Principles of statistical inference

Eric Feigelson

2nd East Asian Workshops on Astrostatistics
Nanjing & Guiyang
July 2018

Outline

- Statistical inference: Definition and use in astronomy
- Nonparametric inference
- Parametric inference (point estimation)
- Methods for point estimation:
  - Least squares estimation
  - Maximum likelihood estimation
  - Bayesian inference (tomorrow)

Statistical Inference

“Statistical inference is the process of drawing conclusions from data that is subject to random variation, for example, observational errors or sampling variation.”

(Wikipedia, 2014)

~~“The outcome of statistical inference may be an answer to the question "what should be done next?", where this might be a decision about making further experiments or surveys, or about drawing a conclusion before implementing some organizational or governmental policy.”(Wikipedia, 2013)~~

Why do astronomers need statistical inference?

1. smooth over discrete observations to estimate the underlying continuous phenomenon **density estimation**
2. quantify relationships between observed properties **regression**
3. explore a multivariate dataset to find relationships among variables **multivariate analysis**
4. divide a sample into subsamples with distinct properties **clustering & classification**
5. try to compensate for flux limits and nondetections **survival analysis**
6. investigate temporal behaviors of variable sources **time series analysis**
7. characterize and model patterns in wavelength, images or space **spatial processes, image analysis**
8. test whether an observation agrees with an astrophysical theory **regression**

Statistical inference is the process of learning something about the populations and processes that underlie a dataset. It is so pervasive throughout astronomy that we are hardly aware of its ubiquitous role.

Nonparametric inference: Motivation

Most standard statistical procedures (e.g. least squares or Bayesian model fitting) are ***parametric*** assuming:

- the underlying dataset and population is homogeneous (no outliers or mixtures)
- the assumed model family correctly and completely explain the physical phenomenon (model specification)
- the accuracy of measurement is invariant across the sample (homoscedasticity)
- the accuracy of measurement improves as sqrt-N (Central Limit Theorem)
- sample residuals from the correct model of the data exhibit a Gaussian distribution (normality).

When any of these assumptions does not hold nonparametric (or more sophisticated parametric) methods are desirable

Many of these nonparametric procedures are hypothesis tests, giving probabilities associated with Yes/No questions

Astronomers should more often ask questions that can be addressed nonparametrically (MSMA Chpt 5)

- Is this sample compatible with that sample? **2-sample test**
- Is this sample compatible with that model? **Goodness-of-fit test**
- Is there a correlation between these variables?
  Correlation coefficient ← ***this should precede regression!***
- Are the populations in the different classes compatible?
  Contingency table tests
- What is the shape of the empirical distribution function?
  EDF, spline, local nonparametric regression (MSMA Chpt 6)
- What are the distinct classes of objects in the sample?
  Hierarchical clustering (MSMA Chpt 9)

Concepts of nonparametric inference

Nonparametric: No assumed parametric model. No point/interval estimation. Hypothesis tests play an important role.

Distribution-free: Does not depend on (asymptotic) normality. Excellent for small samples, classificatory variables, and comparisons between samples. Probabilities often based on binomial/multinomial/combinatorial calculations.

Rank procedures: Operate on location in ordered list rather than actual values. Examples: quantiles (e.g. IQR), rank correlation

Robust (resistant): Results are insensitive to outliers or inhomogeneities. Breakdown point = fraction of outliers permitted for a given procedure.

See Wikipedia for the diverse meanings of `nonparametric statistics'

Limitations of nonparametric procedures

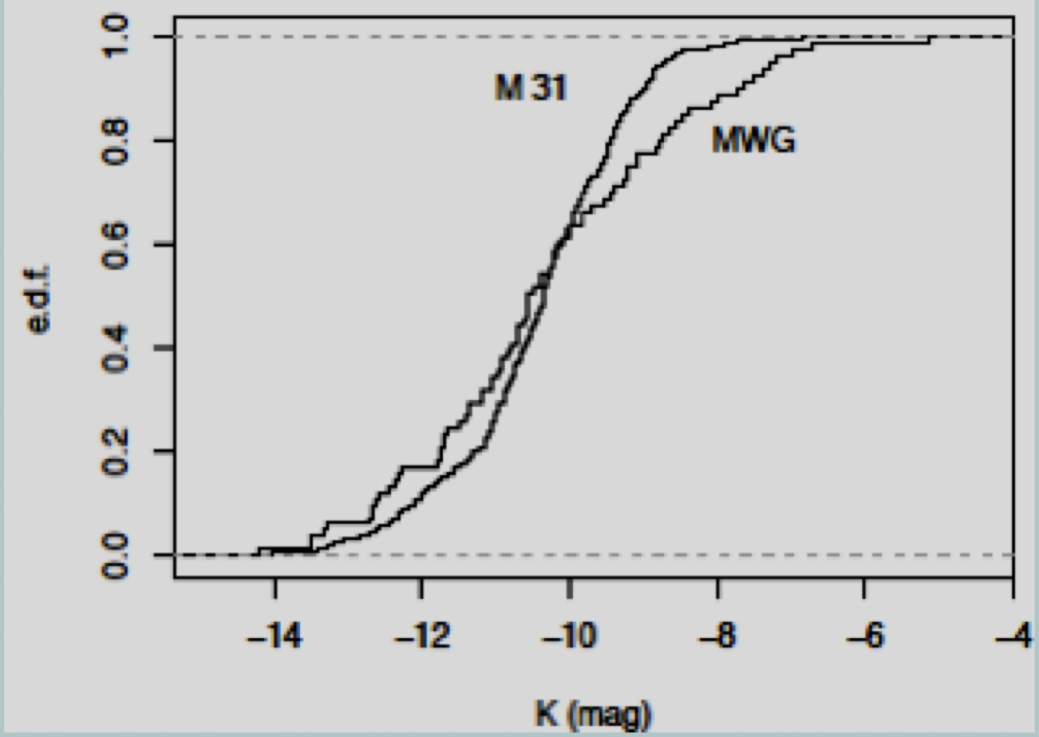
- Only a few statistics (functions) of data have known distributions for arbitrary data that allow probabilities to be estimated. For other statistics, bootstrap resampling can often give accurate distributions.

- Many methods are restricted to one-dimension; e.g., there is no unique ranking or empirical distribution fnction for multidimensional data. However, multidimensional two-sample tests are available with bootstrap probabilities.

- Some methods are heuristic (i.e. without theorems establishing their power, efficiency, etc)

Empirical distribution function

The empirical distribution function (e.d.f.) is the normalized sample cumulative distribution function (c.d.f.) for a univariate random variable. It is the step function from 0 to 1 with a jump of 1/n at the value of every data point.



Magnitude distribution of globular clusters in the Milky Way and Andromeda galaxies

Samples discussed in Appendix A.3. Datasets available at http://astrostatistics.psu.edu/MSMA/datasets

E.D.F.-based statistics and tests

$$M_{KS} = \sqrt{n} \max_x |\hat{F}_n(x) - F_0(x)|$$

$$P_{KS}(M_{KS} > \alpha) \simeq 2 \sum_{r=1}^{\infty} (-1)^{r-1} e^{-2r^2\alpha^2}$$

The **Kolmogorov-Smirnov** statistic. The maximum vertical distance between the e.d.f. and a second dataset or a model F_0.

$$T_{CvM,n} = n \int_{-\infty}^{\infty} [\hat{F}_n(x) - F_0(x)]^2 dF_o(x)$$
$$= \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F(x)i) \right)^2$$

The **Cramer-von Mises** statistic: the sum of squared vertical distances between the e.d.f. and the model.

$$A^2_{AD,n} = n \sum_{i=1}^{n} \frac{[i/n - F_0(X_i)]^2}{F_0(X_i)(1 - F_0(X_i)}.$$

The **Anderson-Darling** statistic: a tail-weighted CvM statistic. This is the most sensitive e.d.f.-based test.

See Beware the Kolmogorov-Smirnov test! page on ASAIP

Robust estimators of location & spread

While the **mean** is commonly used to estimate the central location of a random variable, it is sensitive to non-Gaussianity and outliers. The **median**, or central value, is more **robust**. (If n is even, the median is the mean of the two central values.)

The **variance** is particularly sensitive to outliers with `breakdown' of 1/n% , and more robust estimators of spread are desired. The **interquartile range** (IQR, 25%-ile to 75%-ile distance) has `breakdown' of 25%. The most stable estimator is the **median absolute deviation**

$$MAD = Med|x_i - Med|$$

1.48xMAD is approximately equal to the **standard deviation (√Var)** when the distributions are approximately Gaussian. The variance of the median can also be estimated by bootstrap resampling.

Some hypothesis tests in R and CRAN (many but not all are nonparametric)

```
#   ks.test, wilcox.test, mood.test (R) for univariate 2-sample test
#   chisq.test, fisher.test (R) for contingency tables (categorical data)
#   cor (R) Pearson r, Kendall tau, Spearman rho tests for correlation
#   ad.test (ADGofTest, ksamples) for univariate Anderson-Darling test
#   surv2.ks (surv2sample) for univariate 2-sample test with censoring (upper limits)
#   cenken (NADA) for bivariate correlation test with censoring
#   dip (diptest) for Hartigan's test for univariate multimodality
#   grubbs.test (outliers) test for outliers
#   durbin.watson (car) test for serial autocorrelation
#   cramer.test (Cramer) for multivariate 2-sample test with bootstrap resample
#   mshapiro.test (mvnormtest) for multivariate normality test
#   moran.test (sped) test for randomness vs. autocorrelation in 2 or more dimensions
#   kuiper.test, r.test, rao.test, watson.test (CircStats) tests for uniformity of circular data
```

Parametric inference

The inferential process is based on assumptions that the underlying distributions and relationships have a known functional form (e.g. linear relationships with normal scatter).

Astronomers often use nonlinear models based on astrophysical theory: isothermal sphere, thermal bremsstrahlung spectrum with emission lines, ΛCDM cosmology, etc. This is not common in other fields (e.g. biostatistics, econometrics).

Point estimation (= parameter estimation)

In classical parametric estimation, the observations are assumed to be i.i.d. (independent and identically distributed) values of r.v.'s (random variables) with known probability distributions (statistical models). These distributions are typically characterized by a small number of parameters, often noted by a p-dimensional vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$.

The goal of estimating plausible or `best' values based on the observations is called ***point estimation***. The point estimator (pronounced `theta-hat') is a function of the data $(x_1, x_2, \ldots, x_n)$

$$\hat{\theta} = g(x_1, \ldots, x_n)$$

Such functions are often called ***statistics***. The estimators are random variables, while the true value $\theta = g(\mathbf{X})$ of the underlying population **X** is a fixed number.

Example of point estimation with an astrophysical model

The statistical model for an exoplanet orbit has a vector θ with six parameters: semi-major axis, eccentricity, inclination, ascending node longitude, argument of periastron, and true anomaly. The estimation of `best fit' values of θ is from a dataset is an example of `point estimation'.

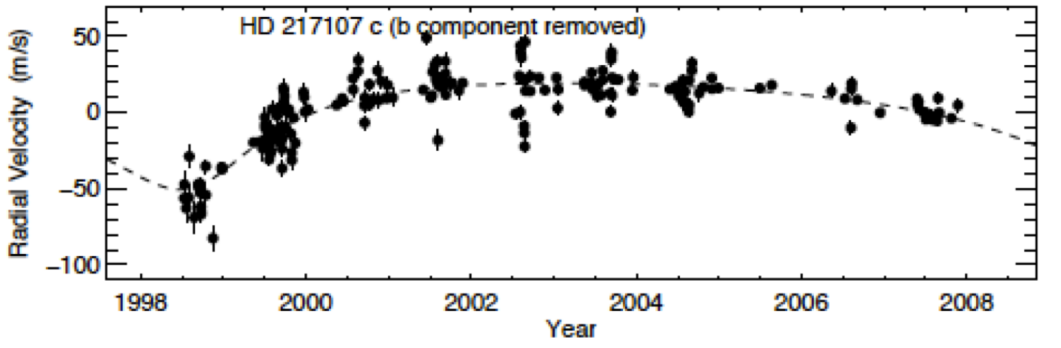


Fig. 7.— RV curves for HD 217107. The data are from Keck Observatory and Lick Observatory, and show the inner planet (top) with $P = 7.1$ d and $M \sin i = 1.4 M_{\rm Jup}$ and the outer planet (bottom) with $P = 11.6$ yr and $M \sin i = 2.6\ M_{\rm Jup}$.

Wright et al. (2009)

Confidence intervals (normal approx)

Point estimates cannot be perfectly accurate, at least due to sampling variations. The estimation of confidence intervals around the $\hat{\theta}$ estimators is a common form of **interval estimation**. The confidence interval of a r.v. statistic Y dependent on the r.v. X has lower and upper values defined by

$$P[l(X) < Y < u(X)] = 1 - \alpha$$

where 0<α<1 is usually chosen α=0.05 (0.01) giving the `95% (99%) confidence interval of Y'. Some statistics exhibit **asymptotic normality** either by parametric assumption or application of the Central Limit Theorem.

Consider estimators of the mean and standard deviation of a normally (Gaussian) distributed variable X. The 95% confidence interval can be expressed three ways (astronomers are more familiar with the third):

$$\begin{aligned} P(-1.96 < \sqrt{n}(\bar{x} - \hat{\mu})/\hat{\sigma} < 1.96) &= 0.95 \\ P(\bar{x} - 1.96\,\hat{\sigma}/\sqrt{n} < \hat{\mu} < \bar{x} + 1.96\,\hat{\sigma}/\sqrt{n}) &= 0.95 \\ \hat{\mu} &= \bar{X} \pm 1.96\,\hat{\sigma}/\sqrt{n} \end{aligned}$$

Confidence intervals (resampling)

However, astronomers encounter many cases where the sample is small and the Central Limit Theorem does not apply, or where the statistic of interest is derived in a non-standard fashion so that its distribution is difficult or impossible to calculate. In such cases, confidence intervals based on asymptotic normality may be very unreliable.

Fortunately, statisticians a class of computationally intensive procedures known as **resampling methods** where simulated samples are drawn from the data in specified ways. Powerful theorems (1980s) demonstrated that they provide inference on a wide range of statistics under very general conditions. These include the **bootstrap** and **cross-validation.**

The simulated populations derived from the observations are analyzed however the original dataset is analyzed to achieve science goals. A simple histogram of the statistic of interest gives confidence intervals due to random variations in the observations.

Bootstrap resampling

Nonparametric bootstrap: samples with replacement from the dataset

Parametric bootstrap: samples with replacement from a model derived from the dataset

Proposed by Bradley Efron in the 1970s, its importance emerged during the 1980s with theorems demonstrating that it gives nearly optimal estimate of the distribution of many statistics under a wide range of circumstances. Computation is often easy as only $N_{sim} \sim O[N(\ln N)^2]$ simulations are needed. $N_{sim} \sim$ 50:2000:50,000 for N=10:100:1000.

Limitations:

- Requires ***pivotal statistics*** (independent of model parameters). These include mean & variance, least squares and maximum likelihood estimators, correlation & regression coefficients, etc, etc.
- Requires ***homoscedasticity*** (treats all data points equally). The simple bootstrap is not applicable to data with heteroscedastic measurement errors or spatial/temporal autocorrelation.

What makes the `best' estimation method?

Often several reasonable estimators are available for measuring very similar properties of a dataset (e.g., mean and median for central location; kernel density estimator or Gaussian Processes regression for data smoothing). A great deal of effort has been exerted to interpret possible meanings of the word `best' because statistical point estimators have several important properties that often can not be simultaneously optimized.

Unbiasedness The ***bias*** an estimator theta-hat is

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

the difference between the expected value (i.e., mean) of the estimator for an ensemble of datasets and the true value of the parameter for the underlying population. This is not the error of a particular instantiation of theta-hat from a particular dataset; rather, this is an intrinsic offset in the estimator, even with an unlimited amount of data. An estimator is unbiased if *B=0*. But as this often cannot be achieved, we often seek to minimize the ***mean square error (MSE or MISE)*** of the estimator,

$$\begin{aligned} E((\hat{\theta} - \theta)^2) &= \mathrm{Var}(\hat{\theta}) + (E(\hat{\theta} - \theta))^2 \\ MSE &= \text{Variance of } \hat{\theta} + (\text{Bias})^2 \end{aligned}$$

Efficiency Among a collection of unbiased estimators, the most efficient one has the lowest asymptotic variance, $Var(\hat{\theta})$. In the best cases, it is equal to the Cramer-Rao bound. This is often called the ***minimum variance unbiased estimator*** **(MVUE)**. An important subset of MVUE's are the `best linear unbiased estimators' (BLUEs).

Consistency A `consistent' estimator will approach the true population parameter value as the sample size increases.

Sufficiency and completeness Technical issues regarding ancillary information in, and functions, of the statistic.

Asymptotic normality Related to the Central Limit Theorem, this requires that an ensemble of consistent estimators $\hat{\theta}(n)$ for sample size n has a distribution around the true population value that approaches a normal (Gaussian) with variance decreasing as 1/n:

$$\hat{\theta}(n) - \theta \rightarrow \frac{1}{\sqrt{n}} N(0, Var(\hat{\theta}))$$

where $N(\mu,\sigma^2)$ is the normal distribution. This is the ***Cramer-Rao bound***.

Estimating non-standard statistics

Mathematical statisticians have worked industriously for decades to establish the biasedness, consistency, asymptotical normality and other properties of various point estimators derived in various ways under various conditions, often with theorems that unequivocably demonstrate MLE, MVUE and other desirable properties.

However, the logical inverse of this situation is also true: when mathematical statisticians have ***not*** established the properties of a statistic, then the properties of that statistic can ***not*** be assumed to be known. Indeed, even subtle changes in the assumptions (e.g., a sample is independently but not identically distributed; a sample is i.i.d. but not normally distributed) can invalidate a critical theorem.

There is thus no guarantee that a new statistic, or a standard statistic examined under non-standard conditions, will have desirable properties such as unbiasedness and asymptotic normality. That is, the astronomer may not be able to assign probabilities to values of the statistic (e.g. that 3xRMS corresponds to P=0.003 confidence levels).

When non-standard statistics are considered, the bootstrap approach to confidence intervals is strongly recommended.

Methods of point estimation

The effort to estimate true values of parametric distributions from limited data is called ***point estimation***. Astronomers often call this ***parameter estimation***. Example: Estimating the mean and variance of a normal (Gaussian) distribution,

$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

The unbiased, consistent, asymptotically normal estimators from a sample of *n* data points are:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

Such estimators can be dreamed up, but are generally derived by several methods developed during the 19-20th centuries.

Method of least squares

Legendre, Laplace, Gauss and a host of astronomers developed least squares estimation during the 19th century. Here the estimator is chosen to minimize the sum of squared residuals about the expected value. For example, the least squares estimator of the expected value E[X]= μ of a univariate distribution is

$$\hat{\mu}_{LS} = \arg\min_{\mu} \sum_{i=1}^{n} (X_i - \mu)^2$$

In this simple case, μ–hat is the sample mean, the intuitive solution. But in more complex estimation problems, particularly in the context of regression, one minimizes $\Sigma[x_i\text{-}f(x_i)]^2$. This method provides solutions that are not intuitively obvious.

For a linear regression problem, if the error variances are different (heteroscadastic) and are known, then one can minimize the weighted sum of squares:

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(X_i - \sum_{j=1}^{k} a_{ij} \beta_j \right)^2$$

This weighted least squares procedure is what astronomers call `*minimum chi squared*' regression that has lots of difficulties with complex problems. E.g. the result is unbiased only when all of the variance is attributed to the measurement errors (no intrinsic scatter). Maximum likelihood estimation is preferred for complex problems like heteroscedastic measurement errors (e.g. B.C. Kelly, Astrophysical Journal, 2007).

NOTE: An important theorem proves that for a wide class models with normal errors, the least squares estimator is the maximum likelihood estimator. For this reason, least squares methods are still in very common use today.

Maximum likelihood method

The young British mathematician R. A. Fisher issued scathing critiques of least squares procedures during 1912-22, and advocated a philosophically different criterion: the `best' estimator would be ***the most probable set of parameter values given the data and the model***. ***This maximizes the likelihood***, the product of the probabilities that each data point arises from the model given a choice of model parameters:

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

The model parameters theta are treated as fixed quantities, while the data points x_i are samples from a random variable. Fisher's ***maximum likelihood estimator (MLE***) θ-hat is the value that maximizes L(θ). For computational convenience, it is often easier to maximize 2ln(L), or minimize -2ln(L), where the product becomes a sum.

Fisher's 1922 paper defining MLE also introduces the concepts of consistency, sufficiency, efficiency and `information'. The MLE is usually (but not always) unbiased and is often the MVUE.

Calculating MLEs: EM Algorithm

Likelihoods can be maximized by any numerical optimization method; e.g. Newton-Raphson. But likelihood functions can often have complicated morphologies, with multiple maxima, flat vs. steep regions, and dependence on starting values. In the 1970s, an alternative method emerged with theorems showing that the algorithm increases the likelihood during each iteration (Dempster, Laird & Rubin JRSS 1977, Wu 1983). This is the **Expectation-Maximization Algorithm.**

The method had been developed earlier for specialized situations, including image deconvolution in astronomy where the point spread function is known (Richardson JOSA 1972, Lucy AJ 1972). The astronomers' **Lucy-Richardson algorithm** is thus an implementation of the statisticians' **EM Algorithm**.

MLE parameter confidence intervals

Recall: whereas ***point estimation*** seeks `best' parameter values based on the data, ***interval estimation*** seeks critical regions consistent with the data,

$$P[l(X) < Y < u(X)] = 1 - \alpha$$

where α=0.05, 0.01 or 0.003. E.g., if the observation were conducted 100 times, 99 times the statistic Y would like within the stated range for α=0.01. These intervals should be valid, optimal and invariant.

In many situations, the estimator θ-hat has an approximate normal distribution, when n is large. For unbiased MLEs, θ-hat has mean θ and variance $1/\mathrm{I}(\theta)$ where I is the ***Fisher information matrix***

$$I(\theta) = nE\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x};\theta)\right)^2$$

That is, the parameter confidence interval is derived from the shape of the likelihood function around the best-fit value. The ***Cramer-Rao inequality*** gives a lower bound to the variance of an unbiased estimator. Note that MLE, MVUE and other confidence interval estimators may differ, and the choice may not be clear.

Resampling methods for point/interval estimation

Although MLE and other traditional estimators have many advantages, astronomers often pursue statistical analyses for which the likelihood cannot be constructed, the distribution of statistics invented by the scientists to measure particular characteristics of the datasets are unknown, the bias is not established, and/or asymptotic normality does not apply (e.g. small-N samples).

In such cases, resampling methods can be effectively used whereby the bias and variance of a statistic are estimated from datasets constructed from the observations by bootstrap Monte Carlo methods.