best-fit parameters are estimated must be chosen. The **method of moments**, **least squares** (LS) and **maximum likelihood estimation** (MLE) are important and commonly used procedures.Second, the method by which best-fit parameters are estimated must be chosen.
- In classical parametric estimation, the observations are assumed to be independently and identically distributed (i.i.d.) random variables with known probability distributions.

- The dataset $x_1, x_2, ..., x_n$ is assumed to be a realization of independent random variables $X_1, X_2, ..., X_n$ having a common **probability distribution function (p.d.f.) f**.
- We now consider distribution functions characterized by a small number of parameters, $\theta = (\theta_1, \theta_2, ..., \theta_p)$. The point estimator of the vector of true parameter values θ is designated $\hat{\theta}$. The estimator $\hat{\theta}$ of θ is a function of the random variables X_i.
- A great deal of mathematics and discussion lies behind the simple goal of obtaining the "best" estimates of the parameters θ.

excerp from History Channel.

- During the 19th century, the **method of moments** and **method of least squares** were developed.
- In the 1910s and 1920s, R. A. Fisher formulated the "likelihood" that a dataset fits a model, and inaugurated the powerful methods of **maximum likelihood estimation** (MLE).
- As computers became more capable, numerically intensive methods with fewer limitations than previous methods became feasible. The most important, developed in the 1970s and 1980s, is the **bootstrap method**.
- With the advent of numerical methods like Markov chain Monte Carlo simulations, nontrivial **Bayesian inferential** computations became feasible during the 1990s. Bayesian computational methods are being actively developed today.

Several important criteria of a point estimator:

- **Unbiasedness** The bias of an estimatoris defined to be: the difference between the mean of estimated parameter and its true value, $B(\hat{\theta}) = E[\hat{\theta}] - \theta$. This is an intrinsic offset in the estimator.
- Heuristically, $\hat{\theta}$ is an unbiased if its long-term average value is equal to θ. If $\hat{\theta}$ is an unbiased estimator of θ, then the variance of the estimator $\hat{\theta}$ is given by $E[(\hat{\theta} - \theta)^2]$.
- The smaller the variance of the estimator, the better the estimation procedure. However, if the estimator $\hat{\theta}$ is biased, then $E[(\hat{\theta} - \theta)^2]$ is not the variance of $\hat{\theta}$. In this case, $E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + (E[(\hat{\theta} - \theta)])^2$ or MSE $=$ Variance of $\hat{\theta}$ $+ (Bias)^2$.

This quantity, the sum of the variance and the square of the bias, is called the mean square error (MSE) and is very important in evaluating estimated parameters. Minimum variance unbiased estimator (MVUE) Among a collection of unbiased estimators, the most desirable one has the smallest variance, Var($\hat{\theta}$).

- Consistency This criterion states that a consistent estimator will approach the true population parameter value as the sample size increases.
- Asymptotic normality This criterion requires that an ensemble of consistent estimators $\hat{\theta}$ has a distribution around the true population value θ that approaches a normal (Gaussian) distribution with variance decreasing as $1/n$.

Techniques of point estimation

- Parameter estimation is motivated by the problem of fitting models from probability distributions or astrophysical theory to data. Many commonly used probability distributions depend only on a few parameters. Once these parameters are known, the corresponding properties of the underlying population, are completely determined.

We will illustrate the common methods of estimation using the two parameters of a population that satisfies a normal (Gaussian) density, the mean μ and standard deviation σ.

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i; \;\; \hat{\sigma}^2 = S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{X})^2$$

are estimators of μ and σ^2, respectively. The factor $n-1$ instead of n in the denominator of the estimator of σ^2 is required for unbiasedness.

Method of moments

- The moments are quantitative measures of the parameters of a distribution: the first moment describes its central location; the second moment its width; and the third and higher moments describe asymmetries.

The k-th moment of a random variable X with distribution function F is given by

$$\mu_k(X) = E[X^k] = \int x^k \mathrm{d}F(x)$$

For the random sample X_i, the k-th sample moment is

$$\hat{\mu_k} = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

Various parameters of a distribution can be estimated by the method of moments if one can first express the parameters as simple functions of the first few moments.

Method of least squares

Parameter estimation using least squares was developed in the early nineteenth century to solve problems in celestial mechanics, and has since been very widely used in astronomy and other fields.

Consider estimation of the population mean μ. The least-squares estimator $\hat{\mu}$ is obtained by minimizing the sum of the squares of the differences $(X_i - \mu)$,

In a simple case, $\hat{\mu} = \hat{X}$ which is the intuitive solution. But in more complex estimation problems, particularly in the context of regression with a functional relationship between two or more variables, this method provides solutions that are not intuitively obvious.

Likelihood methods

Likelihood is the hypothetical probability that a past event would yield a specific outcome.

The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes.

MLE is an enormously popular statistical method for fitting a mathematical model to data.

Developed in the 1920s by R. A. Fisher, MLE is a conceptual alternative to the least-squares method of the early nineteenth century, but is equivalent to least squares under Gaussian assumptions.

The Cramer–Rao inequality, which sets a lower bound on the variance of a parameter, is an important result of MLE theory.

While the likelihood function can be maximized using a variety of computational procedures, the EM algorithm developed in the 1970s is particularly effective.

Maximum likelihood method

The method is based on the **likelihood**, the probability density function viewed as a function of the data given particular values of the model parameters.

Here we use the notation $f(;\theta)$ for a probability density with parameter θ. For example, for an exponential random variable, is given by $f(x;\theta) = \theta exp(-\theta x)$ for $x > 0$, and $f(x;\theta) = 0$ for $x \leq 0$.

For i.i.d. random variables $X_1, X_2, ..., X_n$ with a common density function $f(;\theta)$, the likelihood L and log likelihood l are given by:

$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$ $l = \ln(L(\theta)) = \ln(f(X_i;\theta))$

The likelihood at parameter θ is thus the product of densities evaluated at all the random variables.

MLEs have many strong mathematical properties.

For most probability structures considered in astronomy, the MLE exists and is unique.

A crucial property is that, for many commonly occurring situations, maximum likelihood parameter estimators $\hat{\theta}$ have an approximate normal distribution when n is large.

In most cases, the MLE estimator satisfies: $Var(\hat{\theta}) \sim \frac{1}{I(\theta)}$ where $I(\theta) = nE(\frac{\partial \log(f(X;\theta))}{\partial x})^2$

$I(\theta)$ is called the **Fisher information**. When θ is a vector of parameters, this is the **Fisher information matrix** with off-diagonal terms of the form $\frac{\partial f}{\partial \theta_i \partial \theta_j}$.

Confidence intervals

Point estimates cannot be perfectly accurate as different datasets drawn from the same population will give rise to different inferred parameter values. To account for this, we estimate a range of values for the unknown parameter that is usually consistent with the data.

This is the parameter's confidence interval or confidence set around the best-fit value. The confidence intervals will vary from sample to sample.

Confidence intervals can be estimated for different methods of point estimation including least squares and MLE.

The **confidence interval of a parameter** θ, a statistic derived from a dataset X , is defined by the range of lower and upper values $[l(X), u(X)]$ that depend on the variable(s) X defined such that $P[l(X) < \theta < u(X)] = 1 - \alpha$, where $0 < \alpha < 1$ is usually a small value like $\alpha = 0.05$ or 0.01.

The quality of confidence intervals

is judged using criteria including validity of the coverage probability, optimality (the smallest interval possible for the sample size), and invariance with respect to variable transformations. For $\alpha = 0.05$, the estimated 95% confidence interval of an estimator of some parameter θ is an interval (l, u) such that $P(l < \hat{\theta} < u) = 0.95$.

If there are two or more unbiased estimators, the one with the smaller variance is often preferred. Under some regularity conditions, the Cramer–Rao inequality gives a lower bound on the minimum possible variance for an unbiased estimator. It states that if $\hat{\theta}$ is an unbiased estimator based on i.i.d. random variables $X_1, X_2, ..., X_n$ with a common density function $f(;\theta)$ where θ is a parameter, then the smallest possible value that $Var(\theta)$ can attain is $\frac{1}{I(\theta)}$ where I is the Fisher information (see TNMM page 125).

Calculating MLEs

Likelihoods can be maximized by any numerical optimization method. During the midtwentieth century before computers, simple analytical statistical models were emphasized where the maximum of the likelihood function could be obtained by differential calculus.

For a model with p parameters $\theta_1, \theta_2, .., \theta_p$, the equations: $\frac{\delta L(\theta)}{\delta \theta_i} = 0$ often gave a system of equations that could be solved using algebra.

Alternatively, the maximum of the likelihood could be found numerically using iterative numerical techniques like the Newton–Raphson and gradient descent methods and their modern variants (e.g. Levenberg–Marquardt see NR). These techniques may converge slowly, encountering problems when the derivative is locally unstable and when the likelihood function has multiple maxima.

Other techniques, such as simulated annealing and genetic algorithms, are designed to assist in finding the global maximum in likelihood functions with complex structure.

The Expectation Maximization algorithm

One particular numerical procedure, the EMalgorithm, has been enormously influential in promoting maximum likelihood estimation since the seminal papers of Dempster et al. (1977) and Wu (1983).

The method was independently developed in astronomy for image deconvolution by Richardson (1972) and Lucy (1974). Here the data are the observed image of the sky blurred by the telescope's optics, the model is the telescope point spread function, and the missing dataset is the true sky image.

The EM algorithm considers the mapping of a set of datasets to an unknown complete dataset. One begins the EM algorithm with initial values of the model parameter values θ and the dataset. These might be estimated by least squares, or represent guesses by the scientist.

The algorithm proceeds by iteration of two steps. The **expectation step** (E) calculates the likelihood for the current values of the parameter vector θ.

The **maximization step** (M) updates the missing dataset values with the criterion that the likelihood of the values with respect to the current model is maximized. This updated dataset then takes the place of the original dataset, and the algorithm is iterated until convergence.

The algorithm is successful for many MLE problems because each iteration is guaranteed to increase the likelihood over the previous iteration. Local minima are ignored and convergence is usually rapid. However, there is still no guarantee that the achieved maximum is global over the full parameter space.

Hypothesis testing techniques

As the name implies, the goal here is not to estimate parameters of a function based on the data, but to test whether a dataset is consistent with a stated hypothesis.

The scientist formulates a **null hypothesis** and an **alternative hypothesis**. The result of the test is to either reject or not reject the null hypothesis at a chosen significance level. Note that failure to reject the null hypothesis does not mean that the null hypothesis is correct.

Statistical testing of a hypothesis leads to two types of error: wrongly rejecting the null hypothesis (**Type 1 errors or false positives**) and wrongly failing to reject the null hypothesis (**Type 2 errors or false negatives**). It is impossible to bring these two errors simultaneously to negligible values.

Classical hypothesis testing is not symmetric; interchanging the null and alternative hypotheses gives different results. An important astronomical application is the detection of weak signals in noise.

A traditional choice is to construct the critical regions to keep Type 1 errors under control at the 5 % level, allowing Type 2 errors to be uncontrolled. This choice of 5 % is called **the significance level of the hypothesis test**, and represents the probability of generating false positives; that is, incorrectly rejecting the null hypothesis.

A result of a hypothesis test is called **statistically significant** if it is unlikely to have occurred by chance. That is, the hypothesis is significant at level α if the test rejects the null hypothesis at the prescribed significance level α. Typical significance levels used in many fields are $\alpha = 0.05$ or 0.01. Note that the common standard in astronomy of 3σ, corresponding to $\alpha = 0.003$ for the normal distribution.

A common difficulty is that significance levels must be adjusted when many hypothesis tests are conducted on the same dataset. This situation often occurs in astronomical image analysis. A large image or data cube may be searched at millions of locations for faint sources, so that one must seek a balance between many false positives and sensitivity.

A new procedure for combining multiple hypothesis tests called the false detection rate provides a valuable way to control for false positives (Benjamini & Hochberg 1995).

Resampling methods 2

Understanding the variability of a point estimation is essential to obtaining a confidence interval, or to assess the accuracy of an estimator.

In many situations encountered in the physical sciences, the variance may not have a closed-form expression.

Resampling methods developed in the 1970s and 1980s come to the rescue in such cases. Powerful theorems demonstrated that they provide inference on a wide range of statistics under very general conditions.

Boostrap

Methods such as the **bootstrap** involved constructing hypothetical populations from the observations, each of which can be analyzed in the same way to see how the statistics of interest depend on plausible random variations in the observations.

Resampling the original data preserves whatever structures are truly present in the underlying population, including non-Gaussianity and multimodality. Although they typically involve random numbers, resampling is not an arbitrary Monte Carlo simulation; it is simulation from the observed data.

The **half-sample** method may be the oldest resampling method, where one repeatedly chooses at random half of the data points, and estimates the statistic for each resample. The inference on the parameter can be based on the histogram of the resampled statistics. It was used by P. C. Mahalanobis in 1946 under the name **interpenetrating samples**.

The most important of the resampling methods proved to be the bootstrap or **resampling with replacement** B. Efron (1979). Here

Jacknife

The jackknife method was introduced by M. Quenouille (1949) to estimate the bias of an estimator. The method was later shown to be useful in reducing the bias as well as in estimating the variance of an estimator.

Let $\hat{\theta}$ be an estimator of θ based on n i.i.d. random vectors $X_1, ..., X_n$. That is, $\hat{\theta}_n = f_n(X_1, ..., X_n)$, for some function f_n. Let be the corresponding recomputed statistic based on all but the i-th observation. The jackknife estimator of bias $E(\hat{\theta}_n) - \theta$ is given by $bias_j = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta_{n,-i}} - \hat{\theta}_n)$.

The jackknife estimator θ_J of θ is given by
$\theta_J = \hat{\theta}_n - bias_j = \frac{1}{n} \sum_{i=1}^{n} (n\hat{\theta}_n - n(n-1)\hat{\theta_{n,-i}})$

Model Selection (next class) and Bayesian SI.